Robust Fovea Detection in Retinal OCT Imaging Using Deep Learning

Simon Schürer-Waldheim, Philipp Seeböck, Hrvoje Bogunović, Bianca S. Gerendas, and Ursula Schmidt-Erfurth

***Abstract*—The fovea centralis is an essential landmark in the retina where the photoreceptor layer is entirely composed of cones responsible for sharp, central vision. The localization of this anatomical landmark in optical coherence tomography (OCT) volumes is important for assessing visual function correlates and treatment guidance in macular disease. In this study, the "PRE U-net" is introduced as a novel approach for a fully automated fovea centralis detection, addressing the localization as a pixel-wise regression task. 2D B-scans are sampled from each image volume and are concatenated with spatial location information to train the deep network. A total of 5586 OCT volumes from 1,541 eyes were used to train, validate and test the deep learning method. The test data is comprised of healthy subjects and patients affected by neovascular age-related macular degeneration (nAMD), diabetic macula edema (DME) and macular edema from retinal vein occlusion (RVO), covering the three major retinal diseases responsible for blindness. Our experiments demonstrate that the PRE U-net significantly outperforms state-of-the-art methods and improves the robustness of automated localization, which is of value for clinical practice.**

***Index Terms*—Age-related macular degeneration, deep learning, diabetic macula edema, fovea detection, landmark detection, optical coherence tomography, retinal vein occlusion.**

I. Introduction

THE fovea centralis is an essential anatomical landmark in human retina as it marks the spot which is responsible for central, sharp vision [1]. Macular diseases such as neovascular age-related macular degeneration (nAMD) [2], diabetic macular edema (DME) [3] and macular edema from retinal vein occlusion (RVO) [4] can lead to serious distortion of the fovea, resulting in visual impairment. The identification of the fovea centralis position is an essential step in retinal image modalities [5]–[10] to analyze and monitor macular disease progression as needed for treatment planning and decisions. The fovea has an important role in diseases as subtle changes can have high impact on vision [5]. The fovea localization is required for measuring clinically relevant information such as the retinal subfield thickness [11], [12] or the positioning of an early treatment diabetic retinopathy study (ETDRS) grid [13] on any retinal image to identify the location of retinal pathologies in relation to its distance from the fovea, meaning its influence on visual function. Therefore, the fovea centralis is a diagnostically relevant landmark that is crucial for an appropriate patient management and treatment success.

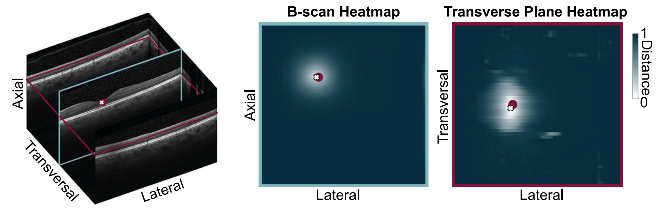

Fig. 1. Fovea localization in retinal OCT. The manual annotated fovea position is marked with a red dot, while the fovea position predicted by our model is depicted with a white asterisk. Heatmaps show the predicted distance to the fovea centralis for each pixel.

In this paper, we propose a novel deep learning based approach to automatically predict the fovea position in optical coherence tomography (OCT) scans. OCT is the current gold standard and most used imaging modality for retinal diseases [14]–[16] and enables to non-invasively acquire three-dimensional scans of the retina in micrometer resolution [17]. To obtain a volumetric scan, one-dimensional A-Scans are concatenated to form a 2D B-scan and multiple 2D B-scans comprise a 3D volume. The major advantage of using automated methods for fovea detection is its objectivity and reproducibility of the results. The labelling will always be done in the exact same manner by the algorithm, whereas the intra- and inter-observer variability need to be considered when human manual detection is performed.

Automated landmark localization in medical imaging is a challenging and an active research topic since decades [18]–[20]. While earlier approaches such as the left ventricular landmark detection algorithm by Garcia-Melendo *et al.* [18] started with handcrafted rules to solve localisation tasks, learning-based methods are currently state-of-the-art [20]–[26]. This paradigm shift can be explained by multiple factors such as the difficulty of defining effective rules and features, the increased computational power and the latest achievements of learning-based methods. The earlier approaches exploited machine learning based methods such as random forests [27]–[37] and in the most cases still depend on handcrafted features. For end-to-end

Manuscript received 5 November 2021; revised 8 March 2022; accepted 2 April 2022. Date of publication 8 April 2022; date of current version 9 August 2022. *(Corresponding author: Philipp Seeböck.)*

The authors are with the Department of Ophthalmology and Optometry, Medical University Vienna, 1090 Vienna, Austria (e-mail: simon.schuerer-waldheim@meduniwien.ac.at; philipp.seeboeck@meduniwien.ac.at; hrvoje.bogunovic@meduniwien.ac.at; bianca.gerendas@meduniwien.ac.at; ursula.schmidt-erfurth@meduniwien.ac.at).

