Analyzing Adjacent B-Scans to Localize Sickle Cell Retinopathy In OCTs

Ashuta Bhattarai[1], Jing Jin[2], and Chandra Kambhamettu[1]

[1] Video/Image Modeling and Synthesis (VIMS) Lab, University of Delaware, USA
{ashutab, chandrak}@udel.edu
[2] Nemours Children's Hospital, Delaware, USA
jing.jin@nemours.org

Abstract. Imaging modalities, such as Optical coherence tomography (OCT), are one of the core components of medical image diagnosis. Deep learning-based object detection and segmentation models have proven efficient and reliable in this field. OCT images have been extensively used in deep learning-based applications, such as retinal layer segmentation and retinal disease detection for conditions such as age-related macular degeneration (AMD) and diabetic macular edema (DME). However, sickle-cell retinopathy (SCR) has yet to receive significant research attention in the deep-learning community, despite its detrimental effects. To address this gap, we present a new detection network called the Cross Scan Attention Transformer (CSAT), which is specifically designed to identify minute irregularities such as SCR in cross-sectional images such as OCTs. Our method employs a contrastive learning framework to pretrain OCT images and a transformer-based detection network that takes advantage of the volumetric nature of OCT scans. Our research demonstrates the effectiveness of the proposed network in detecting SCR from OCT images, with superior results compared to popular object detection networks such as Faster-RCNN and Detection Transformer (DETR). Our code can be found in: github.com/VimsLab/CSAT.

Keywords: Optical coherence tomography · Sickle-cell retinopathy · Object detection · Medical image diagnosis

1 Introduction

Optical Coherence Tomography (OCT) [9] is a medical imaging technique that generates high-resolution cross-sectional images of a retina in real-time. Due to the volumetric nature of the retina, an OCT scan produces multiple cross-sectional images that expose the retinal layers, which serve as a diagnostic tool for ophthalmologists to detect and monitor retinal diseases. Deep learning techniques have been applied to OCT images to analyze various retinal diseases, such as age-related macular degeneration (AMD), diabetic macular edema (DME), and glaucoma. However, despite its significant impact, deep learning-based research on sickle-cell retinopathy (SCR) remains limited [14].

SCR is an ocular condition affecting the retina of sickle cell disease (SCD) patients and can lead to severe vision impairment or blindness. With a global incidence of 300,000 neonates and 100,000 affected individuals in the United States alone, SCR is a significant public health concern [25]. The condition interrupts retinal blood circulation and causes damage to retinal tissues through selective thinning of the inner retinal layers, which can be observed in OCT-generated cross-sectional images (B-scans). The progression of SCR is an important area of research in ophthalmology, as it can provide valuable insights for developing new measures to control its effects. However, diagnosing SCR accurately can be challenging, and incorrect diagnoses can cause significant emotional and financial distress for patients and healthcare systems.

To locate, diagnose, and analyze the severity of SCR, the experts manually study the B-scans of SCD patients. The ophthalmologists specifically look for a pit that signifies abnormal thinning of the inner retinal layers. The impression of retinal thinning caused by SCR can be found across multiple cross-sections. Some example images showing SCR in consecutive OCT scans can be found in our supplemental material. To confirm that a pit results from SCR, the experts try to locate its impression in the adjacent B-scans. By automating this process, we can reduce the potential for human error and increase the speed of diagnosis, ultimately improving patient outcomes. Hence, we present the Cross Scan Attention Transformer (CSAT). CSAT consists of a transformer-based pre-training network and an object detector to detect SCR from OCT scans. It provides a more efficient approach to locating and diagnosing SCR compared to manual inspection by ophthalmologists.

2 Related Work

Deep Learning in OCT Images: Deep learning techniques have been used to perform two main tasks on OCT images: area segmentation and disease detection. Area segmentation involves identifying and separating retinal layers [17, 12, 2], choroidal layers [13], and other areas [23] visible in OCT scans. Disease detection, on the other hand, focuses on specific retinal diseases such as age-related macular degeneration (AMD) [15, 11], diabetic macular edema (DME) [11, 18], glaucoma [22], and microcystic macular edema (MME) [16]. Retinal diseases such as DME and AMD are more widespread and therefore are common research topics. This leaves little attention towards obscure diseases such as SCR seen more commonly in the younger population and people of color [25]. Although Jing et al. [14] have provided a glimpse into the possibility of deep learning-based SCR detection, the lack of a targeted approach and full utilization of OCT volume leaves much to be desired. Our proposed method addresses this issue through the CSAT network.

