Research Article | Vol. 13, No. 10 / 1 Oct 2022 / Biomedical Optics Express 5082
Biomedical Optics EXPRESS

Using deep learning for the automated identification of cone and rod photoreceptors from adaptive optics imaging of the human retina

MENGXI ZHOU,[1,*] NATHAN DOBLE,[2,3] STACEY S. CHOI,[2,3] TIANYU JIN,[1] CHENWEI XU,[4] SRINIVASAN PARTHASARATHY,[1] AND RAJIV RAMNATH[1]

[1]The Ohio State University, Department of Computer Science and Engineering, 2015 Neil Ave., Columbus, OH 43210, USA
[2]The Ohio State University, College of Optometry, 338 W 10th Ave., Columbus, OH 43210, USA
[3]The Ohio State University, Department of Ophthalmology and Visual Science, Havener Eye Institute, 915 Olentangy River Road, Columbus, OH 43212, USA
[4]The Ohio State University, Department of Statistics, 127 Pomerene Hall, 1760 Neil Ave, Columbus, OH 43212, USA
*zhou.2656@osu.edu

Abstract: Adaptive optics imaging has enabled the enhanced in vivo retinal visualization of individual cone and rod photoreceptors. Effective analysis of such high-resolution, feature rich images requires automated, robust algorithms. This paper describes RC-UPerNet, a novel deep learning algorithm, for identifying both types of photoreceptors, and was evaluated on images from central and peripheral retina extending out to 30° from the fovea in the nasal and temporal directions. Precision, recall and Dice scores were 0.928, 0.917 and 0.922 respectively for cones, and 0.876, 0.867 and 0.870 for rods. Scores agree well with human graders and are better than previously reported AI-based approaches.

1. Introduction

Adaptive optics (AO) enhanced imaging modalities, namely flood illuminated [1], confocal laser scanning ophthalmoscopy (SLO) [2] and optical coherence tomography (OCT) [3–6] enable imaging of the human retina with lateral resolutions approaching a few micrometers. These dense, feature rich datasets require the development of automated algorithms for tractable analysis and interpretation. Considering just the cones, peak foveal density can approach 200,000 cones per mm^2 [7] precluding the use of manual identification and necessitating automated methods. The situation is further complicated in disease where their structure may change considerably when compared to healthy controls.

The first approaches towards automated analysis of the cone photoreceptor mosaic employed techniques based on local intensity maxima determinations [8–11], modelling and normalized cross-correlation [12], Hough transform [13], and optical flow and K-means clustering [14]. However, such techniques were often subject to particular conditions, for example a specific retinal eccentricity, imaging modality, resolution, or ocular disease and hence have been challenging to generalize. To tackle this problem, neural network models, especially deep Convolutional Neural Networks (CNNs) [15], have been applied to retinal imaging, as they learn features of interest directly from the training data provided. Various model architectures have been reported, showing reliable and consistent performance in detecting cone photoreceptors over different conditions. For instance, a patch-based CNN model, C-CNN [16] was proposed to detect cone photoreceptors in either confocal or split detector images from healthy subjects. The

#470071 https://doi.org/10.1364/BOE.470071
Journal © 2022 Received 11 Jul 2022; revised 13 Aug 2022; accepted 16 Aug 2022; published 2 Sep 2022

Biomedical Optics EXPRESS

same research group extended their model to a multi-modality version [17] which processed simultaneously acquired confocal and split detector images together, with performance measured in subjects with achromatopsia. Another study used a Multi-Dimensional Recurrent Neural Network (MDRNN) [18] with only split detector images from healthy subjects and subjects with Stargardt's disease. Finally, a U-Net based Fully Convolutional Networks (FCN) model [19] was investigated on the same dataset from [16], showing improvements over the original approach.

The application of neural networks for identifying the rod photoreceptors has been more limited. Cunefare et al. [20] described a RAC-CNN model based on the U-Net architecture whereby the model utilized multi-modality inputs (i.e. both confocal and split detector AO-SLO images) and conducted a semantic segmentation for both cones and rods demonstrating good performance with the retinal eccentricity from 1° to 7°. However, split detector datasets are not available with imaging modalities such as flood illuminated AO cameras or AO-OCT and hence the ability to differentiate photoreceptor type based on just confocal images is unknown.

Moreover, previous studies only analyzed images taken from the central ±10° from the fovea. For example, Cunefare et al. [16] and Davidson et al. [18] used data from 500 to 2800 μm (approximately 1.8° to 10°) and 300 to 2800 μm (approximately 1° to 10°) respectively in all meridians, Cunefare et al. [17] used data from the fovea to 12° in the temporal and superior meridians, and Cunefare et al. [20] used data from 1° to 7° in the temporal retina. However, prevalent diseases such as diabetic retinopathy [21] and retinitis pigmentosa [22] show their first signs of disease in the peripheral retina, highlighting the importance of quantitative analyses at these locations. This work aims to extend the retinal eccentricities over which deep learning models can perform, a necessary first step before the challenging task of differentiating normal and healthy cells.

The work presented here directly differentiates rods and cones from only confocal images and moreover, extends the range of retinal eccentricities to 30° from the fovea. A novel CNN architecture inspired by UPerNet [23], named RC-UPerNet, is employed and results are compared and validated against other commonly used deep learning models as well as experienced human graders.

2. Methods

Human annotations of cones and rods were first acquired for all images in the dataset. These annotated images serve as the ground-truth for the training and testing of the CNN model, Fig. 1. In the training stage, the annotated images were processed to generate label maps. The model was then trained on these maps along with the original registered AO-SLO images. The trained model could then generate probability maps for the three classes (cones, rods and background). A blob finding based method [24] was used to localize cones and rods from the probability maps. In the testing stage, the trained model was used to identify cones and rods in unseen images with the results compared against human annotations.

In the following subsections, the four main components of the method are described, namely: data collection, image pre-processing, the identification model architecture, and post-processing / localization.

2.1. Data collection

Five young, healthy subjects were imaged at 12-14 retinal locations, namely the fovea, 3°, 5°, 10°, 15°, 20°, 25°, 30°, in both the nasal and temporal meridians [25] on a research-grade AO-SLO system described in detail in Wells Gray et al. [26]. Briefly, the AO-SLO system acquires 0.7° x 0.9° images of the retina at 60 frames per second at an imaging wavelength of 680 nm with a lateral resolution of 2 μm. Prior to imaging, a combination of 1% tropicamide and 2.5% phenylephrine was used to dilate the pupil and paralyze accommodation. During imaging, the subject looked at a fixation target displayed on a computer monitor, which corresponded to the