Digital Object Identifier 10.1109/JBHI.2022.3166068

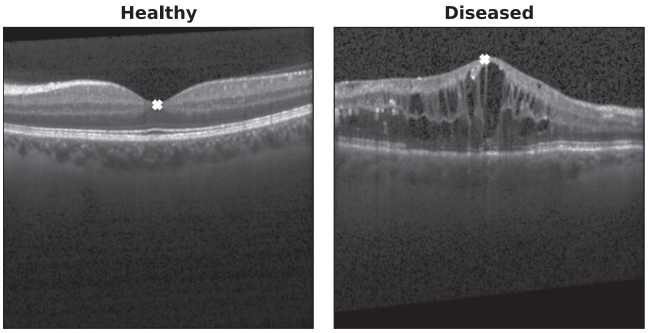

Fig. 2. Difference between healthy (left) and diseased retina (right). In healthy cases, recognizing the fovea center (white cross) as the deepest position of the fovea pit is easy, whereas in diseased cases even expert graders disagree sometimes.

learning, two different design choices exist in the context of landmark detection. Either the landmark coordinates are directly determined [38], or a map [8], [9], [20]–[23], [26] is predicted. Depending on the design of the algorithm, map values can represent pseudo-probabilities of being the target of interest [8], [9] or distances to the landmark [22]. The disadvantage of directly detecting a landmark is that no feedback explaining the prediction choice is given. In contrast, maps provide potential insights, indicating which challenges the algorithm might face.

The automated fovea detection in OCT imaging was addressed in several previous studies [5]–[10]. Most of the approaches reached promising results for healthy eyes, where the retina is showing a consistent morphology and the deepest retinal point (the "fovea pit") represents the landmark of interest. In contrast, fovea detection in diseased subjects is a much more challenging task (automated and manual) due to the highly altered and diverse appearance (Fig. 2). For instance, retinal fluid, pigment epithelial detachments or fibrosis can be found as common pathologies [39]. More importantly, different fovea types can occur [6], [10], with disorganized layer boundaries and abnormal retinal thicknesses in retinal swelling, called edema [10]. Thinning and confluence of certain retinal layers close to the fovea is a typical characteristic humans use to identify the fovea [40]. However, using this anatomical finding to design an algorithm will not always lead to reliable fovea detections as the thinning and confluence can not be observed in all diseased eyes (Fig. 2).

Wu *et al.* [6] propose to solve the fovea detection in OCT images by classifying the foveal configuration type (normal, minor and absent), performing a retinal surface segmentation algorithm and using different layer thickness maps to compute the final landmark position. Montuoro *et al.* [7] use a random forest regressor with layer thickness maps as input to predict the distance to the fovea for every A-scan. For finding the final landmark position, a random sample consensus is applied. Besides a longer run time due to the additionally required layer segmentation, both methods are prone to prediction errors that are caused by imprecise layer segmentations in difficult cases. Furthermore, important context information might exist that is not included in layer thickness maps. Liefers *et al.* [8], [9] introduced two slightly different fully convolutional network architectures that are trained with two-dimensional image patches to solve the landmark detection task. By solving the classification task 'fovea vs. non-fovea,' a landmark likelihood value is predicted for each pixel. The position with the maximum value in the OCT is then used as final landmark position. In contrast, our approach tackles the problem as a pixel-wise regression task, uses full B-scans instead of patches with limited spatial context and utilizes a spatial map as additional input. Li's *et al.*'s [10] approach is based on creating OCT projection maps, using two slightly differently trained lightweight U-net instances to classify the foveal avascular zone (FAZ), a vessel free central retinal area always containing the fovea. By calculating the geometric center of the detected FAZ area the fovea position is obtained. The fast prediction speed of this method, by reducing the 3D OCT volume into a 2D projection map, comes at the expense of: (1) The prediction accuracy depends on the quality of projections, and (2) the 3D high-resolution information is not exploited.

