Medical Image Analysis 93 (2024) 103104

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Anomaly guided segmentation: Introducing semantic context for lesion segmentation in retinal OCT using weak context supervision from anomaly detection

Philipp Seeböck [a,b,*], José Ignacio Orlando [a,c], Martin Michl [a], Julia Mai [a], Ursula Schmidt-Erfurth [a], Hrvoje Bogunović [a]

[a] Lab for Ophthalmic Image Analysis, Department of Ophthalmology and Optometry, Medical University of Vienna, Austria
[b] Computational Imaging Research Lab, Department of Biomedical Imaging and Image-Guided Therapy, Medical University of Vienna, Austria
[c] Yatiris Group at PLADEMA Institute, CONICET, Universidad Nacional del Centro de la Provincia de Buenos Aires, Gral. Pinto 399, Tandil, Buenos Aires, Argentina

ARTICLE INFO

Keywords:
Deep learning
Segmentation
Anomaly detection
Semantic context

ABSTRACT

Automated lesion detection in retinal optical coherence tomography (OCT) scans has shown promise for several clinical applications, including diagnosis, monitoring and guidance of treatment decisions. However, segmentation models still struggle to achieve the desired results for some complex lesions or datasets that commonly occur in real-world, e.g. due to variability of lesion phenotypes, image quality or disease appearance. While several techniques have been proposed to improve them, one line of research that has not yet been investigated is the incorporation of additional semantic context through the application of anomaly detection models. In this study we experimentally show that incorporating weak anomaly labels to standard segmentation models consistently improves lesion segmentation results. This can be done relatively easy by detecting anomalies with a separate model and then adding these output masks as an extra class for training the segmentation model. This provides additional semantic context without requiring extra manual labels. We empirically validated this strategy using two in-house and two publicly available retinal OCT datasets for multiple lesion targets, demonstrating the potential of this generic anomaly guided segmentation approach to be used as an extra tool for improving lesion detection models.

1. Introduction

Optical Coherence Tomography (OCT) scans constitute a gold standard imaging technique in ophthalmology, providing a high-resolution 3D visualization of the retina. It is non-invasive, acquires images in only a few seconds and is therefore one of the most important diagnostic modalities in the context of retinal diseases. 2D cross-sectional slices (also known as B-scans) of 3D OCT volumes are typically qualitatively evaluated by physicians to conduct diagnosis, determine treatments or to infer other clinical decisions (Fujimoto and Swanson, 2016).

Age-related macular degeneration (AMD) is the most common cause of legal blindness in people over age 65 in developed countries (Wong et al., 2014). Its prevalence is steadily increasing due to the aging population, with 288 million people expected to be affected in 2040 (Wong et al., 2014). AMD is a multifactorial disease, leading to several pathological changes in the eye that can be visually assessed through OCT. These include intraretinal cystoid fluid (IRC), subretinal fluid (SRF), pigment epithelial detachment (PED), drusen, pseudodrusen or subretinal drusenoid deposits (SDD), geographic atrophy, subretinal hyperreflective material (SHRM) or hyperreflective foci (HRF) (Schmidt-Erfurth et al., 2017; Liefers et al., 2021; Cao et al., 2021) (Fig. 1, Fig. 3, Fig. 4, Fig. 5). Ophthalmologists usually base their treatment decision on the qualitative assessment of these features at multiple follow-up visits-a task that is prone to subjectivity, errors and biases (Michl et al., 2022). Segmenting all the lesions in a given image might help in defining disease stages more accurately and objectively. However, when done manually, this task is extremely tedious, time consuming and therefore unfeasible in clinical practice. Automated methods, on the other hand, can alleviate this task and allow decisions that are less subjective, more reproducible and individually optimized (Bogunović et al., 2019; Müller et al., 2021; Michl et al., 2022). Accurate automated segmentations might enable large-scale analyses for the discovery of biomarkers, as well as the investigation

* Corresponding author.
E-mail addresses: philipp.seeboeck@meduniwien.ac.at (P. Seeböck), hrvoje.bogunovic@meduniwien.ac.at (H. Bogunović).

https://doi.org/10.1016/j.media.2024.103104
Received 12 October 2022; Received in revised form 1 December 2023; Accepted 5 February 2024
Available online 8 February 2024

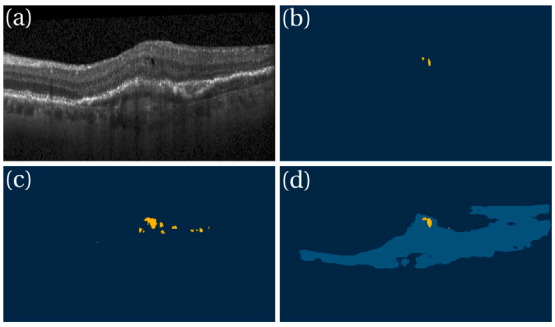

Fig. 1. Anomaly guided segmentation in retinal OCT. (a) Cross-sectional OCT scan, (b) manual annotation of intraretinal cystoid fluid (IRC), (c) prediction of the baseline and (d) segmentation output of the proposed ANGUS-Net approach.

of disease dynamics, individualized risks or new standardized and objective treatment criteria (Schmidt-Erfurth et al., 2021). This might further help in understanding the pathogenesis of complex diseases such as AMD (Schmidt-Erfurth and Waldstein, 2016).

Current existing models for automated lesion segmentation in OCT are mostly based on fully convolutional neural networks, which have shown excellent performances in several image related applications (Litjens et al., 2017). For some complex lesions or real-world datasets, however, existing models are still struggling to get equally good results, e.g. due to the high variability in disease appearance, lesion phenotypes or image quality (Bogunović et al., 2019). Therefore, multiple techniques have been investigated to further improve lesion segmentation performance (Section 1.1).

In this paper, we propose a novel anomaly guided approach for lesion segmentation in retinal (OCT) scans. Anomaly detection is used to provide additional semantic context for the algorithms. In particular, applying an anomaly detection model to the training dataset results in weak anomaly segmentation maps of these images. Subsequently, the manual annotations of the target of interest are merged with these weak predictions, yielding a combined target map. The segmentation model is then trained on this merged map to simultaneously segment the target lesion as well as the anomalous area as an additional target (Fig. 1). Here, we define anomaly detection as the task of capturing deviations from a normal appearance, using only a set of healthy patients without manual labels for training. This idea is appealing for multiple reasons. First, it not only adds meaningful weak supervision to the segmentation model, but also mimics the way physicians analyze medical images: they first screen the image to identify deviations from the normal morphology (i.e. anomalies), and subsequently analyze these regions more closely. Second, as the training of the anomaly detection model does not require manual annotations, this extra weak supervision provides semantic context without additional labeling cost. Third, since the training is independent of disease specific assumptions and only involves healthy samples, its applicability is by definition not limited to a specific lesion type or patient cohort.

1.1. Related work

Current approaches for automated lesion segmentation in retinal OCT images primarily follow a U-net based encoder–decoder architecture that must be trained with large-scale pixel-wise annotated datasets to achieve good results (Ronneberger et al., 2015; Roy et al., 2017; Schlegl et al., 2018; De Fauw et al., 2018; Bogunović et al., 2019; Pekala et al., 2019; Lu et al., 2019; Liefers et al., 2021). The majority of these models operate in 2D using B-scans as inputs (Roy et al., 2017; Bogunović et al., 2019; Gu et al., 2019; Pekala et al., 2019; Lu et al., 2019), with some exceptions that make use of three-dimensional information by working on volume data from multiple cross-sectional scans (De Fauw et al., 2018; Liefers et al., 2021). While 3D data allows the networks to take more spatial context into account, it also restricts the models to be applied only to images coherent with the imaging acquisition protocol used in the training data (that is, with a specific anatomical inter-slice distance and a fixed number of B-scans). Additionally, retinal OCT segmentation methods seem to be mainly focused towards retinal layers (Apostolopoulos et al., 2017; Pekala et al., 2019; Orlando et al., 2019; Gu et al., 2019; He et al., 2021b) or fluid segmentation (Roy et al., 2017; Schlegl et al., 2018; Lu et al., 2019; Bogunović et al., 2019; Ye et al., 2021; Xing et al., 2022; He et al., 2022).

Multiple efforts have been made to improve the performance of segmentation models. These efforts are mostly focusing on enlarging their receptive field (Apostolopoulos et al., 2017; Gu et al., 2019), combining information of multiple scales (Ding et al., 2021; Meng et al., 2021; Xing et al., 2022), applying attention mechanisms (Schlemper et al., 2019; Sinha and Dolz, 2020; Ye et al., 2021), utilizing transformer-based architectures (Chen et al., 2021; Zhang et al., 2021a; Cao et al., 2022), post-processing segmentation prediction outputs (Bai et al., 2017; Lu et al., 2019; Pekala et al., 2019), exploiting uncertainty estimations (Orlando et al., 2019; Nair et al., 2020), introducing boundary information about the targets (Karimi and Salcudean, 2019; Kervadec et al., 2019; He et al., 2021a; Xing et al., 2022), leveraging multi-task settings (Chen et al., 2018; Playout et al., 2019; He et al., 2020, 2021a; Zhang et al., 2021b), and incorporating either auto-context approaches (Montuoro et al., 2017; Venhuizen et al., 2018; Zhou et al., 2019) or prior information about the target shape (He et al., 2021b; Xing et al., 2022). Examples of methods belonging to each of these categories are briefly covered in the sequel.

Apostolopoulos et al. (2017) and Gu et al. (2019) introduced dilated convolutional layers in the encoder–decoder structure to enlarge the receptive field of the model and therefore provide additional context to the network. Similarly, multi-scale approaches help to combine both local and non-local segmentation information to better cope with variations of lesion size. Meng et al. (2021), for instance, introduce a multi-scale module for choroidal neovascularization (CNV) lesion segmentation in OCT scans. While these approaches indeed extend the number of potential pixels that the model takes into account for prediction, it does not explicitly incorporate semantic context. Spatial pyramid pooling modules are combined with attention mechanisms in Xing et al. (2022) to improve segmentation of pathological retinal fluid. The objective of attention modules is to learn potentially irrelevant areas of an input image and put the focus of the model on the relevant features for a specific task. Segmentation networks typically use higher-level features to guide the self-attention mechanism of the actual feature map (Schlemper et al., 2019; Sinha and Dolz, 2020). Conceptually, attention mechanisms are therefore related to multi-scale approaches in the sense that both aim at improving performance by fusing information from different scales. Related to attention, transformer-based architectures for segmentation have been recently proposed (Chen et al., 2021; Zhang et al., 2021a; Cao et al., 2022). For instance, both 'TransUNet' (Chen et al., 2021) and 'Swin-Unet' (Cao et al., 2022) are hybrid architectures combining transformers and U-Net, aiming to combine the extraction of global context with high-resolution convolutional feature maps to enable precise localization. Instead, post-processing techniques directly work on the output of the network and seek to improve the results by applying heuristics or assumptions with respect to the masks themselves. Examples for this are the application of conditional random fields that make use of the spatial correlation of neighboring pixels (as in Bai et al. (2017)); the Gaussian processes regression technique used in Pekala et al. (2019) for segmentation of retinal layers; or the usage of hand-crafted features in combination with random forests in Lu et al. (2019) to reduce the number of false positive fluid pixels. Estimating the pixel-wise uncertainty for segmentation masks has also shown to improve results. Orlando et al. (2019) utilized the Monte-Carlo dropout technique to estimate the epistemic uncertainty of the neural network for photoreceptor layer

segmentation, while Nair et al. (2020) applied the same technique for multiple sclerosis lesion segmentation on a 3D U-Net. Other approaches seek to improve the performance by focusing on the boundary region of the segmentation target of interest. This can be done by introducing an explicit boundary loss, optimizing a distance metric between ground truth and predicted shape instead of solely relying on region based losses (Karimi and Salcudean, 2019; Kervadec et al., 2019; He et al., 2021a; Xing et al., 2022). Besides representing the target itself in a different way to the network, additional supervision from another task can help the generalization performance of the model, as has been shown for skin lesion segmentation (Chen et al., 2018), segmentation of lesions in color fundus images (Playout et al., 2019) or organ segmentation in CT scans (He et al., 2020). The underlying idea is that the features learned for one task can benefit the learning of others, e.g. predicting the overall disease class of an image helps the segmentation of particular lesions. Even though this multi-task learning setting can be helpful, it is not clear which tasks are positively correlated and which tasks, when trained together, negatively impact the performance (Fifty et al., 2021). Another set of methods utilize auto-context where a sequence of models is applied, taking the output of the previous stage as additional input to enable a step-by-step refinement of predicted segmentation maps. This strategy has been utilized in OCT for retinal layer and fluid segmentation (Montuoro et al., 2017; Venhuizen et al., 2018) or natural images (Zhou et al., 2019). This strategy can be seen as providing additional context to a later stage in the sequence, with the context being restricted to the classes which are available in the ground truth annotations. Prior knowledge about the anatomy or specific shape characteristics of lesions can also be encoded as topological constraints into the network. He et al. (2021b) introduce a topology preserving module to ensure an anatomical correct ordering of predicted retinal layers. A diffeomorphic framework is utilized in Wyburd et al. (2021) to ensure preserved topology in myocardium segmentation, by learning to properly warp a binary prior map. However, incorporating prior knowledge is well applicable for regular, relatively well-defined shapes, but cannot be easily extended to segmentation of lesions exhibiting a large variation in shape, location, size or number of connected components.