Transformers for Object Detection: Our proposed method uses transformers for pre-training and object detection. The detection transformer (DETR) [5], one of the first transformers to perform object detection, achieved state-of-the-art results by predicting object class and location without object proposals. Models such as Deformable-DETR [28] and CF-DETR [4] modifies DETR to further improvise the results. Similarly, our proposed method also builds upon DETR by applying semi-supervised pre-training, and modifying the DETR decoder to recognize and share common features among multiple spatially linked images. Transformers have also been researched for medical image diagnosis and segmentation [24, 10, 7]. Along with OCT images, transformers have been used to analyze cross-sectional scans from several imaging modalities, such as computed tomography (CT) [24, 10, 7, 26]. Although there are multiple research studies involving cross-sectional images for medical diagnosis, the potential of using transformers to detect SCR from OCT images is yet to be explored.

Semi-supervised and unsupervised Pre-training: Semi-supervised and unsupervised pre-training techniques are commonly used to enhance the performance of deep learning networks when the available labeled data samples are limited. Several of these methods use contrastive learning such as SimCLR [6], for classification models and MICLe [1], for in medical image segmentation. However, these methods are designed for classification problems and are not preferred for object detection tasks. UP-DETR [8] is an unsupervised pre-training method that enhances the detection and segmentation performance for DETR. Our pre-training method is also based on the DETR model. However, unlike UP-DETR, our method specializes in pre-training a set of spatially related images (such as a set of cross-sections) instead of individual images.

3 Our Approach

The proposed method uses attention mechanisms to focus on the potential locations of SCR in the neighboring B-scans in two phases:

3.1 Phase 1: CSAT Pre-trainer

Phase 1 employs the CSAT pre-trainer, a contrastive learning framework to pre-train the OCT B-scans. The pre-trainer contains a Siamese network [3] with contrastive learning setting (as shown in Figure 1). This network is trained to discern whether two augmented B-scans originate from the same OCT volume of the same patient or from different OCTs of distinct patients. Through this approach, the model gains the ability to compare crucial retinal layer features, such as thickness, shape, and contour. The model takes two separate B-scans as input and runs them through a twin network composed of a feature extractor and a transformer encoder. The feature extractor consists of a ResNet50 backbone pre-trained on the ImageNet dataset. Similar to UP-DETR [8], we freeze the Resnet backbone during pre-training to trade off classification and localization preferences. The features obtained from this extractor are fed into a transformer encoder containing multi-headed self-attention layers. These attention layers focus on the input ResNet features and produce a K dimensional

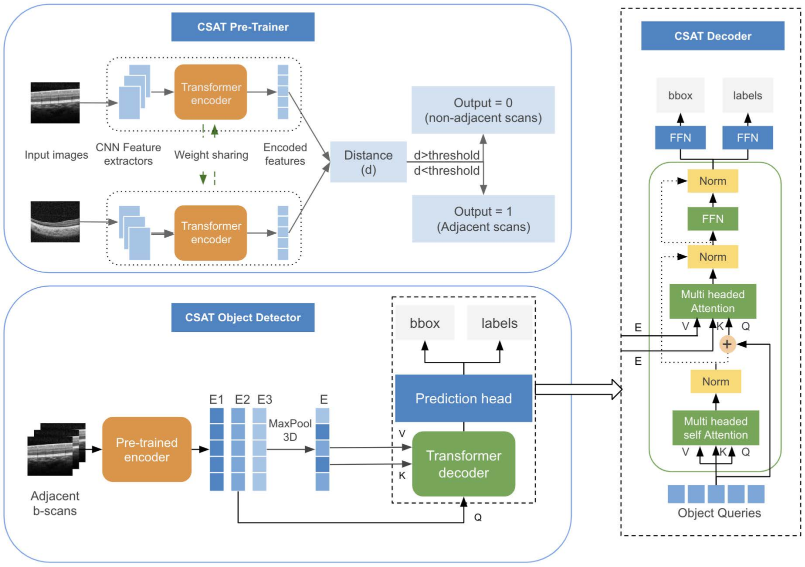

Fig. 1. The CSAT Architecture consists of CSAT pre-trainer (top-left) and CSAT detector (bottom-left). The pre-trainer implements a Siamese network with contrastive focal loss. The detector uses the embeddings from the CSAT pre-trainer as its encoder. The CSAT decoder (right) uses queries from E_m and keys and values from E.

embedding vector of shape (B, L, K). Here, L represents the encoder sequence length, and B represents the batch size. To obtain a shape of (B, K), the embeddings are averaged across L. We use Cosine Similarity (CS) as a distance metric for the loss function. We prefer CS to p-norm for two reasons: First, our problem statement is designed to determine the correlation between the two input vectors since they do not originate from the same image, even in positive examples. Second, CS gives normalized outputs. A contrastive focal loss function described in Equation 1 calculates the loss. This loss is then backpropagated through the network, ultimately causing the attention blocks to concentrate on the inner retinal layers' structure, which is more consistent across adjacent B-scans.