Our proposed approach, which we refer to as prior regularization U-net (*PRE U-net*), addresses previous limitations by tackling the landmark detection task as a distance regression problem for each pixel (Fig. 1). We propose to use a spatial location prior, which is created by using the anatomical pixel information and the volume size to generate a second input for the artificial network. Hence, a three-dimensional spatial context is provided to the model.

The main contributions of this work are fourfold: (1) We addressed the fovea detection as a pixel-wise regression task. (2) Our novel introduced sampling strategy works with full B-scans and does not require using full volumes during training time which saves memory. (3) We introduce a distance map transformation procedure for the target map creation, considering the anatomical pixel size and the network optimization process. (4) We propose the *PRE U-net* architecture that uses the spatial location prior as an additional input map to boost the performance.

II. Methods

In this study, the landmark detection task is addressed as pixel-wise regression task. The model f_θ is trained to predict the distance to the fovea centralis for every pixel of the input image. The pixel with the lowest distance then indicates the predicted location of the fovea. An overview of the proposed approach is shown in Fig. 3.

a) Training and Sampling Strategy: Let χ_n be a dataset of n OCT volumes. For each volume X_i, the model f_θ aims at finding the function $f_\theta : X_i, Z_i \rightarrow V_i$ by optimizing its weights θ, where Z_i is the spatial location prior volume (Section II-B) and V_i the target distance volume (Section II-C). Z_i and V_i have the same size as X_i. For training the *PRE U-Net*, B-scans x_i are sampled from each OCT volume X_i using the following strategy: For each volume, the target B-scan (where the landmark was annotated) and a randomly selected non-target B-scan are extracted. A hyper-parameter d is used to exclude B-scans from the non-target selection that are located less than d planes away

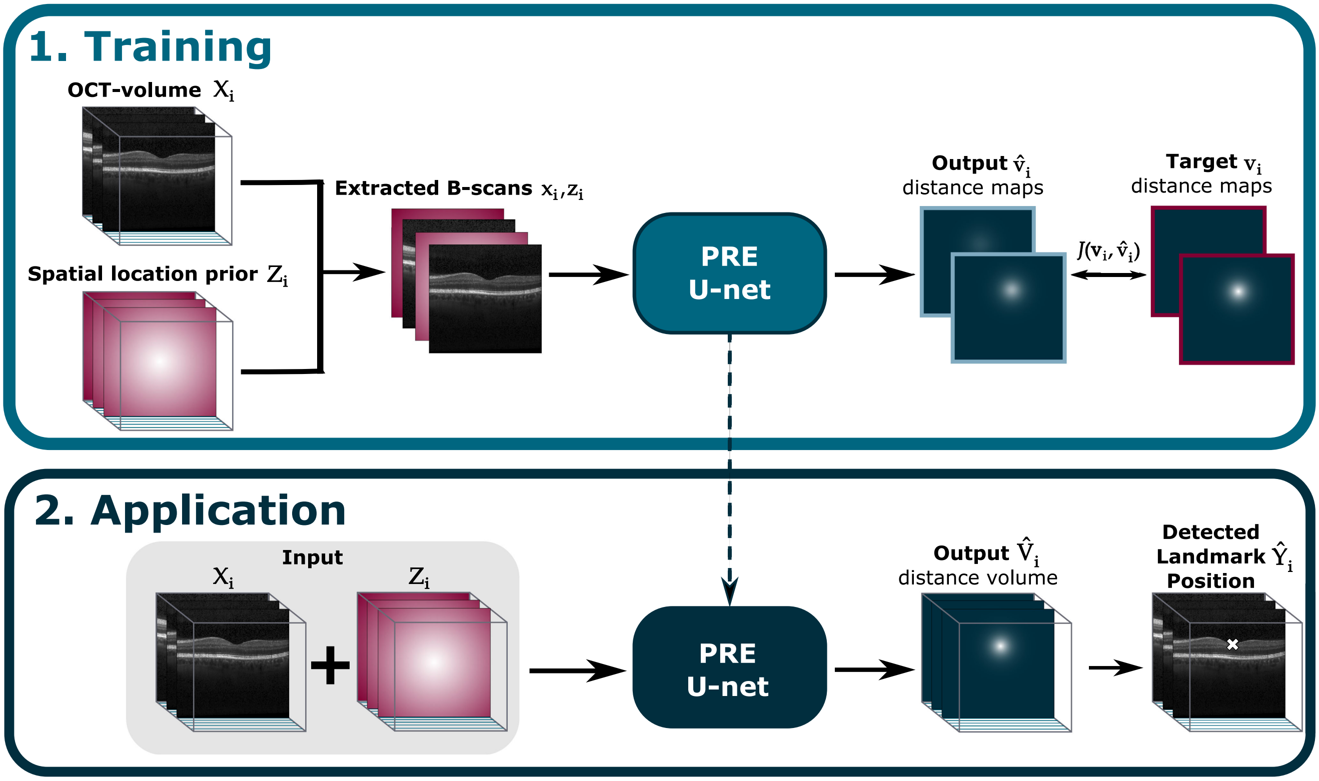
1. Training
OCT-volume X_i
Spatial location prior Z_i
Extracted B-scans x_i, z_i
PRE U-net
Output $\hat{v}_i$ distance maps
$J(v_i, \hat{v}_i)$
Target v_i distance maps
2. Application
Input
X_i
Z_i
+
PRE U-net
Output $\hat{V}_i$ distance volume
Detected Landmark $\hat{Y}_i$ Position