In contrast, a direction of research that has not yet been explored regarding performance improvement of (lesion) segmentation, is the usage of anomaly detection methods to provide additional semantic context. Several powerful approaches for anomaly detection have already been proposed for retinal OCT scans that can be directly utilized, including methods based on shape models (Dufour et al., 2012), Gaussian Mixture Models (GMMs) (Sidibé et al., 2017), auto-encoders combined with one-class support vector machines (SVMs) (Seeböck et al., 2019b), generative adversarial networks (GANs) (Schlegl et al., 2017, 2019; Zhou et al., 2020), CycleGANs (Wang et al., 2021), normalizing flows (Zhao et al., 2022) or Bayesian deep learning with epistemic uncertainty estimates (Seeböck et al., 2019a). In general, unsupervised anomaly detection is an active growing field of research, proposing methods for varying medical imaging domains beyond OCT such as brain imaging (Sato et al., 2018; Pawlowski et al., 2018; Wyatt et al., 2022; Bercea et al., 2023), chest radiographs (Wolleb et al., 2020; Zhang et al., 2020; Nakao et al., 2021), fundus photography (Ouardini et al., 2019) or breast imaging (Quellec et al., 2016; Wei et al., 2018; Burger et al., 2023). In this work we use the anomaly detection model based on epistemic uncertainty estimates (Seeböck et al., 2019a), which to the best of our knowledge represents the state-of-the-art for pixel-wise anomaly detection in retinal OCT imaging.

1.2. Contribution

In this work, we introduce a novel generic concept for improving lesion segmentation models. To the best of our knowledge, this is the first work that proposes to use the output of an anomaly detection model as additional weak supervision for the segmentation algorithm, to provide it with supplementary semantic context. We evaluated our approach using two different in-house and two publicly available OCT image datasets, providing results on clinical-trial, public challenge and real-world data. Moreover, we conducted our evaluation for segmenting six different lesion targets in total, covering the main lesion types of AMD. Results demonstrate a consistent performance improvement of the proposed approach across all datasets and targets. We also provide an extensive evaluation, showing that this improvement holds valid under different architectural backbones, sizes of the training set, single- as well as multi-class settings, external test sets, unseen diseases and lesion-wise detection performance. This demonstrates the robustness of the proposed approach as well as its improved data-efficiency.

2. Method

A schematic overview of the proposed *Anomaly Guided Segmentation (ANGUS)* approach is provided in Fig. 2. First, an anomaly detection model is applied to the training set images in order to create weak anomaly segmentation maps. Secondly, these maps are merged with the corresponding manual ground truth annotations of the target of interest, creating a target map with an additional anomaly class. Finally, these combined maps are used to train the segmentation model to simultaneously segment the target structure and the anomalous area in a multi-class segmentation setting. This transforms the M-class into a $M+1$ class segmentation problem, providing supplementary context and supervision through the extra class without the need of additional manual labels.

2.1. Anomaly guided segmentation: Using anomaly detection for weak supervision

Let $X \in \mathbb{R}^{a \times b}$ be a set of 2D training images with $a \times b$ pixels each and $Y \in \mathcal{Y}^{a \times b}$ a corresponding set of manual ground truth labels of the target of interest. Moreover, let f_{AN} be an anomaly detection model that is able to predict a binary pixel-wise anomaly segmentation map for this image domain. Formally, applying the model f_{AN} to an input image x produces a corresponding binary prediction map of the same size as the input: $f_{\text{AN}}(x) = y_{\text{AN}}$. After applying the model f_{AN} to all images x of the training set X, the resulting anomaly maps y_{AN} can be used as additional weak supervision in the training process of the segmentation model.

Here we propose to merge the automatically created anomaly predictions with the manual annotations of the target of interest y, resulting in a combined target map y_{COM}:

$$y_{\text{COM}} = [y, \mathbb{1}_{\text{AN}}] \tag{1}$$

where y is the one-hot encoded representation of the original M target classes, $[\cdot]$ represents a tensor concatenation in the class dimension and $\mathbb{1}_{\text{AN}}$ is the extra pixel-wise binary map for the anomaly predictions, such that:

$$\mathbb{1}_{\text{AN}}(y_{\text{AN}_{ij}}) = \begin{cases} 1, & \text{if } y_{ij} = 0 \text{ and } y_{\text{AN}_{ij}} = 1 \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where i and j are the pixel coordinates of the maps. This definition allows the model to predict the target lesions and also an additional class for anomalies, without affecting the original ground truth labeling of the targets. In other words, the M class problem is transformed into a $M+1$ class segmentation task using the automatically created weak anomaly prediction maps as additional class, as illustrated in Fig. 2. The segmentation model f_{SEG} is then trained using these combined target maps and a weighted Cross-Entropy loss function L_{CE}, so that it learns to segment both the targets of interest and the anomalous area:

$$L_{\text{CE}} = \sum_{c=0}^{M+1} w_c \cdot y_{\text{COM}_c} \cdot \log\left(f_{\text{SEG}}(x)_c\right) \tag{3}$$

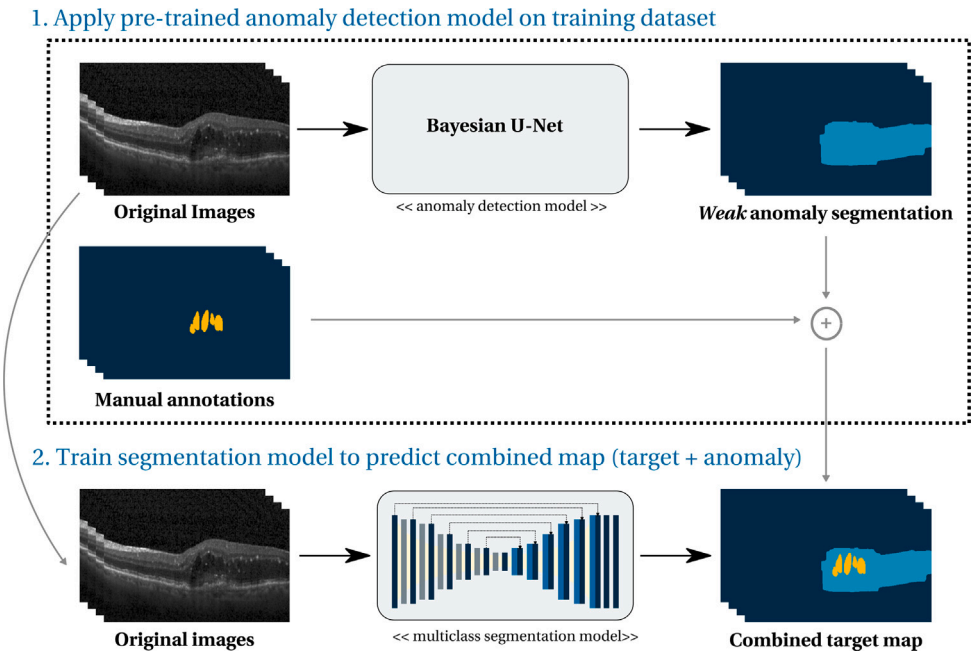

Fig. 2. Overview of our *Anomaly Guided Segmentation (ANGUS)* approach. First, a pre-trained anomaly detection model is applied to segment the OCT training dataset. Second, the predicted anomaly maps are merged with the manual annotations of the target of interest. Finally, this combined target map is used to train a multiclass segmentation model.

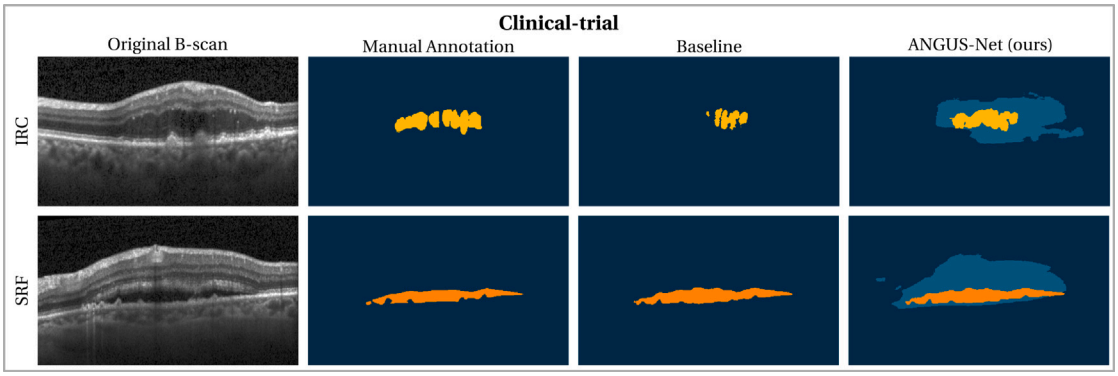

Fig. 3. Qualitative results in the *'clinical-trial'* test set, showing the original B-scan, manual annotation, prediction of the baseline model and prediction of the proposed ANGUS-Net for the lesion targets intraretinal cystoid fluid (IRC) and subretinal fluid (SRF).

$$= \sum_{c=0}^{M+1} w_c \cdot y_{\mathrm{COM}_c} \cdot \log(\hat{y}_{\mathrm{COM}_c}), \tag{4}$$

with w_c denoting the weighting hyperparameter for class c and $f_{\mathrm{SEG}}(x)$ is the probability of each class label predicted by model f_{SEG} for an input image sample x.

2.2. Lesion segmentation in retinal OCT images

The strategy presented in Section 2.1 is applied for segmenting retinal lesions in OCT scans. Section 2.2.1 covers the details of the model used for anomaly detection, while Section 2.2.2 describes the architecture of the model.

2.2.1. Anomaly detection model

To generate the weak anomaly segmentation maps in retinal OCT scans, we make use of the recently published anomaly detection model *WeakAnD* (Seeböck et al., 2019a). This approach can be trained on a dataset of healthy samples without requiring manual annotations. Instead, it utilizes weak labels of the healthy anatomy produced by a graph-based surface segmentation algorithm (Garvin et al., 2009), to train a Bayesian U-Net that learns the healthy anatomical structure of the retina. More specifically, this network is trained to segment 10 retinal layers in healthy subjects, using the automatically generated weak labels as ground truth. During inference, epistemic uncertainty estimates of the segmentation output are obtained by Monte Carlo dropout (Gal and Ghahramani, 2015), i.e. computing the pixel-wise variance across 50 layer segmentations obtained with active dropout during inference. These epistemic uncertainty estimates correlate with deviations from normal layer appearance and result in a first estimate of anomalous areas by applying a threshold on these uncertainty maps. The final pixel-wise anomaly maps are obtained after applying a post processing technique, majority-ray-casting, which extracts blob-shaped segmentations of the anomalous areas.

2.2.2. Architecture of the segmentation model

Following the majority of approaches for automated lesion segmentation in retinal OCTs, we use a standard U-Net based encoder–decoder structure with residual connections as architectural backbone. It comprises seven levels of depth, where in each level the output of the encoder block is connected with the corresponding decoder part using

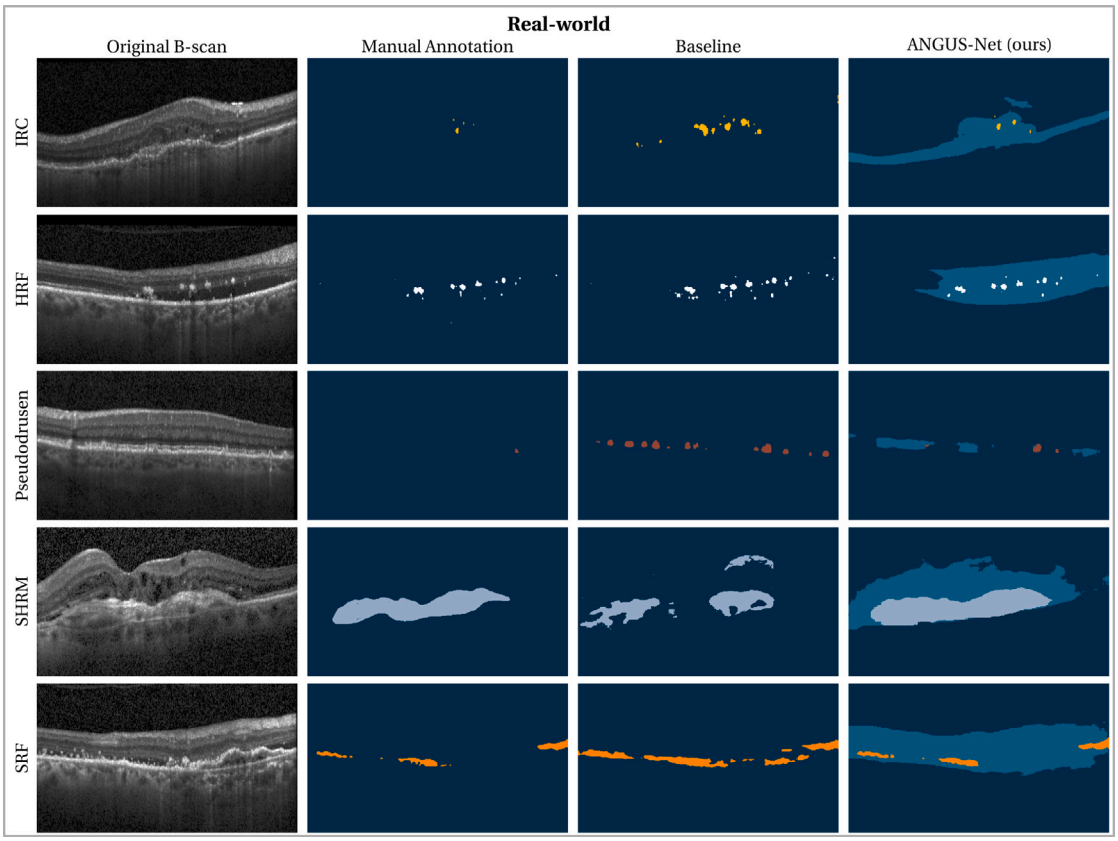

Fig. 4. Qualitative results in the *'real-world'* test set, showing the original B-scan, manual annotation, prediction of the baseline model and prediction of the proposed ANGUS-Net for the lesion targets intraretinal cystoid fluid (IRC), hyperreflective foci (HRF), pseudodrusen, subretinal hyperreflective material (SHRM) and subretinal fluid (SRF).