Biomedical Optics EXPRESS

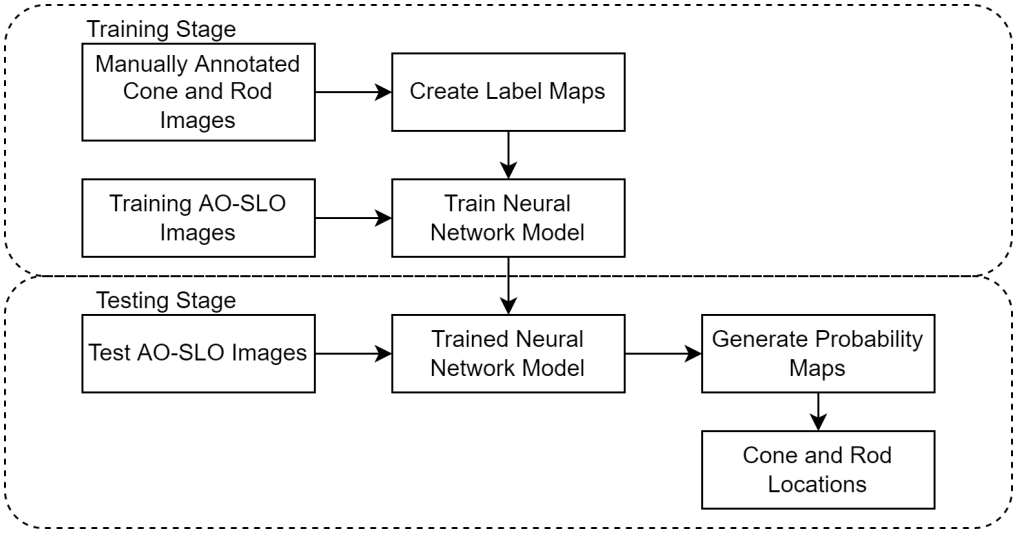


Fig. 1. Workflow of the model training and testing for cone and rod identification.

retinal location of interest. Multiple AO-SLO datasets (4-5 per retinal location, each containing 750-800 AO-SLO images) were acquired and subsequently post-processed to remove eye motion [27]. For the AO imaging, the tenets of the Declaration of Helsinki were observed, and the protocol approved by the Institutional Review Board of The Ohio State University (OSU). Written informed consent was obtained after all procedures were fully explained to the subject and prior to imaging.

For the manual annotation, all 66 registered images were presented to two experienced human graders independently. Both graders used a rule-based algorithm to first generate an initial estimate of the cone and rod photoreceptor centers. The algorithm took local maxima in terms of pixel intensities from the input AO-SLO image and constrained the results based on the user-provided cone and rod spacing. This initial estimation was then manually adjusted (by adding or removing photoreceptors) by the grader. Over the whole dataset, human grader 1 (HG1) marked a total of 6547 cones and 22765 rods, and human grader 2 (HG2) marked a total of 6254 cones and 21163 rods. Annotations from HG1 served as the ground-truth for the training and testing of the proposed neural network model; annotations from HG2 were used for validation purposes only.

2.2. Image preprocessing

Several preprocessing steps were applied before the training of the neural network model. For each AO-SLO image (a 10° nasal retina (NR) AO-SLO image is shown in Fig. 2(a), and a foveal image in Fig. 2(d)), an associated label map of the same size was created with each pixel on the label map representing one of three possible classes viz. 1: cone, 2: rod, or 0: background, according to the grader's manual labeling of the photoreceptor approximate centers (Fig. 2(b) and Fig. 2(e)). As the manual annotation only provided the photoreceptor centers rather than their boundaries, we extended the label to all pixels within a small circular local area to capture the photoreceptor boundary (Fig. 2(c) and Fig. 2(f)). This expansion mimicked the segmentation environment typically used when training a neural network model. As photoreceptor size and spacing varies with retinal eccentricity, using a universal setting for photoreceptor segmentation throughout the whole dataset (as in [20]) would not be suitable. Thus, the expansion was

Biomedical Optics EXPRESS

based on the unique photoreceptor features and distributions within each individual AO-SLO image. For each image, the cone radius and rod radius were first estimated by the photoreceptor center-to-center distances, using the following equations:

$$R_c = \min\Big(\min_{\substack{\forall c_1,c_2\in S_c \\ c_1\neq c_2}} 0.45 * dis(c_1,c_2), \min_{\substack{\forall c_1\in S_c \\ \forall r_1\in S_r}} 0.7 * dis(c_1,r_1)\Big) \tag{1}$$

$$R_r = \min\Big(\min_{\substack{\forall r_1,r_2\in S_r \\ r_1\neq r_2}} 0.45 * dis(r_1,r_2), \min_{\substack{\forall c_1\in S_c \\ \forall r_1\in S_r}} 0.2 * dis(c_1,r_1)\Big) \tag{2}$$

where R_c and R_r denote the estimated cone radius and rod radius in this image; S_c is the set of cone centers manually marked on this image; S_r is the set of rod centers; c_1 denotes any possible cone within S_c while c_2 denotes any other possible cone within S_c; the same applies for rods r_1 and r_2. Function $dis(a,b)$ is the Euclidean distance between two points a and b. All coefficients in the equations were heuristically chosen to maintain small separations between adjacent photoreceptors after expansion and photoreceptor size ratios. Once the radii were estimated, every manual photoreceptor center was then expanded to a circular area with the corresponding radius. All pixels in the area were then marked with the same class as the center pixel. All other pixels were set to be in the background class.

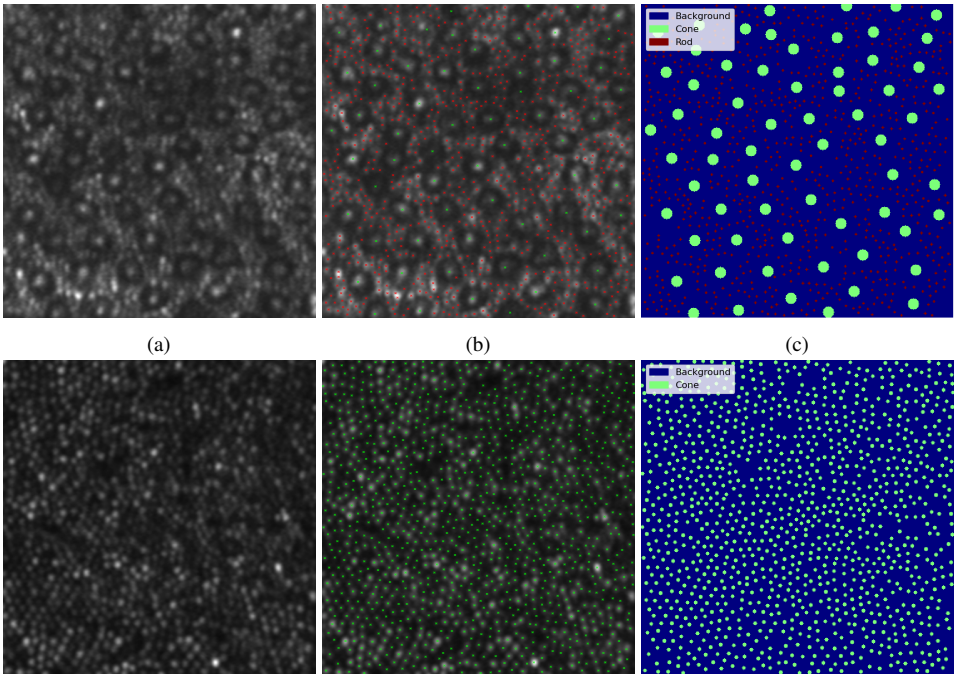


Fig. 2. (a): An example 10° NR AO-SLO image; (b): Manual annotations on the image where green and red marks indicate cones and rods respectively; (c): The label map generated from (b) with expansion; (d): An example foveal AO-SLO image; (e): Manual annotations on (d); (f): The label map generated from (e) with expansion.