Fig. 3. Overview of the proposed approach. During training, the model learns to predict a fovea distance map, taking the OCT B-scan and the spatial location prior as input. The final model is then applied B-scan wise to a new OCT volume during test time, obtaining a fovea distance volume. The pixel with the lowest predicted distance is then used as fovea position estimate.

from the target B-scan. The corresponding spatial location maps $z_i \in Z_i$ of the B-scans are fed as an additional input to the network. The model is optimized by minimizing the following loss:

$$J(\hat{v}_i, v_i) = \frac{1}{n}\sum_{j=1}^{n} |\hat{v}_{i_j} - v_{i_j}| \tag{1}$$

where $\hat{v}_i$ are the output distance maps, v_i are the corresponding target distance maps and n is the number of pixels in these maps.

b) Application: At test time, the trained *PRE U-Net* is applied B-scan wise to a new OCT volume. By taking both the original OCT scan and the corresponding spatial location prior as input, the model predicts a fovea distance volume $\hat{V}_i$. As postprocessing-step a three-dimensional Gaussian smoothing operation is applied as convolution to $\hat{V}$, resulting in a smoothed output volume $\hat{V}^*$ of the same size as the input. The discrete Gaussian kernel is created by selecting the size of the kernel and filling the content according to:

$$G(x, y, z) = \frac{1}{\sqrt{2\pi}^3\,\sigma_1\sigma_2\sigma_3} e^{-\frac{x^2}{2\sigma_1^2}-\frac{y^2}{2\sigma_2^2}-\frac{z^2}{2\sigma_3^2}} \tag{2}$$

where $\sigma_1, \sigma_2, \sigma_3$ are determining the smoothness of the output for the dimensions x, y, z. The minimum value of the smoothed output $\hat{V}^*$ then indicates the detected fovea position $Y_i(y_1, y_2, y_3)$.

A. Spatial Location Prior

With the spatial location prior volume Z, additional three-dimensional spatial context information is provided to the model. To generate Z no additional labels are needed. The spatial location prior volume $Z_i \in \mathbb{R}^{a\times b\times c}$ for an OCT volume $X_i \in \mathbb{R}^{a\times b\times c}$ with an anatomical pixel size of $\gamma_1 \times \gamma_2 \times \gamma_3$ is created by:

$$\begin{aligned} Z_i &: Z_{i_{j_1,j_2,j_3}} \\ &= \sqrt{(\eta_1(j_1 - c_1))^2 + \eta_2(j_2 - c_2))^2 + \eta_3(j_3 - c_3))^2} \end{aligned} \tag{3}$$

where c_1, c_2, c_3 are the coordinates of the 3D volume center C_i, j_1, j_2, j_3 are the pixel positions of Z_i and η_1, η_2, η_3 are normalization hyper-parameters (axial, lateral and transversal). The anatomical pixel size is used to set the normalization parameters: $\eta_1, \eta_2 = \gamma_2$ and $\eta_3 = \gamma_3$.