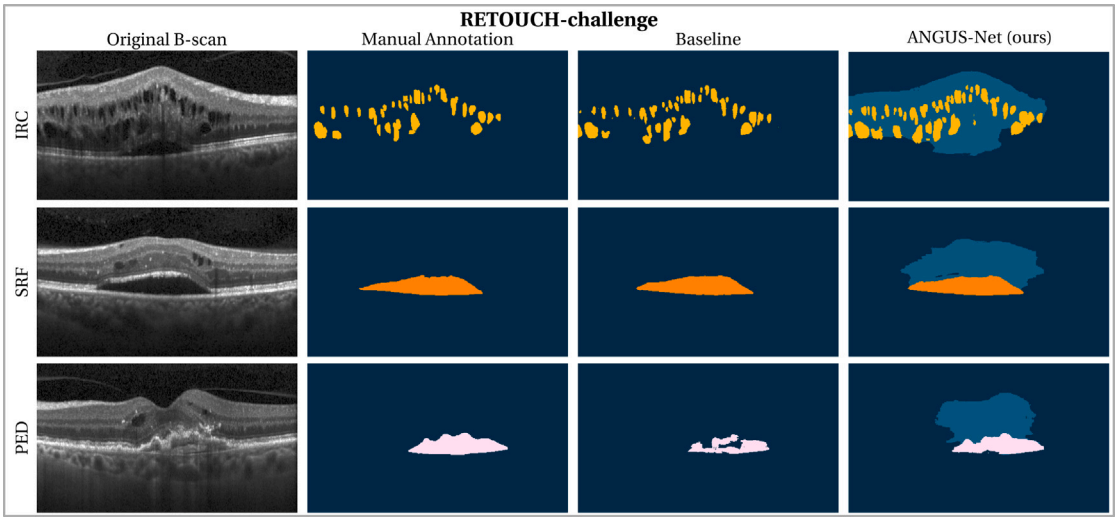

Fig. 5. Qualitative results in the *'RETOUCH challenge'* test set, showing the original B-scan, manual annotation, prediction of the baseline model and prediction of the proposed ANGUS-Net for the lesion targets intraretinal cystoid fluid (IRC), subretinal fluid (SRF) and pigment epithelial detachment (PED).

a skip connection. The number of output channels goes from 64 in the first to 2,048 in the bottleneck layer, using powers of 2. While 2×2 max-pooling is used for downsampling in the encoding path, nearest neighbor upsampling is utilized in the decoder to restore the resolution of the activation maps. Residual convolutional blocks consist of two convolutional layers with 3×3 filters, each followed by a batch normalization layer and a rectified linear unit (ReLU). The residual connection within these blocks links the input with the output of the second batch normalization layer.

3. Experimental setup

We assessed (1) the performance of the proposed method on four different OCT datasets with six individual segmentation targets, (2) the robustness with respect to the used neural network backbone, (3) the influence of the amount of training data, (4) its effectiveness in a multi-target setting, (5) the performance in terms of lesion-wise detection, (6) the impact of the weighting hyperparameter w_c, (7) the generalization performance on an external test set, (8) the performance

in cross-disease settings and (9) the correlation between the size of the anomalous area and the performance.

3.1. Materials

We used two separate in-house datasets and one publicly available dataset (Bogunović et al., 2019) for training and evaluating our model. Moreover, we used one additional publicly available dataset (Kermany et al., 2018) solely for external evaluation, without applying any re-training procedure. The first dataset, *'clinical-trial'* is composed of 66 macula centered Spectralis (Heidelberg Engineering, GER) OCT scans, with 512 × 496 × 49 voxels per volume, covering a retinal area of approximately 6 mm × 2 mm × 6 mm, resulting in an approximate anatomical voxel size of 12μm, 4μm and 122μm, respectively. 59 volumes were acquired from clinical trial patients suffering from neovascular AMD, whereas 7 volumes correspond to patients with retinal vein occlusion (RVO). The dataset was randomly split on a patient-distinct basis and stratified by disease into train, validation and test set, with 46, 6 and 14 volumes, respectively. Pixel-wise manual annotations of IRC and SRF were created by a retina specialists.

The second dataset, *'real-world'* comprises 60 macula centered Spectralis OCT scans, with resolutions ranging from 512 × 496 × 18 voxels to 512 × 496 × 97 voxels each. These images were acquired during clinical routine check-ups in neovascular AMD patients who underwent potential treatment with anti-vascular endothelial growth factor (anti-VEGF) injections. Therefore, this set is representing a less controlled and more realistic scenario. The set was randomly split on a patient-distinct basis into train, validation and test set, using 37, 10 and 13 volumes on each subset, respectively. Five different pathological structures were manually annotated by a retina specialist and include IRC, SRF, pseudodrusen, SHRM and HRF.

The third dataset, *'RETOUCH challenge'*, consists of the publicly available Retinal OCT Fluid Challenge (RETOUCH) dataset with pixel-wise manual annotations of IRC, SRF and PED (Bogunović et al., 2019). For simplicity and not to exclude any pixels from the evaluation, the reference annotation from only one center (MUV) was used. We utilized all 38 macula centered Spectralis OCT scans with 512 × 496 × 49 voxels per volume, covering a retinal area of 6mm × 2mm × 6mm. We utilized the predefined split of 24 training and 14 test volumes. We further randomly split the training set into train and validation set, with 18 and 6 volumes, respectively.

The external evaluation set *'OCT-Kermany'* consists of a subset of the publicly available Retinal OCT dataset (Kermany et al., 2018) with pixel-wise manual annotations of IRC, SRF and PED. In particular, we used the DME' and 'DRUSEN' groups of the original test set, each comprising 250 2D scans. We excluded 24 (DME) and 13 (DRUSEN) samples with invalid manual lesion annotations (e.g. non-closed contours) resulting in a final dataset *'OCT-Kermany'* consisting of 463 scans.

3.2. Experiments

For the three datasets *'clinical-trial'*, *'real-world'* and *'RETOUCH challenge'*, we trained and evaluated the proposed model for each target separately. This means that we obtained two final models in the *'clinical-trial'* dataset, one for IRC and one for SRF, five final models in the *'real-world'* dataset for IRC, SRF, pseudodrusen, SHRM and HRF, and three final models in the *'RETOUCH challenge'* for IRC, SRF and PED, respectively. A binary segmentation model without additional weak anomaly supervision was trained for each target as a comparison baseline (*'U-Net'*), using the same settings and architectural backbone to ensure a fair comparison.

To evaluate the robustness of the proposed approach with respect to (A) the backbone structure, (B) the number of training samples, (C) the number of targets trained on simultaneously, (D) the segmentation performance on a lesion-wise basis, (E) the influence of the weighting hyperparameter w_c, (F) the performance on external data as well as unseen diseases and (G) the influence that the size of the anomalous area has on the final performance we performed multiple evaluation experiments as follows:

- **A:** We used four alternative segmentation model backbones to evaluate the robustness of the approach regarding the architecture. First, we conducted experiments using a backbone structure with fewer parameters, a smaller receptive field and a lower number of layers in total, consisting of a U-Net with 6 levels of depth (*'U-Net-6', 'ANGUS-Net-6'*). In particular, this architecture followed the structure of the architecture described in Section 2.2.2, except that only 6 levels of depth are used, with the number of output channels going from 64 in the first to 1,024 in the bottleneck layer, using powers of 2 (64-128-256-512-1024-1024). Second, we performed experiments using 'DeepLabv3' (Chen et al., 2017) as an alternative architectural backbone (*'DeepLabv3', 'ANGUS-DeepLabv3'*). 'DeepLabv3' is a widely used state-of-the-art segmentation network with significant differences with respect to the U-Net, including depth-wise separable convolutions and atrous spatial pyramid pooling (ASPP). Third, we used the Attention U-Net (Schlemper et al., 2019) as an alternative backbone (*'AttUNet', 'ANGUS-AttUNet'*). As forth alternative, we conducted experiments using the TransUNet (Chen et al., 2021), a model combining Transformers and U-Net for segmentation (*'TransUNet', 'ANGUS-TransUNet'*)). In contrast to the other architectures, the latter two alternative segmentation model backbones explicitly integrate attention mechanisms into their architecture.
- **B:** We conducted evaluation experiments of the models using a reduced number of training samples. Subsets were randomly created with 25%, 50% and 75% of the total number of patients for the two in-house training datasets.
- **C:** We trained an additional model to predict all targets simultaneously in order to assess the effect of the number of target classes. In particular, a baseline model was trained on the *'real-world'* dataset to predict IRC, SRF, pseudodrusen, SHRM and HRF simultaneously, and the proposed approach was trained using the weak anomaly class in addition (*'U-Net-Multi'*, *'ANGUS-Net-Multi'*). Furthermore, we performed the same experiment using the alternative architectures from evaluation experiment *A* (U-Net with 6 levels of depth: *'U-Net-6-Multi'*, *'ANGUS-Net-6-Multi'*, DeepLabv3: *'DeepLabv3-Multi'*, *'ANGUS-DeepLabv3-Multi'*, Attention U-Net: *'AttUNet-Multi', 'ANGUS-AttUNet-Multi'*, and TransUNet: *'TransUNet-Multi', 'ANGUS-TransUNet-Multi'*).
- **D:** In order to investigate the performance of the models on a lesion-wise basis, we computed lesion-detection Recall (LD-Re$_d$) and lesion-detection Precision (LD-Pr$_d$) as proposed in Seeböck et al. (2019a). For each individual lesion defined as a connected component, the Dice with respect to the manual annotation is computed. The amount of true positives is then counted as the number of lesions with a Dice higher than a threshold d and used to calculate LD-Re$_d$ and LD-Pr$_d$. This means that annotated lesions that are missed are counted as false negatives, and predicted lesions without overlap are counted as false positives. By varying the threshold d in the range $[0, 1]$, both LD-Re and LD-Pr curves can be computed. We provide plots of these curves as well as quantitative scores in terms of area under the curve (AUC) to allow a more comprehensive evaluation.
- **E:** We analyzed the influence of the weighting hyperparameter w_c, by evaluating the models *'ANGUS-Net-Multi'*, *'ANGUS-Net-6-Multi'*, *'ANGUS-DeepLabv3-Multi'*, *'ANGUS-AttUNet-Multi'* and *'ANGUS-TransUNet-Multi'* on the 'real-world' dataset with varying weights for the anomaly-class ($w_c = [0.01, 0.1, 1, 2, 5, 10, 50]$), while all other class weights remained constant.

- **F:** We investigated the external generalization performance of the proposed approach by (F.1) directly applying the trained models of experiment *A* on an external publicly available dataset (Kermany et al., 2018) *'OCT-Kermany'* and (F.2) conducting cross-disease experiments using the RETOUCH dataset. For the cross-disease experiment *F.2* we divided the *'RETOUCH challenge'* dataset into two subsets by disease: *'RETOUCH challenge AMD'* (9 training, 3 validation, 7 test volumes), consisting only of AMD patients, and *'RETOUCH challenge RVO'* (9 training, 3 validation, 7 test volumes) comprising only RVO patients. We trained both *'U-Net'* and *'ANGUS-Net'* on one subset and evaluated them on the test set of the other subset respectively. We quantified performance only for IRC and SRF classes, as the OCT volumes of the RVO patients did not contain any PED lesions.
- **G:** To evaluate the influence that the size of the anomalous area has on the final performance, we computed the Spearman's correlation coefficient (Zar, 2005) between Dice values and the set difference between the anomalous area and the target lesion, namely the 'residual anomalous area'. We performed this comparison for all ANGUS models, on all datasets and targets. In Fig. 1(d) this area is represented in light blue.

We used precision, recall and Dice to evaluate the performance of the segmentation models. Paired Wilcoxon signed-rank tests with $\alpha = 0.05$ were used to test for statistical significance in the differences of the evaluation metrics. In the Supplementary Material, we provide details of an additional experiment using both *WeakAnD* (Seeböck et al., 2019a) and the recently proposed diffusion based anomaly detection method *AnoDDPM* (Wyatt et al., 2022) as unsupervised baselines for the task of lesion segmentation (Supplemental Experiment A). Moreover, we also compare both approaches in terms of pixel-level anomaly detection performance in OCT (Supplemental Experiment B).

3.3. Training details

As a pre-processing step, the intensity values of each individual B-scan were rescaled between 0 and 1.