3.2 Phase 2: CSAT Object Detector

Phase 2 utilizes the CSAT object detection network, leveraging the pre-trained model from phase 1 as an encoder. The proposed network, as illustrated in Figure 1, is designed to attend to multiple adjacent B-scans simultaneously, enabling feature sharing to facilitate informed detection of SCR. This approach permits the model to leverage the volumetric nature of OCT, analyzing multiple cross-sections concurrently rather than individually. The model takes in a set of n adjacent B-scans, where n is an odd number between 1 and P, and P is the total number of B-scans in the OCT volume. The input tensor has a shape of (B, n, C, H, W), where B, C, H, and W represent the batch size, channels, height, and width of the input images, respectively. The encoder produces n embedding vector of K dimensions, namely $E_1, E_2, ..., E_n$, each having a shape of (B, K). A $3 \times 1 \times 1$ max pool layer is then applied to obtain a single K dimensional embedding, E. The unique aspect of our method is in the decoder structure. The decoder employs a set of N queries, keys, and values for self-attention. The queries for the encoder-decoder attention is extracted from E_m where $m = 1 + n/2$ whereas the keys and values are extracted from E. This enables the model to compare the features of E_m in the shared features, E. The resulting features are normalized and sent to the prediction heads, which generate N classes (in our case, $N = 2$, representing instances of fovea and SCR) and bounding-box predictions for each query. Each prediction is attributed to the B-scan linked to the embedding vector E_m. The remaining B-scans' predictions are made similarly by placing the targeted image at the center of each tensor per batch. There are three advantages to arranging embedding vectors in this way:

1. Neighboring image features are considered when making predictions.
2. By fine-tuning the CSAT encoder during detection, the weights are updated every time a B-scan is input as a neighbor of other B-scans, leading to faster convergence with fewer iterations.
3. The decoder uses queries from the central vector E_m to ensure that the nearest B-scans receive more attention than those farther apart in E.

3.3 Loss Functions

We trained our self-supervised network using contrastive focal loss [19] described in Equation 1. We applied a penalty of $\gamma = 3$ for incorrect classifications. To balance the number of positive and negative samples, we set the weighing factor, $\alpha = 0.15$, which is the approximate ratio of positive to negative samples in the pre-training dataset.

$$L_{CF} = -log(dist) * \alpha * (1 - dist)^{\gamma} \tag{1}$$

Where, $dist$ is the cosine similarity between the embeddings.

Similar to DETR [5], we used the minimum bipartite matching cost (Equation 2) to determine a one-on-one match between ground truth labels and predicted outputs. We used a weighted negative log-likelihood (NLL) for classification loss, mean square error (MSE) for bounding box regression, and a complete box-iou loss [27] for box predictions. The overall loss was obtained by taking weighted sum of the individual losses which are, 2, 3, and 5, respectively. These weights were obtained through multiple experiments. Similarly, we assigned weights of 0.05, 0.6, and 0.35 to the empty class, fovea, and SCR predictions in the NLL loss function.

Table 1. Confusion matrix for the CSAT pre-trainer. The values are obtained by averaging the results from 5 fold cross-validation.

		Ground Truth	
		True	**False**
Prediction	**Positive**	0.83 ± 0.03	0.19 ± 0.02
	Negative	0.17 ± 0.03	0.81 ± 0.04

$$M = argmin \sum_{i}^{N} C_{match}(y_i, y_i') \tag{2}$$

$$C_{match}(y_i, y_i') = \sum_{i}^{N} -p_i c_i + L_1(b_i, b_i') + L_{iou}(b_i, b_i') \tag{3}$$

Where, y_i and y_i' represent a pair of ground truth and predicted outputs, p_i is the class probability, c_i is the true class labels and b_i and b_i' are a pair of ground-truth and predicted bounding boxes, respectively.