A specific rule for the central fovea was that all cones were considered as rods for the neural network model. In this region the cones have a similar size to the rods and the packing geometry is also similar. These cones can experimentally bring confusion to the model in the learning

Biomedical Optics EXPRESS

stage and lower performance. In the testing stage, it is straightforward to revert all rod predictions back to cones in central fovea images, as no rods exist here [7].

2.3. Model architecture

The semantic segmentation network for the rod and cone identification, termed RC-UPerNet was built on the UPerNet (**U**nified **Per**ceptual Parsing **Net**work) model [23]. The original UPerNet model aims to recognize multiple visual concepts from a given image at once, such as the image scene (classification), object locations (segmentation), object materials (segmentation), etc. Despite its ability to learn multiple concepts, the model also achieves superior performance when used for single concept learning. Here, the model was modified from its original design to only retain one segmentation module, and trained to segment both cones and rods.

The proposed RC-UPerNet model is mainly composed of a Feature Pyramid Network (FPN) [28] with a Pyramid Pooling Module (PPM) [29] (see Fig. 3). The FPN consists of a bottom-up pathway, a top-down pathway, and lateral connections. The bottom-up pathway is often referred as an encoder, in which many convolutional layers are stacked and feature maps are down-sampled to different scales for high level semantic representations. Specifically, ResNet-50 [30] was used in our experiments as the encoder. The set of last feature maps of each stage in ResNet-50 were used to form the feature pyramid, whose down-sampling rates were $\{4, 8, 16, 32\}$. The top-down pathway on the other hand up-samples feature maps from higher pyramid levels to detailed and

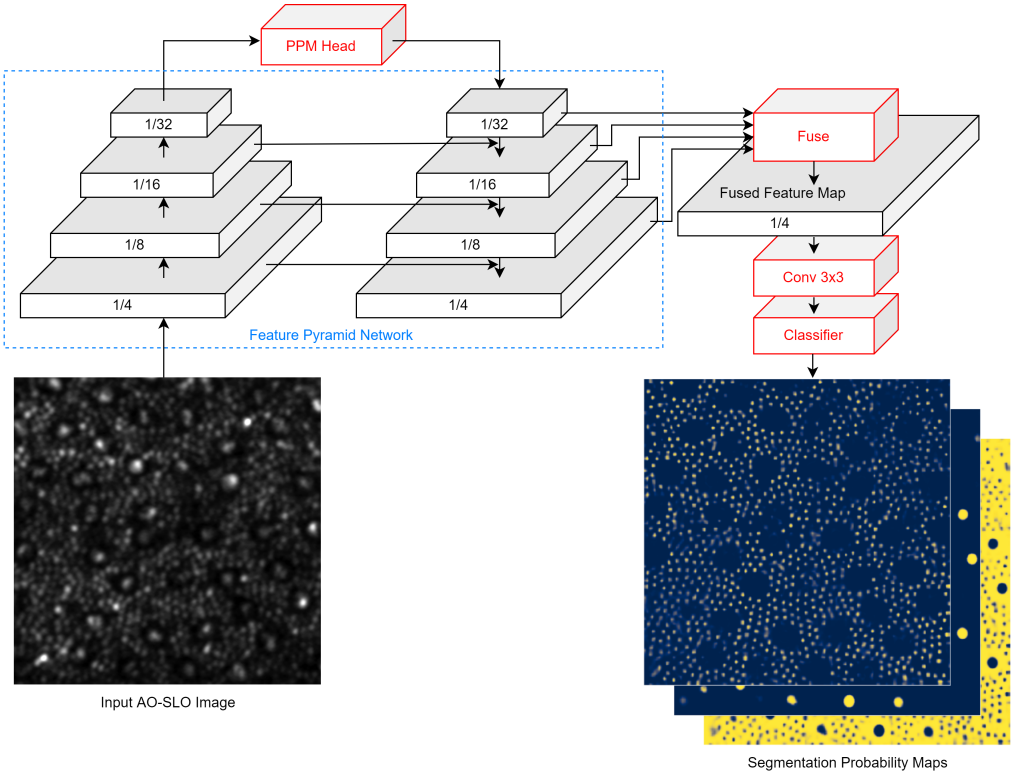


Fig. 3. The proposed RC-UPerNet model architecture. The model mainly contains a Feature Pyramid Network (FPN) [28] with a Pyramid Pooling Module [29]. A fusion module integrates all outputs from the FPN into a fused feature map, which will be consumed by the final classifier.

Biomedical Optics EXPRESS

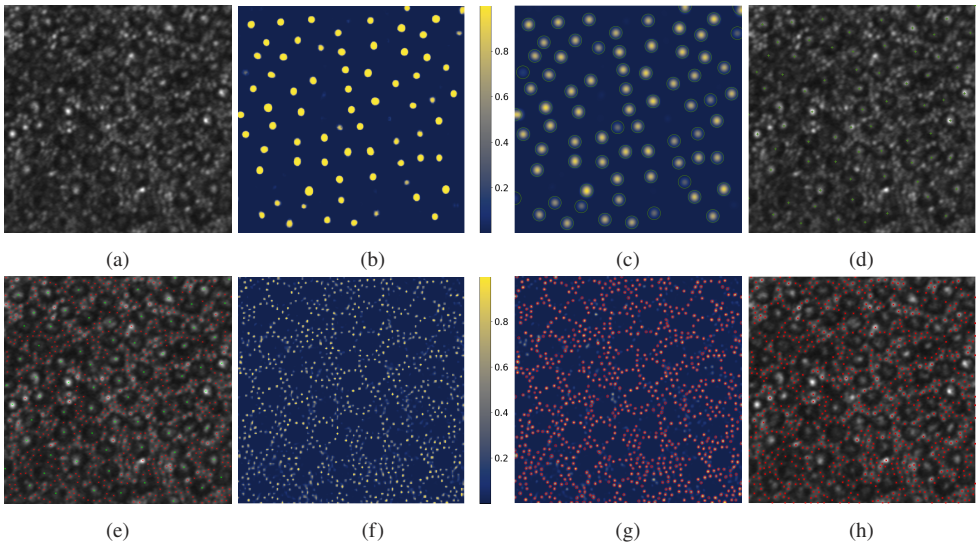


Fig. 4. (a): An example 10° TR AO-SLO image; (b): The cone probability map generated by the trained RC-UPerNet model. Color bar shows the percentage probability; (c): Cone blobs (green circles) found from (b) using the blob finding method; (d): Final detected cone positions (green marks) in the image; (e): Manually labeled photoreceptor centers for the image where green marks indicate cones and red marks indicate rods; (f): The rod probability map generated by the trained RC-UPerNet model; (g): Rod blobs (red circles) found from (f) where blobs close to cone blobs in (c) have been removed; (h): Final detected rod positions (red marks) in the image.

higher resolution features. These features are enriched with features of the same spatial sizes from the bottom-up pathway through lateral connections. Note these lateral connections are not simple skip connections [31], but rather the application of convolutional layers to the features from the bottom-up pathway before connecting them to the features in the top-down pathway. Outputs from each block in the top-down pathway are then fused (interpolated and concatenated) together to generate a comprehensive feature map. A convolutional layer with kernel size 3×3, and a classifier that used a convolutional layer with kernel size 1×1 along with a softmax activation function, are then applied to the above feature map for the final segmentation prediction.