We did not perform any training or optimization of the anomaly detection approach. Instead we utilized the final trained model *WeakAnD (with layer-flattening)* from Seeböck et al. (2019a) in an 'out-of-the-box' setting, using the standard parameters reported in the original paper, namely Monte-Carlo dropout sampling (50 samples), post-processing ($s = 10$, $m_c = 4$, $m_o = 2$), two iterations of majority-ray-casting ($v^{(1)} = 3$, $v^{(2)} = 4$) and a threshold of $t = 0.10$.

For training the segmentation model we used Adam optimization (Kingma and Ba, 2014) with standard parameters $\beta_1 = 0.90$, $\beta_2 = 0.999$, $\epsilon = 1e^{-8}$, weight decay $\lambda = 1e^{-5}$ and a learning rate of $1e^{-4}$. The parameters of the network were initialized using He initialization (He et al., 2015). The weighting parameters w_c were chosen empirically. In the *'clinical-trial'* dataset an equal weighting of $w_0 = w_1 = w_2 = 1$ was chosen for all classes, except for *'ANGUS-Net-6'*, *'ANGUS-DeepLabv3'*, *'ANGUS-AttUNet'* and *'ANGUS-TransUNet'* where $w_1 = 2$ was used for the IRC target class. In the 'real-world' dataset $w_0 = 0.1$ was chosen for the background, $w_1 = 10$ for all targets except SHRM ($w_1 = 15$), and $w_2 = 2$ for the anomaly class except for the SHRM models with $w_1 = 1$. In the 'RETOUCH challenge' dataset $w_0 = w_2 = 1$ and $w_1 = 2$ were chosen for all experiments, except for $w_1 = 1$ for SRF for *'ANGUS-Net-6'*, $w_1 = 5$ for SRF for *'ANGUS-DeepLabv3'* and *'ANGUS-TransUNet'*, $w_1 = 1$ for PED for *'ANGUS-TransUNet'* and $w_1 = 5$ for PED for *'ANGUS-Net-6'* and *'ANGUS-DeepLabv3'*. For all models in experiment C, $w_0 = 0.1$ was used for background, $w_2 = w_3 = w_4 = w_5 = 5$ for all target classes and $w_6 = 2$ for the anomaly class.

During training, random data augmentations were applied to the input images, including horizontal flipping, rotations up to 20°, horizontal/vertical translations up to 20% of the B-scan size, scaling up to 10% and brightness and contrast jitter up to 10%. Furthermore, we also trained the model using speckle and Gaussian noise augmentation, with $\mu = 0$ and $\sigma = 0.1$ parameters each. The models were trained for 150 epochs with a mini batch size of 2. The one with the best Dice on the validation set was selected as the final model for evaluation on the test set.

4. Results

Quantitative results of the proposed method and the baseline in *'clinical-trial'*, *'real-world'* and *'RETOUCH challenge'* test sets are provided in Table 1. For each target class, Dice, precision and recall are provided, and statistically significant differences are indicated with an asterisk. A performance improvement in terms of Dice is observed for all targets in comparison with the baseline model, easily noticeable in the *'real-world'* dataset. For all experimental combinations in Table 1, the proposed method outperforms the baseline with respect to Dice (10 out of 10 combinations). The largest absolute improvement in Dice is observed in this dataset for SHRM segmentation (0.287), while the smallest is seen for SRF in the *'RETOUCH challenge'* test set. This is also reflected in the qualitative results, illustrated in Fig. 3, Fig. 4 and Fig. 5. For each dataset and target, a representative sample with the original B-scan, its manual annotation, the prediction of the baseline model and those obtained with the proposed approach are depicted. In the first row of Fig. 3, the baseline model misses a significant amount of IRC, while the segmentation output of the ANGUS-Net more accurately resembles the manual annotation. This aligns with the quantitative results, showing a higher recall for the proposed model compared to the baseline. A typical example for SRF segmentation in the *'clinical-trial'* dataset is provided in the second row of Fig. 3, with a substantial overlap between manual annotation, baseline and ANGUS-Net prediction. The performance improvements by the proposed method in the *'real-world'* dataset are visually reflected in Fig. 4, including a reduction of outliers for segmenting IRC, HRF and pseudodrusen, an increased overlap between manual annotations and model segmentations for SHRM and a decrease of severe oversegmentation for SRF. Fig. 5 depicts typical results on the *'RETOUCH challenge'* test set, showing higher sensitivity at the expense of lower precision for the ANGUS-Net compared to the baseline for IRC, similar results for SRF, and a clearly improved predictions by the proposed model for PED.

In some cases, the baseline shows either a significantly higher precision or recall but is outperformed in terms of Dice by the proposed approach. As Dice is the harmonic mean between precision and recall, this indicates either a under- or over-segmentation of the baseline model. This is also reflected in the qualitative results. In Fig. 3 the baseline model under-segments IRC, coinciding with a significantly higher precision but lower Dice in Table 1 for IRC in the *'clinical-trial'* dataset. In Fig. 4 the baseline model over-segments SRF, concurring with a significantly higher recall but lower Dice in Table 1 for SRF in the *'real-world'* dataset.

Experiment A. Table 2 provides quantitative results in the *'clinical-trial'*, *'real-world'* and *'RETOUCH challenge'* test sets, using four different backbones. The results are in line with those obtained using the original structure. For almost all architectures and targets, and in all datasets, significant performance improvements for the ANGUS-Net approach are observed. In particular, all targets in the *'real-world'* test set show a significantly higher performance in terms of Dice. The only combinations without improvement are IRC using the U-Net-6 and SRF using the TransUNet architecture in the *'clinical-trial'* dataset, SRF using the DeepLabv3 architecture in the *'RETOUCH challenge'* dataset and IRC using the TransUNet architecture in the *'RETOUCH challenge'* dataset. Overall, in 36 out of 40 combinations the proposed approach outperformed the baseline in terms of Dice, and the average Dice across targets was always higher. The largest absolute increase in Dice is noticed for PED in *'RETOUCH challenge'* (+0.265 for *U-Net-6*, +0.245 for *AttUNet*, +0.219 for *TransUNet*), SRF in clinical trial (+0.242 for *AttUNet*) and SHRM in *'real-world'* dataset (+0.185 for *U-Net-6*, +0.111 for *DeepLabv3*).

Table 1
Quantitative segmentation results on the *'clinical-trial'*, *'real-world'* and *'RETOUCH challenge'* test sets. Dice, precision and recall (± standard deviation) are shown for each target individually and for the average across all targets. Highest Dice values are highlighted in bold, while highest precision and recall values are indicated in italics. The asterisk indicates statistically significant differences.

	Model	Dice	Precision	Recall	Dice	Precision	Recall	Dice	Precision	Recall
		IRC			SRF			Average across targets		
clinical-trial	U-Net	0.645 (±0.24)	*0.685 (±0.26)**	0.695 (±0.24)	0.790 (±0.24)	*0.846 (±0.20)*	0.808 (±0.28)	0.717	*0.766*	0.751
	ANGUS-Net	**0.658** (±0.23)	0.667 (±0.25)	*0.718 (±0.22)**	**0.795 (±0.23)**	0.826 (±0.21)	*0.817 (±0.26)*	**0.727**	0.746	*0.767*
		IRC			HRF			Pseudodrusen		
real-world	U-Net	0.267 (±0.24)	0.237 (±0.25)	0.547 (±0.40)	0.283 (±0.21)	0.196 (±0.17)	*0.727 (±0.36)**	0.129 (±0.08)	0.073 (±0.05)	*0.842 (±0.27)**
	ANGUS-Net	**0.363 (±0.27)***	*0.385 (±0.33)**	*0.593 (±0.36)*	**0.329 (±0.24)***	*0.271 (±0.23)**	0.581 (±0.36)	**0.210 (±0.16)***	*0.142 (±0.12)**	0.609 (±0.34)
		SHRM			SRF			Average across targets		
	U-Net	0.413 (±0.25)	0.463 (±0.32)	0.464 (±0.26)	0.467 (±0.28)	0.360 (±0.25)	*0.850 (±0.32)**	0.312	0.266	*0.686*
	ANGUS-Net	**0.699 (±0.27)***	*0.691 (±0.30)**	*0.781 (±0.28)**	**0.574 (±0.34)***	*0.656 (±0.36)**	0.598 (±0.37)	**0.435**	*0.429*	0.633
		IRC			SRF			PED		
RETOUCH	U-Net	0.723 (±0.23)	*0.800 (±0.21)**	0.708 (±0.25)	0.638 (±0.30)	*0.732 (±0.30)*	*0.630 (±0.32)*	0.424 (±0.33)	0.815 (±0.38)	0.330 (±0.29)
	ANGUS-Net	**0.725 (±0.19)***	0.693 (±0.20)	*0.820 (±0.22)**	**0.640 (±0.28)**	0.730 (±0.28)	0.629 (±0.30)	**0.638 (±0.28)***	*0.855 (±0.27)**	*0.559 (±0.30)**
		Average across targets								
	U-Net	0.595	*0.782*	0.556						
	ANGUS-Net	**0.667**	0.759	*0.669*						

Experiment B. Fig. 6 provides plots, illustrating the performance of both baseline and ANGUS-Net in *'clinical-trial'* and *'real-world'* test sets when trained with fewer samples. For almost all combinations of datasets, targets and number of training samples (in 46 out of 48 combinations), the proposed model consistently outperformed the baseline. For IRC in the *'clinical-trial'* with 50% training data and SRF in the *'real-world'* dataset with 50% training samples the Dice of the baseline is slightly higher. In general, we noticed that a larger number of training samples does not always lead to a better performance in the test set.

Experiment C. Quantitative results are shown in Table 3 for the baseline and the proposed method using both the original and the adapted backbone structures, when trained using a multiclass objective for simultaneously predicting all targets. A clear performance improvement is noticed also under this setting when incorporating the anomalous region as an additional class. In particular, significant performance improvements in Dice were observed for almost all combinations of targets and backbone structures (in 21 out of 25 combinations). Only for the segmentation of IRC using *U-Net-6-Multi* and *DeepLabv3-Multi*, and for SHRM using *AttUNet-Multi* and *TransUNet-Multi* no performance improvement was noticed, while the average Dice across all targets was improved for all backbones.

Experiment D. Fig. 7 illustrates the results. The proposed ANGUS-Net outperformed the baseline in terms of all lesion-wise evaluation scores for all targets and datasets, except for IRC in *'RETOUCH challenge'* (19 out of 20 combinations). SRF in the *'clinical-trial'* test set exhibited the smallest improvement in terms of absolute difference in AUC, while the largest improvement is noticed for IRC in the *'real-world'* test set. Additionally, the discrepancy in LD-Re_d and LD-Pr_d values are clearly larger for the baseline compared to the proposed ANGUS-Net.

Experiment E. Plots illustrating the influence of hyperparameter w_c are provided in Fig. 8. All multi-target backbones showed a similar behavior, with worse performance of the proposed method in case of extreme low or high values for the anomaly class, and superior performance of the proposed method when using the same or a slightly lower weight for the anomaly class than for the target classes (1,2 or 5).

Experiment F. The results of experiment *F.1*, directly applying the trained models on the external evaluation dataset *'OCT-Kermany'*, are depicted in Fig. 9. For all targets, the proposed method achieves an improvement in terms of average Dice performance across all models: 0.68 vs 0.71 for IRC (Fig. 9(a)), 0.57 vs 0.61 for SRF (Fig. 9(b)), and 0.29 vs. 0.58 for PED (Fig. 9(c)).

Cross-disease evaluation results for experiment *F.2* on the *'RETOUCH challenge RVO'* and *'RETOUCH challenge AMD'* test sets are shown in Fig. 10. In both settings, the proposed approach *'ANGUS-Net'* achieves higher Dice scores compared to the baseline *'U-Net'* model. More specifically, the *'U-Net'/'ANGUS-Net'* models trained on *'RETOUCH challenge AMD'* and evaluated on *'RETOUCH challenge RVO'* (Fig. 10(a)) achieve 0.70/0.71 for IRC and 0.67/0.75 for SRF, respectively. The *'U-Net'/'ANGUS-Net'* models trained on *'RETOUCH challenge RVO'* and evaluated on *'RETOUCH challenge AMD'* (Fig. 10(b)) achieve 0.51/0.58 for IRC and 0.34/0.42 for SRF, respectively.

Experiment G. The mean Spearman's correlation coefficient between the Dice values achieved by the *'ANGUS-Net'/'ANGUS-Net-6'/'ANGUS-DeepLabv3'/'ANGUS-AttUNet'/'ANGUS-TransUNet'* models and the 'residual anomalous area' across the three datasets and targets are 0.14/0.10/0.14/0.13/0.16. These indicate a very weak association between a potential performance decrease related to a smaller residual anomalous area. This is in line with the observation in experiment C, showing also consistent performance improvements of our approach for the multi-class segmentation models.

These results clearly demonstrate that the proposed approach is robust with respect to the used backbone structure, quantity of training samples as well as the number of target classes, various out-of-distribution scenarios and reduces the number of false positive and false negative detected lesions.

5. Discussion

We propose a novel way to improve the performance of lesion segmentation models by using anomaly detection as an additional weak supervision signal. We hypothesized that the additional information helps to overcome the lack of explicit guidance and therefore is useful to learn the occurrence of lesions within unhealthy areas. In this study, results show that the proposed strategy achieves better results compared to using a standard supervised learning model with no extra context. The observed segmentation improvements of several challenging targets are consistent across different experiments, indicating that the proposed approach can enhance model's ability to delineate abnormal retinal lesions.