4 Evaluation

4.1 Dataset

Our proposed method utilizes an internal dataset containing 594 OCT volumes from 147 participating patients with varying severity of SCD. The age of the patients on their first OCT test ranged from 4.44 to 20.39, with a mean age of 11.68±4.34 and a median age of 11.23. Among the patients, 70 were male, and 77 were female. The OCT scans were obtained using a Spectralis scanner from Heidelberg Engineering. The posterior pole volume scan involved a 30°×25°cuboid, with 31 raster lines producing 31 cross-sections per OCT. Both SCR and Fovea instances were manually annotated in the B-scans by experts. Detailed statistics on the dataset are presented as supplemental material.

4.2 Training and Experiments

In the pre-training network, we augmented the OCT images using random horizontal flips, color variance, blur, and brightness. The positive samples were generated by pairing B-scans with at most three steps between them, whereas the negative samples were generated by pairing B-scans from different patients. For example, an OCT volume containing B-scans in order: $1, 2, 3, ...n$ may have the following positive pairs: $(1, 2), (1, 3), ...(n-3, n)$. We calculated the confusion matrix of our evaluation by averaging the 5-fold cross-validation results. Table 1 illustrates the confusion matrix obtained from this experiment. We obtained an average precision of 81% and an average recall of 83%.

Table 2. Comparing the mean average precision (mAP) of three CSAT models: CSAT$_a$ (no pre-training), CSAT$_b$ (pre-training but no fine-tuning), and CSAT$_c$ (pre-training and fine-tuning) with YOLO [20], Faster RCNN [21], and DETR [5] for SCR and fovea detection. Bold and underlined numbers represent the highest and second-highest SCR detection results, respectively.

mAP		YOLO v8	Faster RCNN	DETR	CSAT a	CSAT b	CSAT c
@.5	SCR	0.78	0.78	0.81	0.80	0.83	**0.86**
	Fovea	0.85	0.84	0.86	0.85	0.89	0.91
@.5:.95	SCR	0.72	0.69	0.76	0.78	0.81	**0.83**
	Fovea	0.75	0.74	0.81	0.82	0.84	0.85

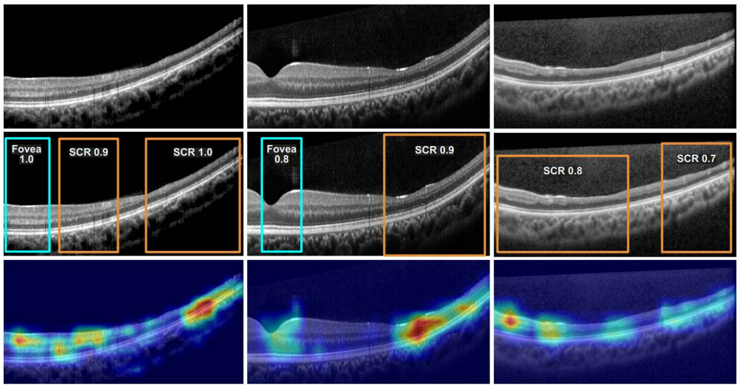

Fig. 2. SCR detection using CSAT. Top: Original B-scans. Middle: SCR and fovea instances detected by CSAT. Bottom: Attention plot of the CSAT detections.

We trained the CSAT detector to identify fovea and SCR instances in B-scans. We used the pre-trained embedding weights to train and test our CSAT detector. We used mean average precision (mAP) as our evaluation metric and compared the performances of CSAT to some state-of-the-art object detection methods such as YOLOv8 [20], Faster RCNN [21], and DETR [5], using our OCT dataset. These models were not pre-trained on the dataset. We trained and tested three different models of CSAT, namely CSAT a, b, and c, based on whether they were trained without pre-training, with pre-training but without fine-tuning, and with both pre-training and fine-tuning, respectively. The results depicted in Table 2 showed that our method outperformed other object detection networks in SCR detection. Among our three models, CSAT$_c$, which fine-tuned the pre-trained network, performed the best, followed by CSAT$_b$ and CSAT$_a$. A detailed description of CSAT implementation (pre-trainer and detector) can be found in our supplemental material.

4.3 Ablation Study

In Sub-section 3.2, we discussed the three advantages of the proposed architecture. This section will present the experiments we conducted to support our conclusions.