In addition to FPN, PPM from PSPNet [29] is applied on the last layer of the bottom-up network before feeding it into the top-down pathway. PPM is designed to overcome the issue that the empirical receptive field of a deep CNN is relatively much smaller than in theory [32]. The module applies several pooling operations with different sizes to a starting feature map, aiming to harvest different sub-region representations. These pooled feature maps are then put through convolutional and upsampling layers and concatenated together along with the starting feature map. This forms a final feature representation in which both local and global contextual information are embedded. It was found that PPM is highly compatible with the FPN architecture by bringing effective global prior representations [23]. The original model design implementation described in [23] was used in our experiments. Other model settings, such as different numbers of layers, different filter sizes, individual FPN or PSPNet model, are yet to be explored. Readers are referred to [28] and [29] for further details on FPN and PPM.

Biomedical Optics EXPRESS

2.4. Post-processing and localization

In order to generate photoreceptor positions from the output probability maps of the neural network model, additional post-processing is required to localize the photoreceptors. To understand this, suppose the model yields an output of size $H \times W \times 3$, where $H \times W$ is the size of the input image. Each channel of the output may be thought of as a probability map corresponding to one of the three classes: cones, rods, or background. Each probability map indicates how likely each pixel from the original image belongs to a certain class. For localization, a blob finding method [24,33] was employed where cone and rod probability maps were processed separately. As an example, a 10° temporal retina (TR) AO-SLO image and its corresponding manually marked photoreceptor centers are shown in Fig. 4(a) and Fig. 4(e). For cone localization, a series of Laplacian of Gaussian (LoG) filters with increasing standard deviations were applied to the cone probability map (Fig. 4(b)) and the resulting images were stacked into a cube. With local maxima being extracted from the cube, thresholding and an overlapping condition, which eliminated smaller blobs if they overlapped with a larger blob by a fraction of 0.5, were applied to only retain the most proper blobs, Fig. 4(c). The corresponding center positions of the final blobs in the original image were then marked as cone locations, Fig. 4(d). The same process was applied for rods, Fig. 4(f)–4(h). This blob finding method outperformed the typical local-maximum finding method by a noticeable margin.

As the cone and rod probability maps were processed separately, it was possible that within the range of a cone blob, rod blobs were also detected. In such cases, the one that had the higher probability was retained. In other words, if a cone blob had a probability P_c (the maximum probability from this cone blob range in the cone probability map) and all other detected rod blobs had lower probabilities than P_c, then only the cone blob would be retained while all rod blobs would be removed. On the other hand, if one of the rod blobs had a higher probability than P_c, then all rod blobs in this range would be retained with the cone blob being removed.

3. Experiments and results

3.1. Model training

Owing to the size of the dataset, a five-fold cross-validation was implemented for the comprehensive analysis of the model performance [34], instead of a train-validation-test split on the dataset which would result in less number of images in each set. For each fold in the cross-validation, the data were split in an 80-20 manner where data from four subjects were used for model training, with one subject's data remaining unseen during training and was only used for testing. Each specific subject's data would serve as the test once in a specific fold. In the training stage, only annotations from the first human grader (HG1) were used as the ground-truth; results from the second human grader (HG2) were used only for validation.

The model was developed and run in Python 3.7 with the deep learning packages PyTorch [35] and MMSegmentation [36]. The model was trained for a total of 500 epochs with Stochastic Gradient Descent (SGD) as the optimizer. The batch size was set as 2 during training. The learning rate was set as 0.01 with a momentum of 0.9 and a weight decay of 0.0005. The loss function, L was based on the cross-entropy, that is:

$$L = -\sum_{(x,y)\in\Omega}\sum_{c=0}^{2} q(x,y,c)\cdot \log p(x,y,c), \tag{3}$$

where (x, y) denotes a pixel in the image Ω. c denotes a class, where the representations are 0 for background, 1 for cone, and 2 for rod. $q(x, y, c)$ is a binary indicator (0 or 1) showing if class label c is the correct classification for pixel (x, y), and $p(x, y, c)$ is the predicted probability that pixel (x, y) is of class c.

Biomedical Optics EXPRESS

In particular, note that the model was pre-trained on the ADE20K [37,38] dataset, and then fine-tuned on our cone and rod datasets. The ADE20K dataset contains 27,574 general scene images with 707,868 unique objects annotated from 3,688 categories. This common Transfer Learning practice [39,40] aims to transfer the information gained by a neural network model from a source domain where data is abundant to a target domain where the amount of training data may be limited. Additionally, random flipping with a chance of 50% on both vertical and horizontal directions and rotation with a chance of 25% for 0°, 90°, 180°, 270°, respectively were performed on images as data augmentation steps during the model training on the AO-SLO datasets.

Our computational infrastructure consisted of an NVIDIA Tesla P100 GPU with 16 GB of memory. Training averaged a time of 16.6 seconds per epoch, resulting in a total of 2.31 hours for a 500 epochs setting. Model inferencing on the test set averaged 0.18 seconds per image.

3.2. Performance measures

To measure the performance of the RC-UPerNet model, standard precision, recall and Dice coefficient scores were used. Each cone and rod detected by the algorithm was matched to the HG1's manual annotations for the identification of all true positives (TP), false positives (FP) and false negatives (FN), following the approaches presented in [17,20]. Taking the cones as an example, a predicted cone was considered a true positive if it was close enough to a manually marked cone by a distance d_c. The distance d_c was empirically set to $1.5 \times R_c$ from Eq. (1), as $1 \times R_c$ was found to be too strict for the measurement and created spurious FPs and FNs (FPs and FNs that were actually TPs in terms of human evaluation), while a larger number (e.g. $2 \times R_c$) resulted in excessive false TPs. Predicted cones that failed to match to a manually marked cone were considered false positives, and manually marked cones that did not have a matching predicted cone were considered false negatives. If multiple predicted cones matched a same manually marked cone, only the closest one was considered a true positive with the rest considered false positives. The same process was applied to rod photoreceptors, however with a different distance d_r set to $1.5 \times R_r$ from Eq. (2). To avoid border effects, predicted and manually marked photoreceptors within 10 pixels of the edges of the image were ignored during measurement. Finally, the precision, recall and Dice coefficients were computed based on each photoreceptor type, using the following equations:

$$Precision = \frac{N_{TP}}{N_{TP} + N_{FP}}, \tag{4}$$

$$Recall = \frac{N_{TP}}{N_{TP} + N_{FN}}, \tag{5}$$

$$Dice = \frac{2N_{TP}}{2N_{TP} + N_{FP} + N_{FN}}, \tag{6}$$

where N_{TP}, N_{FP} and N_{FN} denote the number of true positives, false positives and false negatives, respectively. Unless specifically stated, the precision, recall and Dice coefficients were computed in an aggregated manner over the dataset. In other words, the true positives, false positives and false negatives were first counted and summed over all images in the dataset, following which precision, recall and Dice coefficients were calculated.