In particular, we hypothesize that the incorporation of an additional anomaly map provides spatial context to better localize the structures of interest. This can also be described as 'attention through anomaly-supervision', as the extra labels guide the network's focus towards the lesion. The results in our experiments indicate that this 'attention through anomaly-supervision' provides complementary information compared to attention mechanisms built into the architecture of the network, as the segmentation performance is also consistently improved for the Attention U-Net (Schlemper et al., 2019) and the TransUNet (Chen et al., 2021). Moreover, the network not only learns to distinguish the target class from everything else, but also to differentiate between the target class, all other types of anomalies and the healthy retina, making the model much more semantically powerful.

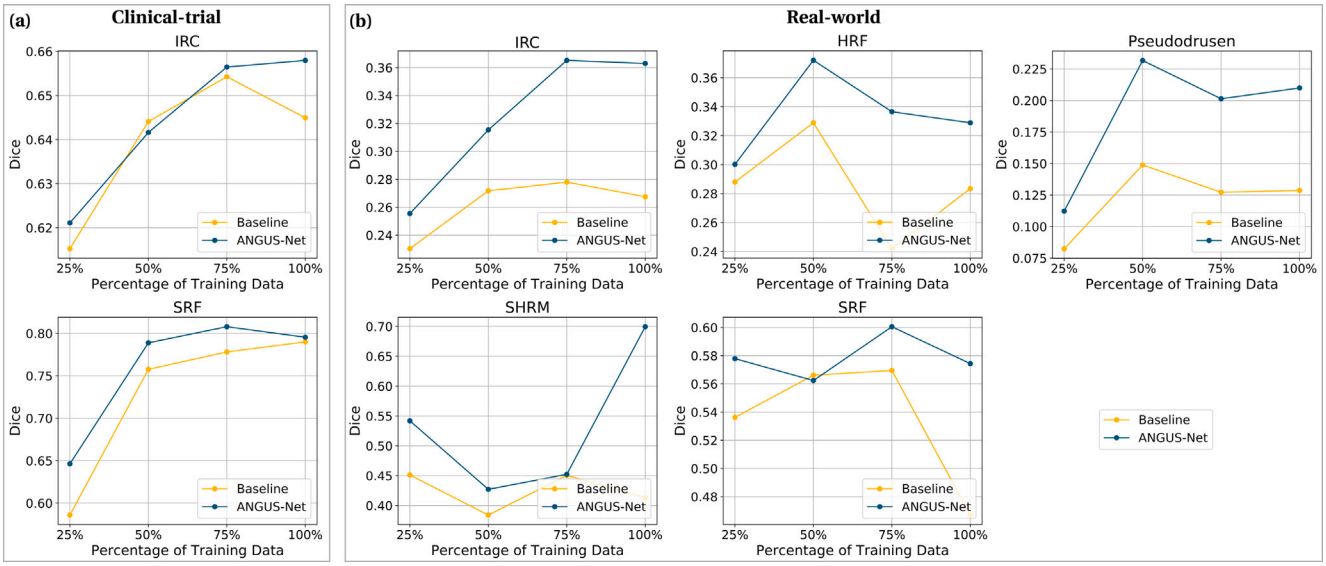

Fig. 6. Experiment B: Plots showing the quantitative performance in both (a) *'clinical-trial'* and (b) *'real-world'* test sets for models trained with 25%, 50%, 75% and 100% of the training data. While the results of the baseline are illustrated in orange, the ones of the proposed ANGUS-Net are highlighted in blue.

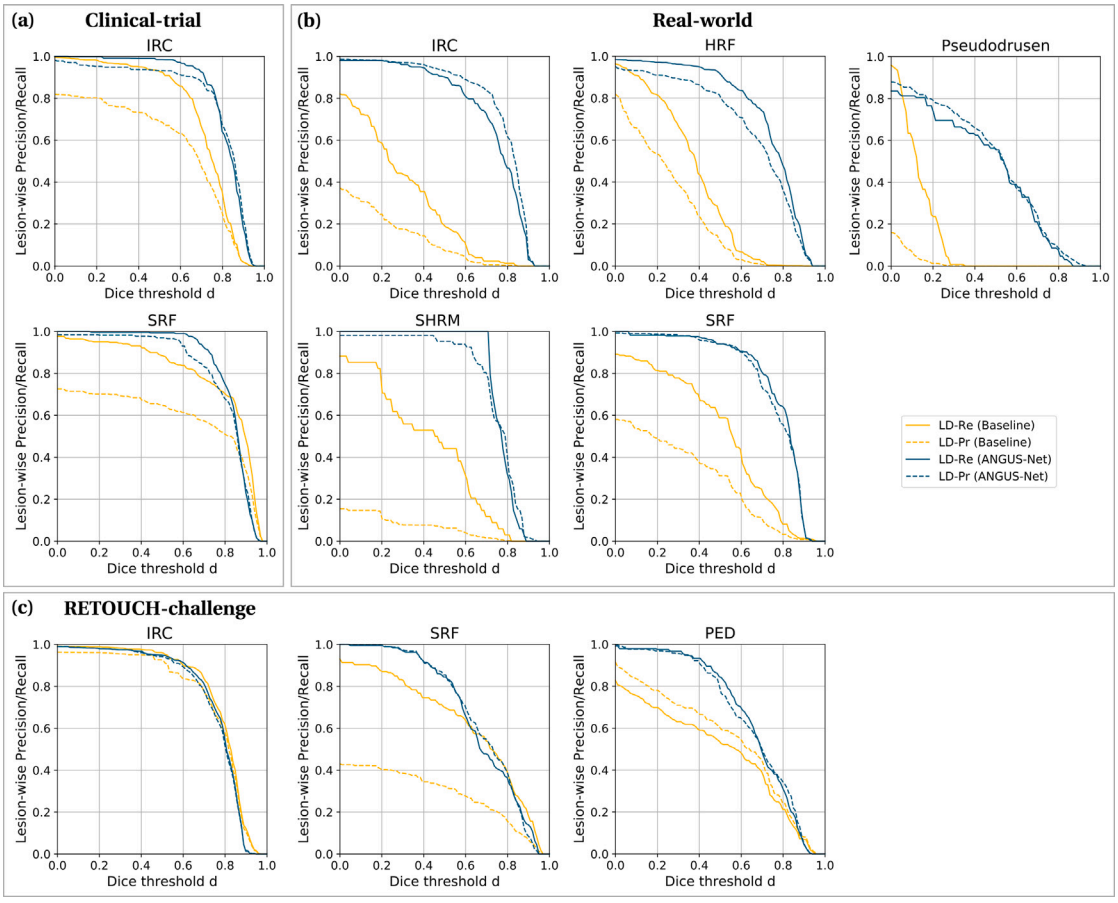

Fig. 7. Experiment D: Lesion-wise detection performance in (a) *'clinical-trial'*, (b) *'real-world'* and (c) *'RETOUCH challenge'* test sets. The lesion-wise precision/recall curves (dashed/solid) are illustrated both for the baseline model (orange) and the proposed ANGUS-Net (blue). In (a) *'clinical-trial'*, the baseline achieved an lesion-wise precision/recall area under the curve (AUC) of 0.55/0.71 for IRC and 0.58/0.79 for SRF, while the proposed method outperformed the baseline in terms of this measure with 0.79/0.81 and 0.80/0.83, respectively. In (b) *'real-world'*, the baseline is clearly outperformed by the proposed method, which is reflected in lesion-wise precision/recall AUC values of 0.12/0.28 vs. 0.77/0.73 for IRC, 0.23/0.36 vs. 0.65/0.73 for HRF, 0.02/0.13 vs. 0.47/0.44 for Pseudodrusen, 0.06/0.40 vs. 0.74/0.76 for SHRM and 0.28/0.49 vs. 0.76/0.77 for SRF. In (c) *'RETOUCH challenge'*, the baseline is outperformed by the proposed method except for a single value in terms of lesion-wise precision/recall AUC: 0.744/0.778 vs. 0.750/0.744 for IRC, 0.280/0.618 vs. 0.686/0.677 for SRF and 0.526/0.468 vs. 0.658/0.666 for PED.

Table 2
Experiment A: Quantitative segmentation results using a different structure for the backbone on the *'clinical-trial'*, *'real-world'* and *'RETOUCH challenge'* test sets. Dice, precision and recall (± standard deviation) are shown for each target individually and for the average across all targets. Highest Dice values are highlighted in bold, while highest precision and recall values are indicated in italics. The asterisk indicates statistically significant differences.

	Model	Dice	Precision	Recall	Dice	Precision	Recall	Dice	Precision	Recall
		IRC			SRF			Average across targets		
clinical-trial	U-Net-6	**0.662 (±0.24)***	*0.695 (±0.26)**	0.709 (±0.24)	0.791 (±0.25)	*0.845 (±0.20)**	0.801 (±0.28)	0.727	*0.770*	0.755
	ANGUS-Net-6	0.659 (±0.23)	0.664 (±0.26)	*0.717 (±0.22)*	**0.805 (±0.20)***	0.801 (±0.21)	*0.858 (±0.22)**	**0.732**	0.732	*0.788*
	DeepLabv3	0.622 (±0.22)	*0.627 (±0.25)**	0.691 (±0.21)	0.782 (±0.25)	*0.827 (±0.22)**	0.794 (±0.28)	0.702	*0.727*	0.743
	ANGUS-DeepLabv3	**0.631 (±0.21)**	0.578 (±0.23)	*0.779 (±0.19)**	**0.803 (±0.20)**	0.824 (±0.21)	*0.825 (±0.22)**	**0.717**	0.701	*0.802*
	AttUNet	0.647 (±0.24)	*0.639 (±0.27)**	0.741 (±0.21)	0.491 (±0.26)	0.824 (±0.30)	0.390 (±0.25)	0.569	*0.732*	0.566
	ANGUS-AttUNet	**0.645 (±0.22)**	0.591 (±0.24)	*0.799 (±0.22)**	**0.733 (±0.27)***	*0.860 (±0.20)**	*0.713 (±0.30)**	**0.689**	0.725	*0.756*
	TransUNet	0.628 (±0.24)	*0.734 (±0.28)**	0.603 (±0.23)	**0.808 (±0.21)***	0.848 (±0.19)	*0.818 (±0.24)**	**0.718**	*0.791*	0.711
	ANGUS-TransUNet	**0.652 (±0.22)***	0.665 (±0.25)	*0.713 (±0.23)**	0.771 (±0.25)	*0.855 (±0.19)**	0.767 (±0.28)	0.711	0.760	*0.740*
		IRC			HRF			Pseudodrusen		
real-world	U-Net-6	0.313 (±0.22)	0.245 (±0.21)	*0.769 (±0.31)**	0.260 (±0.22)	0.180 (±0.17)	*0.675 (±0.38)**	0.134 (±0.09)	0.076 (±0.06)	*0.813 (±0.32)**
	ANGUS-Net-6	**0.373 (±0.27)***	*0.328 (±0.28)**	0.679 (±0.34)	**0.345 (±0.25)***	*0.292 (±0.24)**	0.565 (±0.37)	**0.177 (±0.12)***	*0.107 (±0.08)**	0.741 (±0.33)
	DeepLabv3	0.149 (±0.22)	0.229 (±0.30)	0.211 (±0.32)	0.167 (±0.15)	0.108 (±0.11)	*0.553 (±0.39)**	0.104 (±0.07)	0.059 (±0.05)	*0.812 (±0.29)**
	ANGUS-DeepLabv3	**0.237 (±0.27)***	*0.241 (±0.29)*	*0.343 (±0.39)**	**0.201 (±0.18)***	*0.149 (±0.15)**	0.415 (±0.35)	**0.132 (±0.11)***	*0.080 (±0.08)**	0.667 (±0.36)
	AttUNet	0.238 (±0.23)	*0.266 (±0.266)*	*0.440 (±0.40)**	0.274 (±0.21)	0.190 (±0.17)	*0.688 (±0.37)**	0.096 (±0.08)	0.054 (±0.05)	0.867 (±0.28)
	ANGUS-AttUNet	**0.242 (±0.31)**	0.243 (±0.31)	0.346 (±0.412)	**0.336 (±0.24)***	*0.263 (±0.21)**	0.617 (±0.35)	**0.139 (±0.09)***	*0.079 (±0.06)**	*0.869 (±0.24)*
	TransUNet	0.210 (±0.27)	0.222 (±0.29)	*0.304 (±0.39)*	0.307 (±0.24)	0.229 (±0.20)	*0.618 (±0.37)**	0.123 (±0.13)	0.077 (±0.11)	0.786 (±0.35)
	ANGUS-TransUNet	**0.246 (±0.31)***	*0.278 (±0.35)**	0.290 (±0.37)	**0.321 (±0.24)***	*0.250 (±0.21)**	0.58837 (±0.)	**0.186 (±0.13)***	*0.119 (±0.10)**	*0.788 (±0.30)*
		SHRM			SRF			Average across targets		
	U-Net-6	0.473 (±0.28)	*0.708 (±0.31)*	0.503 (±0.34)	0.444 (±0.27)	0.335 (±0.24)	*0.843 (±0.33)**	0.325	0.309	*0.721*
	ANGUS-Net-6	**0.658 (±0.26)***	0.706 (±0.27)	*0.717 (±0.29)**	**0.581 (±0.33)***	*0.639 (±0.36)**	0.615 (±0.34)	**0.427**	*0.414*	0.663
	DeepLabv3	0.480 (±0.26)	0.570 (±0.35)	0.484 (±0.26)	0.438 (±0.27)	0.347 (±0.25)	*0.818 (±0.35)**	0.268	0.263	*0.576*
	ANGUS-DeepLabv3	**0.591 (±0.27)***	*0.739 (±0.27)**	*0.613 (±0.29)**	**0.533 (±0.34)***	*0.590 (±0.34)**	0.568 (±0.37)	**0.339**	*0.360*	0.521
	AttUNet	0.471 (±0.25)	0.536 (±0.33)	*0.619 (±0.35)**	0.519 (±0.29)	0.410 (±0.26)	*0.848 (±0.33)**	0.319	0.291	*0.692*
	ANGUS-AttUNet	**0.480 (±0.35)**	*0.655 (±0.41)**	0.440 (±0.36)	**0.597 (±0.32)***	*0.646 (±0.34)**	0.634 (±0.34)	**0.359**	*0.377*	0.581
	TransUNet	0.529 (±0.)22	0.536 (±0.28)	*0.675 (±0.22)**	0.462 (±0.28)	0.391 (±0.27)	*0.780 (±0.36)**	0.326	0.291	*0.633*
	ANGUS-TransUNet	**0.532 (±0.34)**	*0.656 (±0.39)**	0.497 (±0.34)	**0.586 (±0.31)***	*0.619 (±0.31)**	*0.678 (±0.35)**	**0.374**	*0.384*	0.568
		IRC			SRF			PED		
RETOUCH	U-Net-6	0.729 (±0.23)	*0.800 (±0.21)**	0.713 (±0.25)	**0.652 (±0.28)***	*0.800 (±0.25)**	0.625 (±0.31)	0.461 (±0.32)	*0.825 (±0.37)**	0.365 (±0.30)
	ANGUS-Net-6	**0.747(±0.19)**	0.740(±0.187)	*0.806 (±0.213)**	0.644 (±0.26)	0.752 (±0.26)	*0.654 (±0.30)*	**0.726 (±0.25)***	0.799 (±0.24)	*0.720(±0.27)**
	DeepLabv3	0.685 (±0.22)	*0.732 (±0.20)**	0.701 (±0.25)	**0.616 (±0.27)***	*0.716 (±0.28)*	0.624 (±0.31)	0.532 (±0.31)	*0.842 (±0.34)**	0.433 (±0.29)
	ANGUS-DeepLabv3	**0.690 (±0.20)***	0.645 (±0.21)	*0.798 (±0.22)**	0.587 (±0.28)	0.605 (±0.28)	*0.689 (±0.33)**	**0.639 (±0.28)***	0.792 (±0.27)	*0.610 (±0.30)**
	AttUNet	0.690 (±0.25)	*0.782 (±0.24)**	0.662 (±0.27)	0.578 (±0.31)	*0.785 (±0.29)**	*0.523 (±0.32)*	0.294 (±0.30)	0.717 (±0.45)	0.215 (±0.25)
	ANGUS-AttUNet	**0.733 (±0.18)**	0.714 (±0.19)	*0.811 (±0.22)**	**0.612 (±0.28)***	0.732 (±0.29)	*0.594 (±0.31)**	**0.539 (±0.27)***	*0.823 (±0.31)**	*0.441 (±0.26)**
	TransUNet	**0.715 (±0.21)***	*0.788 (±0.20)**	0.701 (±0.24)	0.625 (±0.29)	*0.721 (±0.30)**	0.640 (±0.31)	0.410 (±0.32)	0.810 (±0.38)	0.317 (±0.28)
	ANGUS-TransUNet	0.698 (±0.19)	0.634 (±0.20)	*0.849 (±0.21)**	**0.637 (±0.24)**	0.610 (±0.27)	*0.799 (±0.26)**	**0.629 (±0.29)***	*0.845 (±0.28)**	*0.544 (±0.30)**
		Average across targets								
	U-Net-6	0.614	*0.808*	0.567						
	ANGUS-Net-6	**0.705**	0.764	*0.726*						
	DeepLabv3	0.684	*0.791*	0.677						
	ANGUS-DeepLabv3	**0.721**	0.755	*0.757*						
	AttUNet	0.521	*0.762*	0.467						
	ANGUS-AttUNet	**0.628**	0.756	*0.615*						
	TransUNet	0.583	*0.773*	0.552						
	ANGUS-TransUNet	**0.654**	0.696	*0.731*						