B. Target Map Creation

The target distance volume V is used as ground truth map during training. To generate V, the annotated landmark position and the anatomical voxel size is needed. First, a corresponding ground truth distance volume $D_i \in \mathbb{R}^+$ of the same size as X_i is created by:

$$\begin{aligned} D_i &: D_{i_{j_1,j_2,j_3}} \\ &= \sqrt{(\zeta_1(j_1 - y_1))^2 + \zeta_2(j_2 - y_2))^2 + \zeta_3(j_3 - y_3))^2} \end{aligned} \tag{4}$$

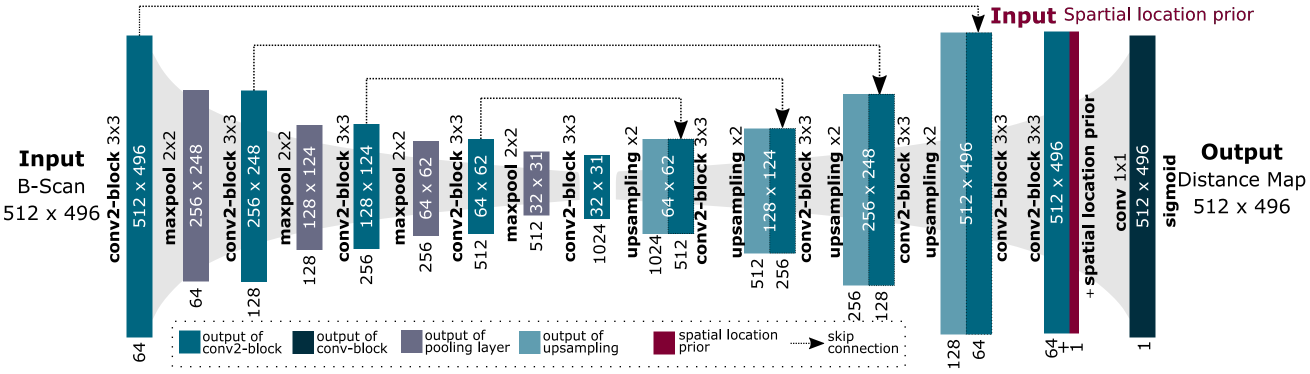

Fig. 4. U-net based architecture of the proposed method (*PRE U-net*) with five levels of depth, using the spatial location prior (red) as an additional feature channel.

where y_1, y_2, y_3 are the coordinates of the ground truth landmark Y_i, j_1, j_2, j_3 are the pixel positions of D_i and $\zeta_1, \zeta_2, \zeta_3$ are normalization hyper-parameters($\zeta_1, \zeta_2 = \gamma_2$, $\zeta_3 = \gamma_3$). A logistic function $g(.)$ with an offset term is used to transform D_i into the target map $V_i \in \mathbb{R}^{+[0,1]}$:

$$V_i = g(D_i) = \frac{L}{1 + e^{-k(D_i - x_0)}} - 0.5L \tag{5}$$

with hyper-parameters L, k and x_0. $L/2$ determines the maximum output value, k the steepness of the curve and x_0 defines the horizontal translation of the function. The offset term $-0.5\ L$ is used to set the constraint that a perfect prediction has a distance of 0.

C. Architecture

In this study, a U-net [41] based architecture is used as a backbone (Fig. 4). The encoding and decoding part are connected by skip connections. The *conv2-block* used is a double sequence of a zero-padded convolution operation (conv) with a kernel size of 3 × 3, followed by a batch norm layer [42] and a leaky rectified linear unit (leaky ReLU) [43], [44]. The leaky ReLU is used instead of the ReLU to address the dying ReLU problem [43]. 2 × 2 max pooling operations reduce the spatial resolution in the encoding part. Bilinear upsampling operations with a scale factor of 2 regain the spatial resolution. After each upsampling-block the feature maps are concatenated (skip connection) and fed to a conv2-block. Our network has five levels of depth, with 64, 128, 256, 512 and 1024 output channels each. The spatial location prior is added in form of a feature channel just before the last 1 × 1 convolution operation (Fig. 4, red block). Finally, a sigmoid activation function is applied in order to predict distance values between 0 and 1.

III. Experiment Setup

A. Materials

This work was conducted in adherence to the tenets of the Declaration of Helsiniki, and ethics approval was obtained by the Ethics Committee of the Medical University of Vienna Submission Nr 1246/2016. A total number of 5,586 OCT volumes obtained with a Spectralis OCT device (Heidelberg Engineering, Heidelberg, Germany) from 1,541 eyes of 1,541 different subjects were used for development and evaluation (Table I). The volumes were acquired with voxel dimensions of 496 × 512 × 49. The volume covers an anatomical area of approximately 2 × 6 × 6 mm, resulting in an voxel size of about $\gamma_1 = 0.004$, $\gamma_2 = 0.012$ and $\gamma_3 = 0.122$.