3.3. Results

The performance of our RC-UPerNet implementation was compared against other state-of-the-art approaches, i.e. C-CNN [16], RAC-CNN [20], and a UNet++ [41] model. It should be noted that as the datasets and problem statements differed, the exact approaches presented in [16] and [20] could not be exactly applied. For example, C-CNN [16] only tried to classify between cones

Biomedical Optics EXPRESS

and backgrounds, without considering rods. Also, RAC-CNN [20] used both confocal and split detector images to train their model, whereas our dataset only contained confocal images. Hence, the following minimum adaptions were made to the reproduced models: for the C-CNN, the filter size was doubled in all convolutional layers and the final fully-connected layer altered to make a 3-class classification; For RAC-CNN, the pathway for split detector images was simply excised. Note that the capability of the full RAC-CNN model cannot be compared due to the lack of the split detector information, which can be a fundamental element to the model's superior performance. However, we made the minimum adaption above to the model so that they were most comparable on our dataset. More discussion on having additional input information is provided in Section 4. Moreover, an additional UNet++ based model was also implemented for further comparison. This UNet++ model was similar to the RAC-CNN model except that the core was changed from a U-Net architecture to the UNet++ architecture. Local-maximum finding was used as the post-processing for these models to localize the photoreceptors. Finally, the inter-grader agreement was measured by treating HG1's annotation as the ground-truth and HG2's annotation as the detection.

Table 1 presents the mean performance and standard deviations of all methods from the cross-validation experiments. It can be seen that the RC-UPerNet algorithm received the best performance where the mean Dice coefficient was 0.922 for cones and 0.870 for rods. Also, the standard deviation for the proposed method is much lower than the other methods, suggesting that RC-UPerNet is more robust when trained with different portions of the dataset. Moreover, the RC-UPerNet method had the closest performance to HG2.

Table 1. Performance of the proposed RC-UPerNet model against other competitive methods and HG2. All scores are averaged across the 5-fold cross-validation. Standard deviations are shown in the second row of each cell. The best score within each column is shown in bold and the second best is shown in italics.

Model	Cone			Rod		
	Precision	**Recall**	**Dice**	**Precision**	**Recall**	**Dice**
C-CNN [16]	0.891	0.825	0.856	0.785	0.839	0.810
	± 0.068	± 0.041	± 0.044	± 0.064	± 0.027	± 0.037
RAC-CNN [20]	*0.904*	0.896	*0.900*	*0.808*	0.851	0.829
	± 0.038	± 0.029	± 0.031	± 0.047	± 0.036	± 0.037
UNet++ [41]	0.895	*0.906*	*0.900*	0.798	**0.895**	*0.843*
	± 0.025	± 0.022	± 0.018	± 0.057	± 0.042	± 0.027
RC-UPerNet (Ours)	**0.928**	**0.917**	**0.922**	**0.876**	*0.867*	**0.870**
	± 0.012	± 0.016	± 0.008	± 0.029	± 0.040	± 0.014
HG2	0.953	0.935	0.944	0.919	0.830	0.869
	± 0.030	± 0.015	± 0.013	± 0.071	± 0.073	± 0.041

Figure 5 shows photoreceptor detection examples from our RC-UPerNet method and the RAC-CNN method for AO-SLO images from 3 representative retinal locations, i.e. 5° NR, 10° TR, and 30° NR. The detected cones and rods are measured against HG1's annotation where green marks stand for true positives, yellow for false positive, and red for false negative. The Dice coefficient for each method on each individual image is also shown. For conciseness, results from the C-CNN method are not shown in Fig. 5, but were generally poorer compared to the RAC-CNN and RC-UPerNet methods.

Biomedical Optics EXPRESS

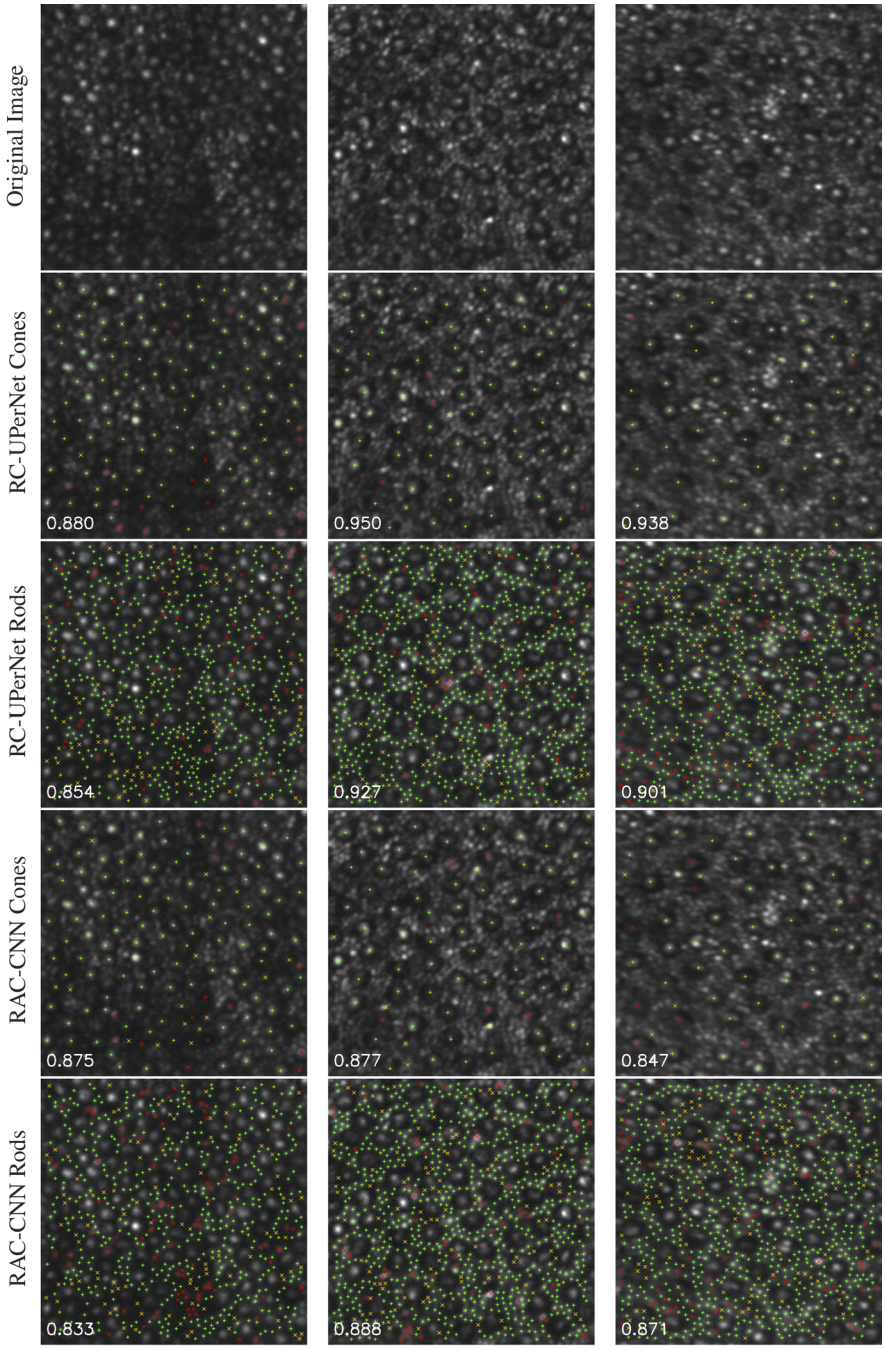


Fig. 5. Performance of the RC-UPerNet and RAC-CNN models on sample AO-SLO images taken from 5° NR, 10° TR, and 30° NR. All photoreceptor centers are measured by HG1's annotation where green marks denote true positives, yellow marks denote false positives, red marks denote false negatives. The Dice coefficient for each image is shown in the left bottom corner.