Table 3
Experiment C: Quantitative segmentation results for the multi-class models on the *'real-world'* test set, trained on all targets simultaneously (IRC, HRF, Pseudodrusen, SHRM, SRF). Dice, precision and recall (± standard deviation) are shown for each target individually and for the average across all targets. Highest Dice values are highlighted in bold, while highest precision and recall values are indicated in italics. The asterisk indicates statistically significant differences.

Model	Dice	Precision	Recall	Dice	Precision	Recall	Dice	Precision	Recall
	IRC			HRF			Pseudodrusen		
U-Net-6-Multi	**0.267 (±0.26)**	0.213 (±0.25)	*0.640 (±0.42)**	0.262 (±0.21)	0.183 (±0.17)	*0.657 (±0.38)**	0.136 (±0.09)	0.077 (±0.06)	*0.894 (±0.20)**
ANGUS-Net-6-Multi	0.251 (±0.30)	*0.247 (±0.31)**	0.417 (±0.44)	**0.335 (±0.26)***	*0.287 (±0.25)**	0.553 (±0.35)	**0.176 (±0.10)***	*0.104 (±0.06)**	0.750 (±0.29)
DeepLabv3-Multi	**0.235 (±0.24)**	0.189 (±0.23)	*0.523 (±0.43)**	0.498 (±0.27)	0.396 (±0.25)	*0.803 (±0.34)**	0.099 (±0.06)	0.054 (±0.04)	*0.870 (±0.21)**
ANGUS-DeepLabv3-Multi	0.231 (±0.29)	*0.260 (±0.32)**	0.273 (±0.35)	**0.559 (±0.31)***	*0.611 (±0.32)**	0.597 (±0.34)	**0.146 (±0.09)***	*0.085 (±0.06)**	0.744 (±0.29)
AttUNet-Multi	0.205 (±0.27)	0.193 (±0.27)	*0.403 (±0.45)**	0.193 (±0.18)	0.131 (±0.15)	*0.692 (±0.36)**	0.053 (±0.04)	0.027 (±0.02)	*0.949 (±0.14)**
ANGUS-AttUNet-Multi	**0.291 (±0.32)***	*0.291 (±0.33)**	0.381 (±0.40)	**0.351 (±0.26)***	*0.313 (±0.27)**	0.536 (±0.35)	**0.166 (±0.10)***	*0.097 (±0.07)**	0.846 (±0.22)
TransUNet-Multi	0.206 (±0.26)	0.167 (±0.23)	*0.409 (±0.44)**	0.218 (±0.19)	0.148 (±0.15)	*0.614 (±0.38)**	0.115 (±0.08)	0.064 (±0.05)	*0.869 (±0.23)**
ANGUS-TransUNet-Multi	**0.244 (±0.31)***	*0.274 (±0.34)**	0.317 (±0.40)	**0.318 (±0.26)***	*0.283 (±0.26)**	0.461 (±0.36)	**0.160 (±0.11)***	*0.094 (±0.07)**	0.832 (±0.27)
U-Net-Multi	0.321 (±0.27)	0.272 (±0.27)	*0.691 (±0.39)**	0.300 (±0.24)	0.226 (±0.21)	*0.639 (±0.36)**	0.125 (±0.08)	0.069 (±0.05)	*0.911 (±0.16)**
ANGUS-Net-Multi	**0.326 (±0.29)***	*0.300 (±0.30)**	0.532 (±0.40)	**0.325 (±0.24)***	*0.286 (±0.25)**	0.518 (±0.34)	**0.142 (±0.10)***	*0.082 (±0.06)**	0.771 (±0.32)
	SHRM			SRF			Average across targets		
U-Net-6-Multi	0.609 (±0.28)	0.720 (±0.29)	0.583 (±0.30)	0.531 (±0.28)	0.438 (±0.27)	*0.811 (±0.33)**	0.361	0.326	*0.717*
ANGUS-Net-6-Multi	**0.683 (±0.24)***	*0.731 (±0.26)*	*0.698 (±0.27)**	**0.582 (±0.31)***	*0.647 (±0.32)**	0.616 (±0.35)	**0.405**	*0.403*	0.607
DeepLabv3-Multi	0.194 (±0.27)	0.446 (±0.46)	0.166 (±0.26)	0.153 (±0.15)	0.097 (±0.10)	*0.483 (±0.38)**	0.236	0.236	*0.569*
ANGUS-DeepLabv3-Multi	**0.308 (±0.30)**	*0.651 (±0.45)**	*0.236 (±0.27)*	**0.176 (±0.18)***	*0.159 (±0.19)**	0.282 (±0.31)	**0.284**	*0.353*	0.426
AttUNet-Multi	**0.447 (±0.29)***	0.524 (±0.37)	*0.479 (±0.28)**	0.447 (±0.29)	0.383 (±0.28)	*0.734 (±0.36)**	0.269	0.252	*0.651*
ANGUS-AttUNet-Multi	0.299 (±0.26)	*0.633 (±0.43)**	0.221 (±0.21)	**0.583 (±0.31)***	*0.680 (±0.32)**	0.610 (±0.34)	**0.339**	*0.403*	0.519
TransUNet-Multi	**0.564 (±0.24)***	*0.710 (±0.30)*	*0.543 (±0.25)**	0.501 (±0.28)	0.410 (±0.27)	*0.810 (±0.33)**	0.321	0.300	*0.649*
ANGUS-TransUNet-Multi	0.371 (±0.29)	0.682 (±0.43)	0.290 (±0.26)	**0.603 (±0.31)***	*0.584 (±0.31)**	0.748 (±0.33)	**0.339**	*0.383*	0.530
U-Net-Multi	0.491 (±0.30)	0.706 (±0.35)	0.4335 (±0.30)	0.503 (±0.28)	0.430 (±0.27)	*0.770 (±0.34)**	0.348	0.341	*0.689*
ANGUS-Net-Multi	**0.608 (±0.25)***	*0.759 (±0.28)*	*0.605 (±0.28)**	**0.558 (±0.32)***	*0.647 (±0.33)**	0.596 (±0.35)	**0.392**	*0.415*	0.604

Surprisingly, this comparatively simple idea has not been explored before. We believe this may be related to the fact that anomaly detection in medical imaging is in itself an active area of research, that has only recently been enriched with the possibilities of deep learning. From a conceptual perspective, the proposed approach implicitly inherits the advantages of anomaly detection models. First of all, it gains additional supervision without requiring any extra manual labels. Furthermore, its applicability is not limited to a specific disease,

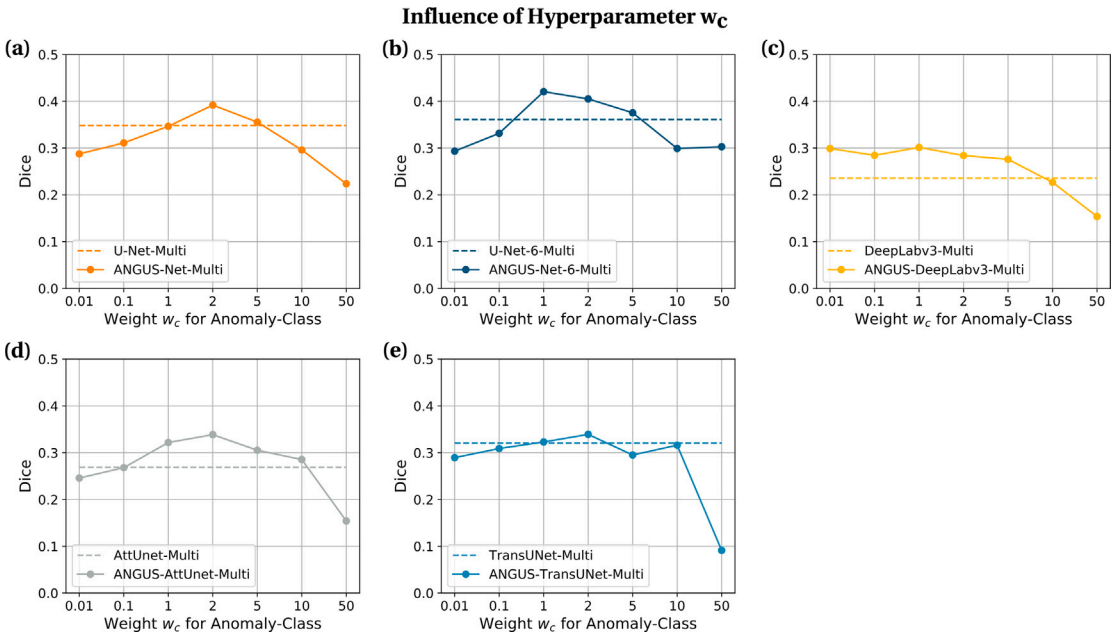

Fig. 8. Experiment E: Plots showing the influence on the average Dice when varying the weighting hyperparameter for the anomaly-class ($w_c = [0.01, 0.1, 1, 2, 5, 10, 50]$). Results are illustrated on the 'real-world' test set for the models (a) *'ANGUS-Net-Multi'*, (b) *'ANGUS-Net-6-Multi'*, (c) *'ANGUS-DeepLabv3-Multi'*, (d) *'ANGUS-AttUnet-Multi'* and (e) *'ANGUS-TransUNet-Multi'* versus their corresponding baselines. For better readability, the values are plotted at equal intervals on the horizontal axis.

patient cohort or appearance, as by definition the anomaly detection model is trained only on normal samples, therefore independently of a particular disease. This last point was previously shown in Seeböck et al. (2019a) to hold for multiple diseases and anomaly appearances. Our proposed approach exhibits similar properties, with performance improvements across different datasets, diseases, and lesion targets (Section 4, experiment *A*, *C* and *F*).