Our first experiment involved training the CSAT model with different values of n. The precision-recall curve in the Figure 3 (a) demonstrates the comparison between models trained with $n = 1$, $n = 3$, and $n = 5$. Our findings indicate that models with $n = 5$ and $n = 3$ outperformed the model with $n = 1$, which only used one B-scan at a time. This suggests that including neighboring B-scans in the training process improves detection results. However, increasing the value of n also increases the model complexity. As we can observe in the Figure 3 (a), the rate of improvement in detection performance decreases as the value of n increases. This indicates that although a higher n may enhance the detection performance, the B-scans closest to the image with E_m have more influence on the detection made for E_m.

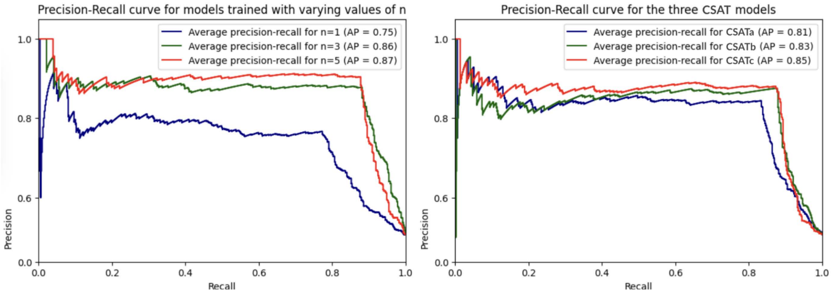

Fig. 3. Ablation results: (a): P-R curve for different values of n. (b): P-R curve for CSAT models - CSAT$_a$, CSAT$_b$, and CSAT$_c$

Our second experiment involved training and evaluating three CSAT models, CSAT a, b, and c. The Table 2 shows the mAP values for these models, while the Figure 3 (b) displays the precision-recall curves for each model. These results highlight the effectiveness of model c compared to b and a, indicating that pre-training and fine-tuning significantly enhance the detection performance.

5 Discussion

In the proposed work, the CSAT pre-trainer extracts the fundamental features from the B-scans to help the detector converge more quickly with less computation. It is trained to classify a pair of B-scans into two classes. Positive: the pairs are adjacent B-scans from the same patient and, negative: the B-scans belong to different patients (hence, they are not adjacent). Through this classification, we ensure that while learning to identify adjacent B-scans, the model focuses on the inherent features unique to these B-scans. By learning these features, the encoder provides the information that the detector needs i.e., the similar artifacts between neighboring B-scans. This allows the detector to utilize these common features to detect SCR with higher accuracy, as depicted by our results. To further verify this theory, we extensively visualized attention maps between the positive and negative pairs, like in Figure 2. We observed that the common features in the adjacent B-scans were highlighted in the case of positive pairs, whereas there were few to no highlights in the negative pairs.

6 Conclusion

Our proposed network, the Cross Scan Attention Transformer (CSAT), is a unique extension to the Detection Transformer (DETR) that primarily serves three purposes. Firstly, it detects subtle patterns within the retinal layers of OCT images using a transformer-based pre-training network. Secondly, it extracts and analyzes similar features from adjacent B-scans using pre-trained embeddings and attention mechanisms. Finally, it presents the first deep learning-based framework dedicated to sickle-cell retinopathy (SCR) detection from OCT images, outperforming several state-of-the-art object detection networks. Our research aims to improve the diagnosis and treatment of SCR, a condition affecting many individuals worldwide. With the introduction of the CSAT network, we hope to contribute to developing more accurate and reliable medical image diagnosis tools. In subsequent studies, we plan to expand the CSAT network to other areas of medical image analysis and imaging modalities.

Acknowledgments. The research in this paper was supported by a gift award from the Nemours Children's Hospital, Delaware.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Bibliography

[1] Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., Natarajan, V., Norouzi, M.: Big self-supervised models advance medical image classification. In: International Conference on Computer Vision (ICCV) (2021)

[2] Bhattarai, A., Kambhamettu, C., Jin, J.: Cu-net: Towards continuous multi-class contour detection for retinal layer segmentation in oct images. In: 2022 IEEE International Conference on Image Processing (2022)

[3] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. In: Advances in Neural Information Processing Systems. vol. 6. Morgan-Kaufmann (1993)

[4] Cao, X., Yuan, P., Feng, B., Niu, K.: Cf-detr: Coarse-to-fine transformers for end-to-end object detection. Proceedings of the AAAI Conference on Artificial Intelligence **36**(1) (2022)

[5] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. European Conference on Computer Vision (2020)

[6] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. ICML (2020)