Biomedical Optics EXPRESS

4. Discussion

A key differentiator of this work is the wide range of retinal eccentricities that our datasets cover, namely out to 30° from the fovea. Previous efforts were limited to retinal locations within 10° (see Section 1) [16–18,20]. In the peripheral retina, the identification task becomes more challenging as both the size, spacing, luminance, and modal structure of the cones can vary significantly while the number of rods dramatically increase. Our method proves to be robust to such noise effects achieving stable results throughout the retina. However, the performance can diverge depending on retinal locations and types of the photoreceptor, e.g. the weaker rod identification shown in the 5° NR image in Fig. 5. A detailed breakdown of the model's performance for different retinal eccentricities is provided in Fig. 6. For rod identification, our RC-UPerNet model achieved better performance in the more peripheral retinal locations ($\geq 10°$) than in the more central retinal areas ($<10°$). On the other hand, for cone identification, the model generally attained satisfying performance (~ 0.9 Dice coefficient). However, it made slightly poorer predictions in the peripheral locations as some cones become bimodal in appearance. Also notice the lower Dice coefficients and larger standard deviations reported for cone identification at 15° and 30°, Fig. 6. These were primarily due to one lower quality image at each location which had individual RC-UPerNet Dice coefficients of 0.698 and 0.522 respectively. These two images also received lower Dice coefficients from RAC-CNN (0.686 and 0.588), UNet++ (0.686 and 0.556), and HG2 (0.808 and 0.667), suggesting that lower image quality also created uncertainties to other competitive models and even human graders. Removing these two images in the cones calculation resulted in improved mean Dice coefficients of 0.889 at 15° and 0.854 at 30°, comparable to the cone Dice scores at the other locations. Standard deviations of the Dice coefficients also reduced from 0.097 to 0.068 at 15° by removing the image, and from 0.144 to 0.062 at 30°.

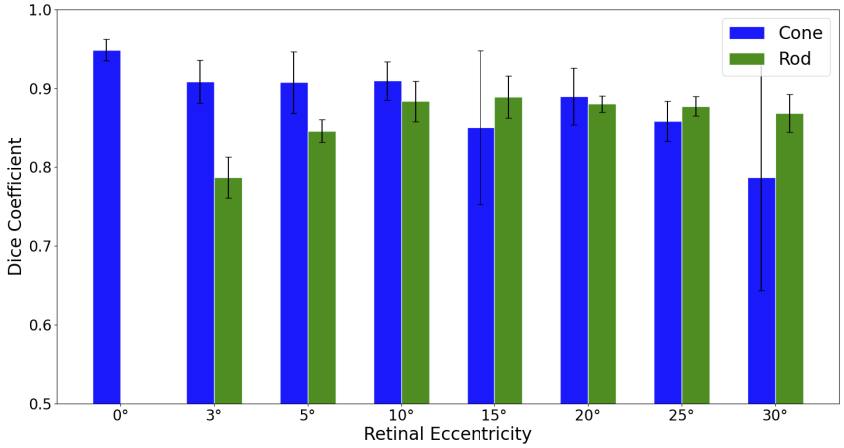
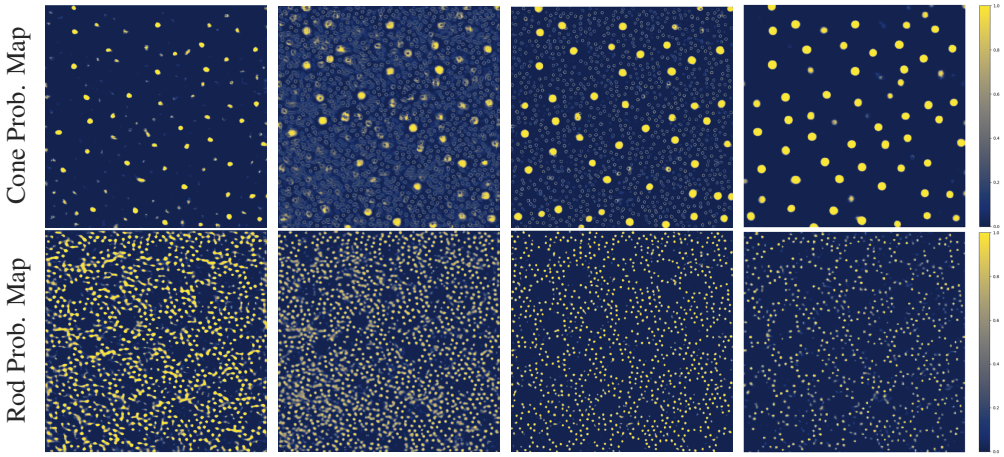


Fig. 6. Performance of RC-UPerNet model for different retinal eccentricities.

An underlying advantage of the proposed RC-UPerNet model was that it produced more precise probability maps (before post-processing) than the other models. Taking the 30° NR image in Fig. 5 for example, the probability map generated by the RC-UPerNet showed perceptible benefits for segmentation, as it cleanly and reliably put high probabilities for pixels to their correct classes (see Fig. 7). In contrast, the cone probability map from the C-CNN model [16] was noisier, i.e. some rod regions also had relatively high probabilities in the cone map. Also, its rod probability map had more connected areas, making it harder to locate each individual cell. The RAC-CNN model [20] had issues in that it gave relatively high probability to rod boundaries in the cone map

Biomedical Optics EXPRESS

and even higher probabilities to cone centers in the rod map. The UNet++ model [41] was better than the RAC-CNN model in that it had less wrong probabilities in the rod map for the cone center areas, however, it still it mishandled rod boundaries in the cone map. Poor probability maps from these models in turn require more complex post-processing procedures to properly handle the undesired patterns and more careful hand-tuning for good identification results (e.g. tuning on the thresholds, which can be again time-consuming for users). Hence, the proposed RC-UPerNet outperformed other models also in the sense that it produced more accurate and post-processing friendly results.