Moreover, specific (supervised) lesion segmentation and anomaly detection models complement each other well from a conceptual point of view. While anomaly detection methods seek to detect all deviations from normal appearance by learning from normal samples only, (supervised) lesion segmentation approaches aim to detect a particular pathological structure which typically depicts only a very specific anomaly subgroup. This conceptual difference is also reflected in experimental results when using anomaly detection models as unsupervised baselines for the task of segmenting specific lesions, achieving a relatively high recall and low precision at the same time.

The proposed approach is also generic in the sense that it does not rely on a specific network architecture, which allows its usage in other more complex architectures as well as in combination with other sophisticated techniques that aim at improving segmentation performances (Section 1.1). This assumption is supported by the results of experiment *A*, which shows robustness when changing the structure of the underlying backbone. This holds not only in the single target context but also in a multi-class setting for the simultaneous segmentation of multiple lesions (Section 4, experiment *C*).

Furthermore, the results of experiment *B* indicate that the proposed method is more data efficient. As visualized in Fig. 6, a larger number of training samples does not necessarily lead to better generalization performance. We hypothesize that this might be caused by variations in the overall class balancing when de- or increasing the number of images. This results in changes of the general distribution of the samples in the training set, making it more or less similar to the class distribution in the test set.

In some specific experimental combinations, the Dice could not be improved by the proposed method. However, a higher Dice for the proposed method was observed in 132 out of 143 combinations across all experiments 'A', 'B', 'C' and 'D' (Section 4).

Despite the clear performance improvements in our experiments, we noticed that both the baseline and our model reported low Dice values for IRC, HRF and pseudodrusen in the real-world dataset. This might be due to multiple factors: the heterogeneity of the real world data, the difficulty of manually annotating lesions, and the difficulty of achieving a high Dice when the lesion is small (Reinke et al., 2021). Nevertheless, the lesion-wise evaluation metrics for the proposed approach are at the same time high, even for these targets. This indicates that the low Dice is mainly due to delineation and not localization errors of the lesions.

The clear performance improvements regarding the lesion-wise evaluation metrics also imply that the proposed approach greatly reduces the amount of lesion regions that are either missed or falsely detected by the baseline model. Since the LD-Re$_d$ and LD-Pr$_d$ curves of the ANGUS-Net are closer to each other compared to the baseline, we hypothesize that the proposed approach leads to models which are better balanced in terms of weighting false positives and false negatives.

Even though these numbers cannot be directly compared, our model achieved higher Dice values than the mean of the competing methods reported by Bogunović et al. (2019) in the *RETOUCH challenge* Spectralis test set. Furthermore, the best ANGUS based models achieved numbers close to or better than the reported mean inter-observer Dice (IRC: 0.74, SRF: 0.66, PED: 0.80) in absolute values. To increase the reproducibility of our work, our predicted anomaly maps on the RETOUCH challenge dataset will be made available online by the time of publication.[1]

Regarding computational effort, our approach adds limited overhead, as it only requires applying the anomaly detection model once on the training dataset, to produce the weak anomaly labels for training. We also did not notice any systematic effect on the convergence behavior during training. Finally, the inference time is the same for both the baseline and the proposed segmentation model.

One potential limitation is the situation in which a complete manual annotation of target structures is available, as in this case the pixel-wise binary map for the anomaly predictions ($\mathbb{1}_{\text{AN}}$, Section 2.1) could be mostly empty. Even though in this theoretical scenario the proposed method may be of limited value, this is not a realistic situation. In

[1] https://github.com/tbd

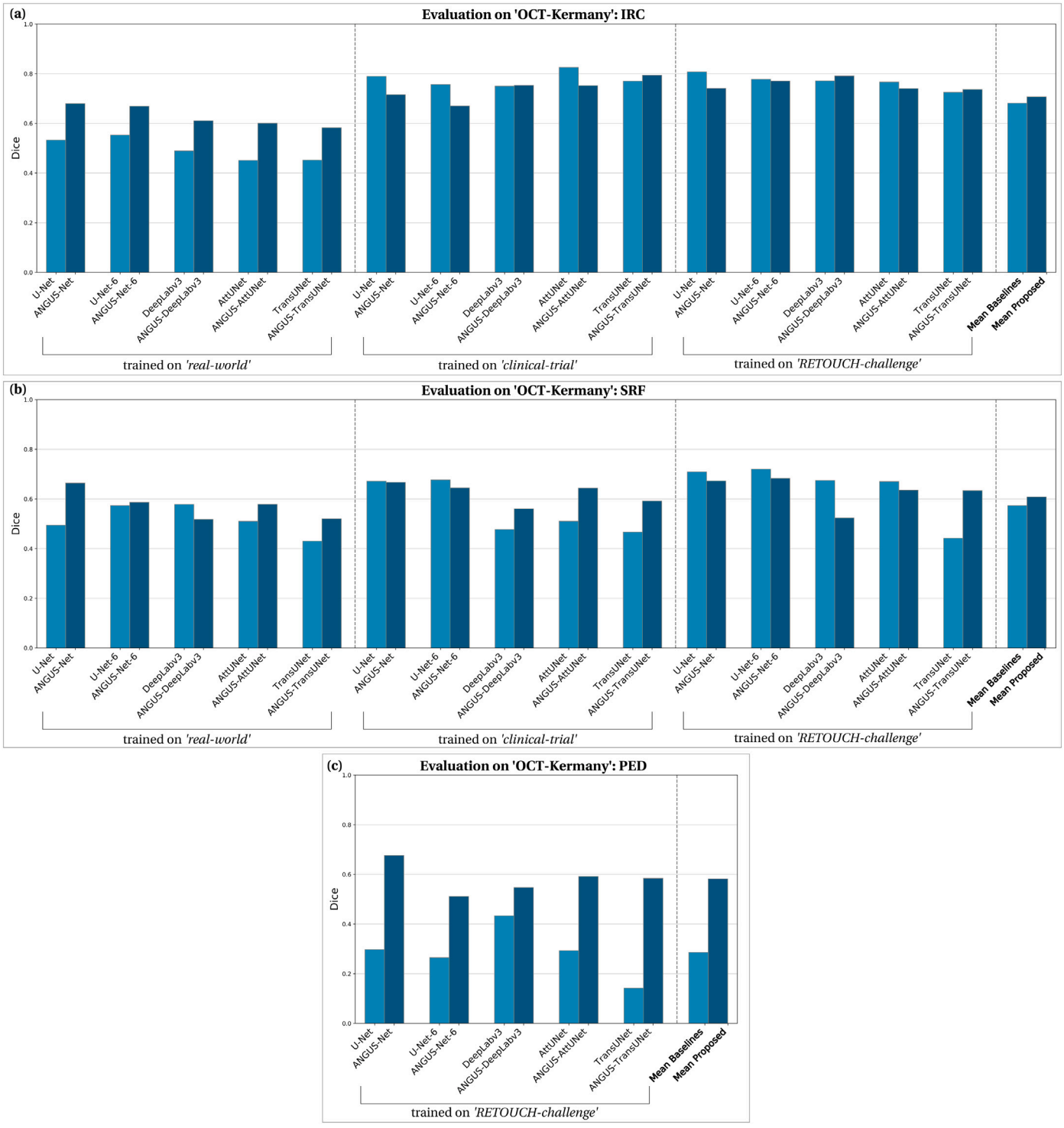

Fig. 9. Experiment F.1: Plots illustrating the performance in terms of Dice when directly applying the trained models on the external publicly available dataset (Kermany et al., 2018) *'OCT-Kermany'*. (a) IRC, (b) SRF and (c) PED segmentation results. For each target, the mean performance is additionally visualized on the right hand side.

particular, it is extremely costly to obtain a comprehensive detailed pixel-wise annotation covering every potential lesion type separately, particularly in the medical domain. Furthermore, in many clinical scenarios reaching a complete definition of all lesion types is not possible, as not all factors, patterns or lesion types relevant for a specific disease are known. Moreover, the anomaly detection model segments anomalous areas that do not necessarily correspond to a specific lesion type, meaning that even a complete annotation of all potential lesions may not have a total overlap with the anomalous area. For instance, this could be an anatomical alteration produced by the presence of a lesion nearby, such as a general swelling of the retina caused by IRC (Fig. 3, top row). Moreover, the results indicate a very weak association between a potential performance decrease related to a smaller size of the 'residual anomalous area' (experiment *G*).

The need to explicitly choose the weighting hyperparameter w_c is another potential drawback. At the same time, it may be beneficial to be able to explicitly control the influence of the anomaly context on the training loss. We also observed that even though our proposed method

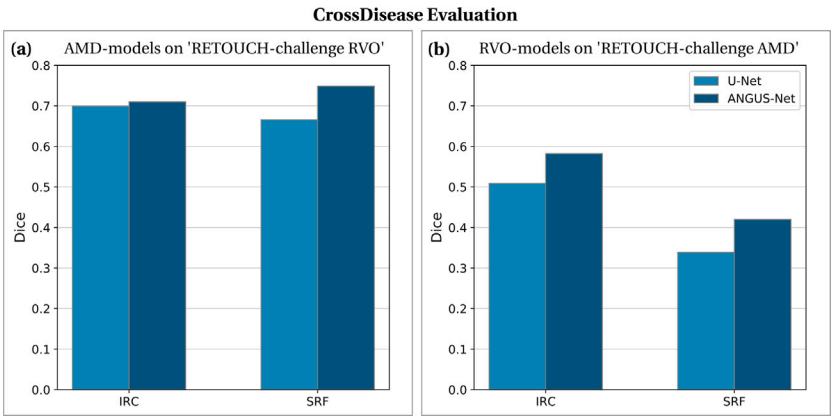

Fig. 10. Experiment F.2: Barplots of cross-disease evaluations. Results of U-Net/ANGUS-Net models trained on the *'RETOUCH challenge AMD'* training set and evaluated on the *'RETOUCH challenge RVO'* test set are illustrated in (a). Results of U-Net/ANGUS-Net models trained on the *'RETOUCH challenge RVO'* training set and evaluated on the *'RETOUCH challenge AMD'* test are illustrated in (b).

improved the performance for the multi-class models in experiment *C* compared to the baselines, the average absolute performance in terms of Dice was lower compared to the proposed single target models, for all backbones. We hypothesize that this difference in performance may stem from increased complexity of the multi-class problem.

Another limitation of this work is that we used the anomaly maps of only a single anomaly detection model for training the segmentation models in our experiments (WeakAnD (Seeböck et al., 2019a)). However, to the best of our knowledge, the utilized anomaly detection model is state-of-the-art in pixel-wise anomaly detection in OCT. In the Supplement, we also demonstrate superior performance in terms of anomaly detection of the utilized WeakAnD model compared to AnoDDPM, a recently published diffusion based anomaly detection method. This further strengthens the choice of the anomaly detection method used in our experiments. At the same time, we believe that the underlying simplicity of our approach allows to seamlessly exchange the underlying anomaly detection model and backbone architecture, and even applying it to other medical imaging domains. Here, we conjecture that a more accurate anomaly mask is beneficial for the performance of the final lesion segmentation model. However, this hypothesis needs to be investigated in future work.

Notice that adding the anomalous area as an additional weak label is significantly different from supervision with dilated lesion masks. The anomaly detection model identifies abnormal areas not always linked to a specific lesion but possibly associated with anatomical changes caused by a lesion. However, this abnormal alteration does not necessarily correspond to a consistently dilated area around the lesion, as reflected in all qualitative examples of this work.

6. Conclusion

In this paper, we proposed an anomaly guided approach for improving lesion segmentation in retinal OCT scans. Our method uses anomalies as an auxiliary task to provide additional weak supervision. By adding the semantic context, the approach achieved performance improvements in four different datasets and demonstrated robustness in terms of the underlying structure of the backbone, number of training samples and target classes, generalization to external test sets and unseen diseases as well as an improved lesion-wise detection. The proposed approach is generic enough to be combined with any underlying architecture or other techniques to enhance the segmentation performance and to be extended to other imaging domains, diseases or anatomical structures in future work. We believe that the simplicity of our method contributes to its efficacy and potential impact across diverse applications, as it is conceptually easy to extrapolate to other use case scenarios, can be easily combined with other techniques, is efficient in terms of computational resources and can serve as a foundation for further developments.

In this context, further efforts should not only seek to combine our proposed approach with other techniques, but also focus on other strategies to utilize anomaly detection for performance improvement, including the use of anomaly maps as pre-training task or as direct attention-mechanism. This could support clinicians in patient management, including early diagnosis, tracking disease progression and making treatment decisions.

CRediT authorship contribution statement

Philipp Seeböck: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **José Ignacio Orlando:** Conceptualization, Formal analysis, Investigation, Supervision, Validation, Writing – review & editing. **Martin Michl:** Data curation, Validation, Writing – review & editing. **Julia Mai:** Data curation, Validation, Writing – review & editing. **Ursula Schmidt-Erfurth:** Funding acquisition, Validation, Writing – review & editing. **Hrvoje Bogunović:** Conceptualization, Funding acquisition, Project administration, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Hrvoje Bogunovic reports financial support was provided by FWF Austrian Science Fund (grant number FG 9-N). Hrvoje Bogunovic reports a relationship with Heidelberg Engineering Inc that includes: funding grants. Ursula Schmidt-Erfurth reports a relationship with Genentech, Novartis, Roche, Heidelberg Engineering, Kodiak, RetInSight that includes: consulting or advisory. Hrvoje Bogunovic is editorial board member of the Medical Image Analysis journal.