[7] Chen, Y., Fan, D.P., Cheng, M.M., Zhang, T., Liu, J., Qian, C., Shen, J.: Transunet: Transformers make strong encoders for medical image segmentation. In: Computer Vision and Pattern Recognition (2021)

[8] Dai, Z., Cai, B., Lin, Y., Chen, J.: Up-detr: Unsupervised pre-training for object detection with transformers. Computer Vision and Pattern Recognition (CVPR) (2020)

[9] Fujimoto, J.G., Pitris, C., Boppart, S.A., Brezinski, M.E.: Optical coherence tomography: An emerging technology for biomedical imaging and optical biopsy. Neoplasia **2**(1) (2000)

[10] Han, C., Wang, J., Xu, J., Zhang, Z., Liu, C., Shi, Y., Zhang, Q.: After-unet: Axial fusion transformer unet for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2021)

[11] He, T., Chen, S., Xu, D., Wang, Y., Zhang, X., Chen, X., Chen, W.: Automatic detection of age-related macular degeneration based on deep learning and local outlier factor algorithm. Frontiers in Bioengineering and Biotechnology **8** (2020)

[12] He, Y., Carass, A., Liu, Y., Jedynak, B.M., Solomon, S.D., Saidha, S., Calabresi, P.A., Prince, J.L.: Structured layer surface segmentation for retina oct using fully convolutional regression networks. Medical Image Analysis (2021)

[13] Hsia, W.P., Tse, S.L., Chang, C.J., Huang, Y.L.: Automatic segmentation of choroid layer using deep learning on spectral domain optical coherence tomography. Applied Sciences **11**(12) (2021)

[14] Jin, J., Bhattarai, A., Miller, R., Kolb, E., Kambhamettu, C.: A deep learning system for sickle cell retinopathy detection using retinal oct images from children with sickle cell disease (2022)

[15] Kaymak, S., Serener, A.: Automated age-related macular degeneration and diabetic macular edema detection on oct images using deep learning. International Journal of Computer Assisted Radiology and Surgery **15**(5) (2020)

[16] Lang, A., Carass, A., Swingle, E.K., Al-Louzi, O., Bhargava, P., Saidha, S., Ying, H.S., Calabresi, P.A., Prince, J.L.: Automatic segmentation of microcystic macular edema in oct. Biomed. Opt. Express **6**(1) (Jan 2015)

[17] Li, Q., Li, S., He, Z., Guan, H., Chen, R., Xu, Y., Wang, T., Qi, S., Mei, J., Wang, W.: DeepRetina: Layer Segmentation of Retina in OCT Images Using Deep Learning. Translational Vision Science and Technology **9**(2) (2020)

[18] Li, X., Hu, X., Yu, L., Zhu, L., Fu, C.W., Heng, P.A.: Canet: Cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. IEEE Transactions on Medical Imaging **39** (2019)

[19] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection (2017)

[20] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection (2015)

[21] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks (2015)

[22] Soltanian-Zadeh, S., Kurokawa, K., Liu, Z., Zhang, F., Saeedi, O., Hammer, D.X., Miller, D.T., Farsiu, S.: Weakly supervised individual ganglion cell segmentation from adaptive optics oct images for glaucomatous damage assessment. Optica **8**(5) (May 2021)

[23] Soltanian-Zadeh, S., Liu, Z., Liu, Y., Lassoued, A., Cukras, C.A., Miller, D.T., Hammer, D.X., Farsiu, S.: Deep learning-enabled volumetric cone photoreceptor segmentation in adaptive optics optical coherence tomography images of normal and diseased eyes. Biomed. Opt. Express **14**(2) (Feb 2023)

[24] Wang, H., Cao, P., Wang, J., Zaiane, O.R.: Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. Proceedings of the AAAI Conference on Artificial Intelligence **36**(3) (Jun 2022)

[25] Wastnedge, E., Waters, D., Patel, S., Morrison, K., Goh, M., Adeloye, D., Rudan, I.: The global burden of sickle cell disease in children under five years of age: a systematic review and meta-analysis. Journal of Global Health **8** (Dec 2018)

[26] Zhao, J., Xiao, X., Li, D., Chong, J., Kassam, Z., Chen, B., Li, S.: mftrans-net: Quantitative measurement of hepatocellular carcinoma via multi-function transformer regression network. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021 (2021)

[27] Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-iou loss: Faster and better learning for bounding box regression (2019)

[28] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: 9th International Conference on Learning Representations, ICLR 2021 (2021)