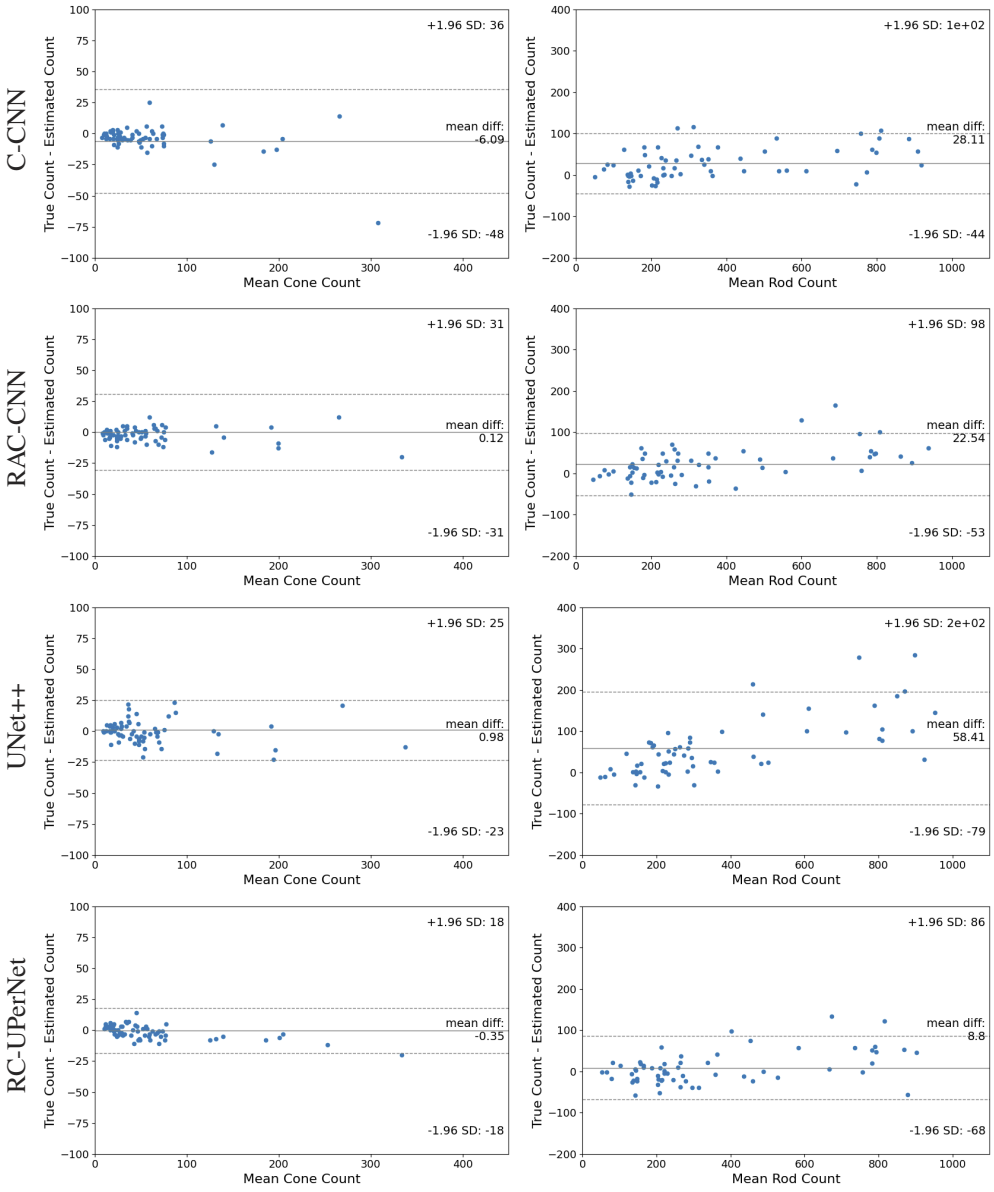
(a) C-CNN (b) RAC-CNN (c) UNet++ (d) RC-UPerNet

Fig. 7. Comparison of the model predicted probability maps between C-CNN [16], RAC-CNN [20], UNet++ [41], and the proposed RC-UPerNet. Cone probability maps are on the top row and rod probability maps are on the bottom row. Probability maps from the proposed RC-UPerNet are more precise and reliable for segmentation and localization of the photoreceptors.

For measuring and comparing model performance, the number of photoreceptors, a common index used in diagnosis, was further evaluated for our proposed RC-UPerNet and other models. Bland-Altman plots for cone and rod photoreceptor counts from each model against HG1 are shown in Fig. 8. For RC-UPerNet, the 95% confidence limits of agreement for the average difference in cone counts and rod counts were -0.35 ± 9.22 and 8.80 ± 39.18, respectively. The same metrics were -6.09 ± 21.40 and 28.11 ± 37.01 for the C-CNN model; -0.12 ± 15.74 and 22.54 ± 38.61 for the RAC-CNN model; 0.98 ± 12.39 and 58.41 ± 69.89 for the UNet++ model. The lower mean differences and narrower confidence limits from our method gave further evidence that it agrees better with the human grader, and hence is a more reliable photoreceptor detector.

Human graders can help improve the identification accuracy through manual adjustments on cones which also results in improved identification of rods. As mentioned in the post-processing (see Section 2.4), there were situations where the model made both cone and rod predictions in a small local area. In such cases, only the photoreceptor that had the higher probability would be retained. However, no model is perfect, and may give a higher probability to the wrong photoreceptor type, resulting in both incorrect cone and rod identification. Examples can be found in Fig. 5 where a false positive (yellow) cone prediction spawns several false negative (red) rods due to the removal of rod predictions close to this cone. On the other hand, a false negative cone could cause a false positive rod potentially because the intensity at the center makes the model confuse a cone with a rod. Hence, to test each model with no such confusing

Biomedical Optics EXPRESS



(a) Cone (b) Rod

Fig. 8. Bland-Altman plots of (a) cone and (b) rod photoreceptor counts for C-CNN [16], RAC-CNN [20], UNet++ [41], and the proposed RC-UPerNet. The mean difference is shown as the middle solid line, and the 95% confidence limits of agreement is shown as dotted lines.

Biomedical Optics EXPRESS

areas, experiments were conducted with the assumption that all cone locations were known beforehand. From such experiments, the overall Dice coefficient for rods was improved from 0.870 to 0.884. Precision and recall scores were also improved from 0.876 to 0.887 and 0.867 to 0.882, respectively. This suggested that further improvements on cone identification was more crucial since rod predictions would also benefit. We recommend the use of a semi-automated method where human graders iteratively adjust cone predictions; rod predictions will also be improved and require less manual correction. This can be especially helpful in the peripheral retina areas where the model has more confusion about cones (with the caveat that cones are large, sparsely distributed and thus only need a few manual adjustments).

Future work will seek improvements of the proposed method in the following directions. Firstly, a larger annotated AO-SLO image dataset will benefit our RC-UPerNet model, as the model is highly data-driven. Having the model see more training samples, especially for regions that currently received weaker predictions, will enable it to better learn the features and produce more accurate predictions. Because photoreceptor features and distributions vary over different retinal eccentricities, encoding meta information (such as retinal eccentricity and imaging parameters) into the model can potentially also improve performance. Ideally, the model would learn to focus on specific features, such as photoreceptor sizes or how rods are distributed among cones, for each different scenario, hence producing identifications of higher quality. Finally, if the imaging system allows the simultaneous acquirement of images from other modalities (e.g. split detector or an OCT channel) along with each confocal image, these images could serve as additional inputs and provide information from different perspectives to the model. For example, images from a modality that emphasizes cones while depressing rods will help better capture the cones. Note that we can use the same UPerNet architecture for each additional input image and concatenate the "Fused Feature Maps" before feeding them into the final two blocks. Promising improvements along this direction have been shown in [17] and [20].

5. Conclusion

This work describes and evaluates RC-UPerNet, a novel neural network model for the identification of rod and cone photoreceptors from AO-SLO confocal images. RC-UPerNet was trained and tested on images that covered a larger range of retinal eccentricities (0-30°) than hitherto considered in prior work, and where the identification task was more challenging due to the variation in photoreceptor features. Our proposed RC-UPerNet model received overall Dice coefficients of 0.922 for cones and 0.870 for rods, surpassing other published automated methods by a noticeable margin. Further contributions of this work include the collection of a larger annotated dataset with encoded meta information, and the use of other imaging modalities.