Data availability

As stated in our manuscript, the predicted anomaly maps on the RETOUCH challenge dataset will be made available online once the manuscript has been accepted.

Acknowledgments

This research was funded in whole, or in part, by the Austrian Science Fund (FWF) [10.55776/FG9]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2024.103104.

References

Apostolopoulos, S., Zanet, S.D., Ciller, C., Wolf, S., Sznitman, R., 2017. Pathological OCT retinal layer segmentation using branch residual u-shape networks. In: Proc. of MICCAI. Springer, pp. 294–301.

Bai, F., Marques, M.J., Gibson, S.J., 2017. Cystoid macular edema segmentation of optical coherence tomography images using fully convolutional neural networks and fully connected CRFs. arXiv preprint arXiv:1709.05324.

Bercea, C.I., Neumayr, M., Rueckert, D., Schnabel, J.A., 2023. Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models. arXiv preprint arXiv:2305.19643.

Bogunović, H., et al., 2019. RETOUCH: the retinal OCT fluid detection and segmentation benchmark and challenge. IEEE Trans. Med. Imaging 38 (8), 1858–1874.

Burger, B., Bernathova, M., Seeböck, P., Singer, C.F., Helbich, T.H., Langs, G., 2023. Deep learning for predicting future lesion emergence in high-risk breast MRI screening: a feasibility study. Eur. Radiol. Exp. 7 (1), 32.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision. Springer, pp. 205–218.

Cao, D., et al., 2021. Hyperreflective foci, optical coherence tomography progression indicators in age-related macular degeneration, include transdifferentiated retinal pigment epithelium. Invest. Ophthalmol. Vis. Sci. 62 (10), 34.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. TransUNet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

Chen, S., Wang, Z., Shi, J., Liu, B., Yu, N., 2018. A multi-task framework with feature passing module for skin lesion classification and segmentation. In: ISBI. IEEE, pp. 1126–1129.

De Fauw, J., et al., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. 24 (9), 1342.

Ding, X., Shen, C., Che, Z., Zeng, T., Peng, Y., 2021. Scarf: A semantic constrained attention refinement network for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3002–3011.

Dufour, P.A., Abdillahi, H., Ceklic, L., Wolf-Schnurrbusch, U., Kowal, J., 2012. Pathology hinting as the combination of automatic segmentation with a statistical shape model. In: Proc. of MICCAI. Springer, pp. 599–606.

Fifty, C., Amid, E., Zhao, Z., Yu, T., Anil, R., Finn, C., 2021. Efficiently identifying task groupings for multi-task learning. Adv. Neural Inf. Process. Syst. 34.

Fujimoto, J., Swanson, E., 2016. The development, commercialization, and impact of optical coherence tomography. Invest. Ophthalmol. Vis. Sci. 57 (9).

Gal, Y., Ghahramani, Z., 2015. Bayesian convolutional neural networks with Bernoulli approximate variational inference. arXiv preprint arXiv:1506.02158.

Garvin, M.K., Abràmoff, M.D., Wu, X., Russell, S.R., Burns, T.L., Sonka, M., 2009. Automated 3-D intraretinal layer segmentation of macular spectral-domain optical coherence tomography images. IEEE Trans. Med. Imaging 28 (9), 1436–1447.

Gu, Z., et al., 2019. Ce-net: Context encoder network for 2d medical image segmentation. IEEE Trans. Med. Imaging 38 (10), 2281–2292.

He, X., Fang, L., Tan, M., Chen, X., 2022. Intra-and inter-slice contrastive learning for point supervised OCT fluid segmentation. IEEE Trans. Image Process. 31, 1870–1881.

He, T., Hu, J., Song, Y., Guo, J., Yi, Z., 2020. Multi-task learning for the segmentation of organs at risk with label dependence. Med. Image Anal. 61, 101666.

He, K., Zhang, X., Ren, S., et al., 2015. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proc. of IEEE ICCV. pp. 1026–1034.

He, K., et al., 2021a. HF-UNet: learning hierarchically inter-task relevance in multi-task U-net for accurate prostate segmentation in CT images. IEEE Trans. Med. Imaging 40 (8), 2118–2128.

He, Y., et al., 2021b. Structured layer surface segmentation for retina OCT using fully convolutional regression networks. Med. Image Anal. 68, 101856.

Karimi, D., Salcudean, S.E., 2019. Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks. IEEE Trans. Med. Imaging 39 (2), 499–513.

Kermany, D., Zhang, K., Goldbaum, M., et al., 2018. Labeled optical coherence tomography (oct) and chest x-ray images for classification. Mendeley Data 2 (2), 651.

Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., Ayed, I.B., 2019. Boundary loss for highly unbalanced segmentation. In: International Conference on Medical Imaging with Deep Learning. PMLR, pp. 285–296.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Liefers, B., et al., 2021. Quantification of key retinal features in early and late age-related macular degeneration using deep learning. Am. J. Ophthalmol. 226, 1–12.

Litjens, G., et al., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Lu, D., et al., 2019. Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network. Med. Image Anal. 54, 100–110.

Meng, Q., et al., 2021. MF-Net: Multi-scale information fusion network for CNV segmentation in retinal OCT images. Front. Neurosci. 1192.

Michl, M., et al., 2022. Automated quantification of macular fluid in retinal diseases and their response to anti-VEGF therapy. Br. J. Ophthalmol. 106 (1), 113–120.

Montuoro, A., Waldstein, S.M., Gerendas, B.S., Schmidt-Erfurth, U., Bogunović, H., 2017. Joint retinal layer and fluid segmentation in OCT scans of eyes with severe macular edema using unsupervised representation and auto-context. Biomed. Opt. Express 8 (3), 1874–1888.

Müller, P.L., et al., 2021. Reliability of retinal pathology quantification in age-related macular degeneration: Implications for clinical trials and machine learning applications. Transl. Vis. Sci. Technol. 10 (3), 4.

Nair, T., Precup, D., Arnold, D.L., Arbel, T., 2020. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Med. Image Anal. 59, 101557.

Nakao, T., Hanaoka, S., Nomura, Y., Murata, M., Takenaga, T., Miki, S., Watadani, T., Yoshikawa, T., Hayashi, N., Abe, O., 2021. Unsupervised deep anomaly detection in chest radiographs. J. Dig. Imaging 34, 418–427.

Orlando, J.I., et al., 2019. U2-net: A Bayesian U-Net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological OCT scans. arXiv preprint arXiv:1901.07929.

Ouardini, K., Yang, H., Unnikrishnan, B., Romain, M., Garcin, C., Zenati, H., Campbell, J.P., Chiang, M.F., Kalpathy-Cramer, J., Chandrasekhar, V., et al., 2019. Towards practical unsupervised anomaly detection on retinal images. In: Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data: First MICCAI Workshop, DART 2019, and First International Workshop, MIL3ID 2019, Shenzhen, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13 and 17, 2019, Proceedings 1. Springer, pp. 225–234.

Pawlowski, N., Lee, M.C., Rajchl, M., McDonagh, S., Ferrante, E., Kamnitsas, K., Cooke, S., Stevenson, S., Khetani, A., Newman, T., et al., 2018. Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders.

Pekala, M., Joshi, N., Liu, T.A., Bressler, N.M., DeBuc, D.C., Burlina, P., 2019. Deep learning based retinal OCT segmentation. Comput. Biol. Med. 114, 103445.

Playout, C., Duval, R., Cheriet, F., 2019. A novel weakly supervised multitask architecture for retinal lesions segmentation on fundus images. IEEE Trans. Med. Imaging 38 (10), 2434–2444.

Quellec, G., Lamard, M., Cozic, M., Coatrieux, G., Cazuguel, G., 2016. Multiple-instance learning for anomaly detection in digital mammography. IEEE Trans. Med. Imaging 35 (7), 1604–1614.

Reinke, A., et al., 2021. Common limitations of image processing metrics: A picture story. arXiv preprint arXiv:2104.05642.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Proc. of MICCAI. Springer, pp. 234–241.

Roy, A.G., et al., 2017. ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. Biomed. Opt. Express 8 (8), 3627–3642.

Sato, D., Hanaoka, S., Nomura, Y., Takenaga, T., Miki, S., Yoshikawa, T., Hayashi, N., Abe, O., 2018. A primitive study on unsupervised anomaly detection with an autoencoder in emergency head CT volumes. In: Medical Imaging 2018: Computer-Aided Diagnosis, Vol. 10575. SPIE, pp. 388–393.

Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U., 2019. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. 54, 30–44.

Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G., 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: Proc. of IPMI. Springer, pp. 146–157.

Schlegl, T., et al., 2018. Fully automated detection and quantification of macular fluid in OCT using deep learning. Ophthalmology 125 (4), 549–558.

Schlemper, J., et al., 2019. Attention gated networks: Learning to leverage salient regions in medical images. Med. Image Anal. 53, 197–207.

Schmidt-Erfurth, U., Klimscha, S., Waldstein, S., Bogunović, H., 2017. A view of the current and future role of optical coherence tomography in the management of age-related macular degeneration. Eye 31 (1), 26–44.

Schmidt-Erfurth, U., Waldstein, S.M., 2016. A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration. Prog. Retin. Eye Res. 50, 1–24.

Schmidt-Erfurth, U., et al., 2021. AI-based monitoring of retinal fluid in disease activity and under therapy. Prog. Retin. Eye Res. 100972.

Seeböck, P., et al., 2019a. Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal OCT. IEEE Trans. Med. Imaging 39 (1), 87–98.

Seeböck, P., et al., 2019b. Unsupervised identification of disease marker candidates in retinal OCT imaging data. IEEE Trans. Med. Imaging 38 (4), 1037–1047.

Sidibé, D., et al., 2017. An anomaly detection approach for the identification of DME patients using spectral domain optical coherence tomography images. Comput. Methods Programs Biomed. 139, 109–117.

Sinha, A., Dolz, J., 2020. Multi-scale self-guided attention for medical image segmentation. IEEE J. Biomed. Health Inform. 25 (1), 121–130.
Venhuizen, F.G., et al., 2018. Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography. Biomed. Opt. Express 9 (4), 1545–1569.
Wang, J., et al., 2021. Weakly supervised anomaly segmentation in retinal OCT images using an adversarial learning approach. Biomed. Opt. Express 12 (8), 4713–4729.
Wei, Q., Ren, Y., Hou, R., Shi, B., Lo, J.Y., Carin, L., 2018. Anomaly detection for medical images based on a one-class classification. In: Medical Imaging 2018: Computer-Aided Diagnosis, Vol. 10575. SPIE, pp. 375–380.
Wolleb, J., Sandkühler, R., Cattin, P.C., 2020. Descargan: Disease-specific anomaly detection with weak supervision. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23. Springer, pp. 14–24.
Wong, W.L., et al., 2014. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. Lancet Glob. Health 2 (2), e106–e116.
Wyatt, J., Leach, A., Schmon, S.M., Willcocks, C.G., 2022. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 650–656.
Wyburd, M.K., Dinsdale, N.K., Namburete, A.I., Jenkinson, M., 2021. TEDS-Net: Enforcing diffeomorphisms in spatial transformers to guarantee topology preservation in segmentations. In: Proc. of MICCAI. Springer, pp. 250–260.
Xing, G., et al., 2022. Multi-scale pathological fluid segmentation in OCT with a novel curvature loss in convolutional neural network. IEEE Trans. Med. Imaging.
Ye, Y., Chen, X., Shi, F., Xiang, D., Pan, L., Zhu, W., 2021. Context attention-and-fusion network for multiclass retinal fluid segmentation in OCT images. In: Medical Imaging 2021: Image Processing, Vol. 11596. International Society for Optics and Photonics, 1159622.
Zar, J.H., 2005. Spearman rank correlation. In: Encyclopedia of Biostatistics. vol. 7, Wiley Online Library.
Zhang, Y., Liu, H., Hu, Q., 2021a. Transfuse: Fusing transformers and cnns for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer, pp. 14–24.
Zhang, J., Xie, Y., Pang, G., Liao, Z., Verjans, J., Li, W., Sun, Z., He, J., Li, Y., Shen, C., et al., 2020. Viral pneumonia screening on chest X-rays using confidence-aware anomaly detection. IEEE Trans. Med. Imaging 40 (3), 879–890.
Zhang, Y., et al., 2021b. 3D multi-attention guided multi-task learning network for automatic gastric tumor segmentation and lymph node classification. IEEE Trans. Med. Imaging 40 (6), 1618–1631.
Zhao, Y., Ding, Q., Zhang, X., 2022. AE-FLOW: Autoencoders with normalizing flows for medical images anomaly detection. In: The Eleventh International Conference on Learning Representations.
Zhou, K., Gao, S., Cheng, J., Gu, Z., Fu, H., Tu, Z., Yang, J., Zhao, Y., Liu, J., 2020. Sparse-gan: Sparsity-constrained generative adversarial network for anomaly detection in retinal oct image. In: 2020 IEEE 17th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1227–1231.
Zhou, Y., Sun, X., Zha, Z.-J., Zeng, W., 2019. Context-reinforced semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4046–4055.