Funding. American Academy of Optometry - Allergan Foundation; Translational Data Analytics Institute at The Ohio State University; National Science Foundation (2133650).

Acknowledgements. We would like to acknowledge Sophie Araujo-Hernandez for her work in the manual grading of the images presented here. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the respective funding agencies.
Software Availability. The software will be made available for research purposes upon request to the corresponding author.

Disclosures. The authors declare that there are no conflicts of interest related to this article.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. J. Liang, D. R. Williams, and D. T. Miller, "Supernormal vision and high-resolution retinal imaging through adaptive optics," J. Opt. Soc. Am. A **14**(11), 2884–2892 (1997).
2. A. Roorda, F. Romero-Borja, W. J. Donnelly III, H. Queener, T. J. Hebert, and M. C. Campbell, "Adaptive optics scanning laser ophthalmoscopy," Opt. Express **10**(9), 405–412 (2002).

Biomedical Optics EXPRESS

3. B. Hermann, E. Fernández, A. Unterhuber, H. Sattmann, A. Fercher, W. Drexler, P. Prieto, and P. Artal, "Adaptive-optics ultrahigh-resolution optical coherence tomography," Opt. Lett. **29**(18), 2142–2144 (2004).
4. D. T. Miller, J. Qu, R. S. Jonnal, and H. Zhao, "Optical coherence tomography for an adaptive optics retina camera," in *19th Congress of the International Commission for Optics: Optics for the Quality of Life*, vol. 4829 (SPIE, 2003), pp. 641–643.
5. R. J. Zawadzki, S. M. Jones, S. S. Olivier, M. Zhao, B. A. Bower, J. A. Izatt, S. Choi, S. Laut, and J. S. Werner, "Adaptive-optics optical coherence tomography for high-resolution and high-speed 3D retinal in vivo imaging," Opt. Express **13**(21), 8532–8546 (2005).
6. Y. Zhang, J. Rha, R. S. Jonnal, and D. T. Miller, "Adaptive optics parallel spectral domain optical coherence tomography for imaging the living retina," Opt. Express **13**(12), 4792–4811 (2005).
7. C. A. Curcio, K. R. Sloan, R. E. Kalina, and A. E. Hendrickson, "Human photoreceptor topography," J. Comp. Neurol. **292**(4), 497–523 (1990).
8. B. Xue, S. S. Choi, N. Doble, and J. S. Werner, "Photoreceptor counting and montaging of en-face retinal images from an adaptive optics fundus camera," J. Opt. Soc. Am. A **24**(5), 1364–1372 (2007).
9. K. Y. Li and A. Roorda, "Automated identification of cone photoreceptors in adaptive optics retinal images," J. Opt. Soc. Am. A **24**(5), 1358–1363 (2007).
10. D. H. Wojtas, B. Wu, P. K. Ahnelt, P. J. Bones, and R. Millane, "Automated analysis of differential interference contrast microscopy images of the foveal cone mosaic," J. Opt. Soc. Am. A **25**(5), 1181–1189 (2008).
11. S. J. Chiu, Y. Lokhnygina, A. M. Dubis, A. Dubra, J. Carroll, J. A. Izatt, and S. Farsiu, "Automatic cone photoreceptor segmentation using graph theory and dynamic programming," Biomed. Opt. Express **4**(6), 924–937 (2013).
12. A. Turpin, P. Morrow, B. Scotney, R. Anderson, and C. Wolsley, "Automated identification of photoreceptor cones using multi-scale modelling and normalized cross-correlation," in *International Conference on Image Analysis and Processing*, (Springer, 2011), pp. 494–503.
13. D. M. Bukowska, A. L. Chew, E. Huynh, I. Kashani, S. L. Wan, P. M. Wan, and F. K. Chen, "Semi-automated identification of cones in the human retina using circle hough transform," Biomed. Opt. Express **6**(12), 4676–4693 (2015).
14. Y. Chen, Y. He, J. Wang, W. Li, L. Xing, X. Zhang, and G. Shi, "Automated cone cell identification on adaptive optics scanning laser ophthalmoscope images based on tv-l1 optical flow registration and k-means clustering," Appl. Sci. **11**(5), 2259 (2021).
15. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances Neural Proc. Syst. **25**, 1097–1105 (2012).
16. 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," Sci. Rep. **7**(1), 6620 (2017).
17. D. Cunefare, C. S. Langlo, E. J. Patterson, S. Blau, A. Dubra, J. Carroll, and S. Farsiu, "Deep learning based detection of cone photoreceptors with multimodal adaptive optics scanning light ophthalmoscope images of achromatopsia," Biomed. Opt. Express **9**(8), 3740–3756 (2018).
18. 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," Sci. Rep. **8**(1), 7911 (2018).
19. J. Hamwood, D. Alonso-Caneiro, D. M. Sampson, M. J. Collins, and F. K. Chen, "Automatic detection of cone photoreceptors with fully convolutional networks," Trans. Vis. Sci. Tech. **8**(6), 10 (2019).
20. 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," Biomed. Opt. Express **10**(8), 3815–3832 (2019).
21. J. Lechner, O. E. O'Leary, and A. W. Stitt, "The pathology associated with diabetic retinopathy," Vision Res. **139**, 7–14 (2017).
22. D. T. Hartong, E. L. Berson, and T. P. Dryja, "Retinitis pigmentosa," The Lancet **368**(9549), 1795–1809 (2006).
23. T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), pp. 418–434.
24. T. Lindeberg, "Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention," Int. J. Comput. Vis. **11**(3), 283–318 (1993).
25. E. Wells-Gray, S. Choi, A. Bries, and N. Doble, "Variation in rod and cone density from the fovea to the mid-periphery in healthy human retinas using adaptive optics scanning laser ophthalmoscopy," Eye **30**(8), 1135–1143 (2016).
26. E. M. Wells-Gray, S. S. Choi, R. J. Zawadzki, S. C. Finn, C. A. Greiner, J. S. Werner, and N. P. Doble, "Volumetric imaging of rod and cone photoreceptor structure with a combined adaptive optics-optical coherence tomography-scanning laser ophthalmoscope," J. Biomed. Opt. **23**(03), 1 (2018).
27. S. B. Stevenson and A. Roorda, "Correcting for miniature eye movements in high-resolution scanning laser ophthalmoscopy," in *Ophthalmic Technologies XV*, vol. 5688 (SPIE, 2005), pp. 145–151.
28. T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017, pp. 2117–2125.
29. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017, pp. 2881–2890.

Biomedical Optics EXPRESS

30. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016, pp. 770–778.
31. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015, pp. 3431–3440.
32. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016, pp. 2921–2929.
33. "Scikit-image blob detection," https://scikit-image.org/docs/stable/auto_examples/features_detection/plot_blob.html. Accessed: 2022-08-06.
34. R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, vol. 14 (Montreal, 1995), pp. 1137–1145.
35. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds. (Curran Associates, Inc., 2019), pp. 8024–8035.
36. M. Contributors, "MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark," https://github.com/open-mmlab/mmsegmentation (2020).
37. B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2017), pp. 633–641.
38. B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," Int. J. Comput. Vis. **127**(3), 302–321 (2019).
39. C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *International conference on artificial neural networks*, (Springer, 2018), pp. 270–279.
40. M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2014), pp. 1717–1724.
41. Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, (Springer, 2018), pp. 3–11.