TABLE I
Data Setup Used Including the Distribution Over the Four Groups: Healthy, Wet AMD, DME and RVO

Volumes/Subjects	Training	Validation	Test	Altogether
Healthy	144/144	34/34	100/100	278/278
wet AMD	1750/336	34/34	100/100	1884/470
DME	177/177	34/34	100/100	311/311
RVO	2979/348	34/34	100/100	3113/482
Sum	5050/1005	136/136	400/400	5586/1541

The dataset consists of 278 healthy volumes, 1884 volumes with neovascular age-related macular degeneration (AMD), 311 volumes with diabetic macular edema (DME) and 3113 volumes with retinal vein occlusion (RVO). The data was randomly split on a patient-distinct basis into training ($n_{subjects} = 1005$), validation ($n_{subjects} = 136$) and test set ($n_{subjects} = 400$), as depicted in Table I. Up to 13 volumes per eye were available, representing different stages of treatment and disease severity. In order to obtain a challenging evaluation set, only treatment-naive volumes (showing most retinal swelling and most retinal layer disorganization, thus a high level of foveal architechture distortion) were used in the validation and test set. The fovea position was manually labelled according to a standardized and predefined process for all volumes by certified expert graders (gold standard for manual fovea positioning) of the Vienna Reading Center (VRC) at the Medical University of Vienna.

B. Training Details

The OCT voxel values were rescaled to [0,1]. In the experiments of this work we used $L = 2$, $k = 3.5$ and $x_0 = 0$ as hyper-parameters to create the target maps. L=2 was chosen to receive distance values in the range of 0 and 1. k was empirically determined by taking into account the area which can be identified as close to the fovea center by humans, intuitively reflecting the increasing difficulty of estimating the correct regression value with increasing distance to the fovea center (Appendix I, Fig. 9). x_0 was set to 0 since a horizontal translation of the transformation function g(.) was not intended. The sampling hyper-parameter d was set to 3, in order to avoid sampling of non-target B-scans that are close to the target B-scan and would therefore represent ambiguous samples. In other words, B-scans further away than d from the fovea center should be unambiguous and identifiable as non-targets. The network was trained using a learning rate of 0.0001 and Adam optimizer [45]. Kaiming [46] was picked as network-weights initialization method. A mini-batch size of 8 was chosen, containing the target and a non-target B-scan from 4 different volumes each. The model was trained for 8 epochs corresponding to 10,104 iterations. Data augmentation with horizontal flipping, rotation with $\pm 10°$as well as translation with $\pm 5\%$ laterally and $\pm 15\%$ axially was applied during training at a chance of 50%. Moreover, the B-scans were elastically deformed ($\alpha = 0.05, \sigma = 0.75$) at a chance of 10% and Gaussian noise (μ=0, σ=0.25) was added at a chance of 5%. During test time, the post-processing parameters for a $3 \times 5 \times 5$ Gaussian smoothing kernel were set to $\sigma_1 = 0.33, \sigma_2 = 0.45$ and $\sigma_3 = 0.42$. The deep-learning models were developed using PyTorch 1.0.0 [47] and Python 3.6 (Python Software Foundation, Delaware, USA).

C. Evaluation

In our experiments, we evaluated (1) the performance of our model in comparison with a heuristic that has been used in previous work [8], two state-of-the-art deep learning approaches [8], [10] and (2) the contribution of each of the individual components of our proposed approach in the final results. For quantitative evaluation, the Euclidean distance between the predicted and the manually annotated fovea location was calculated. Following the recommended evaluation strategy of previous work [8], [9], we only evaluated the distance in the transversal and lateral dimension. A distance-threshold of 0.750 mm to the manual location is used for determining the number of outliers as this value corresponds to the actual fovea size of about 1.5 mm in diameter [48]. For a more detailed evaluation, we also used two more restrictive distance-thresholds that are 1) the size corresponding to the foveolar 0.175 mm [48] and 2) the distance used by the VRC to detect outliers. According to this definition, the difference between detected and manually annotated positions must lie within a range of 1 B-scan and 5 A-scans (euclidean distance of 0.135 mm) to call the distance as clinically acceptable. This cutoff was the most stringent cutoff we used for evaluation of the distance between the detected and the manually annotated position, claiming that if the detected distance is in this range it can be called clinically acceptable.

The one-sided Wilcoxon signed-rank test [49] was conducted at a significance level of α=2.5% to test if the performance of our proposed method is statistically significantly better than the baseline approaches: (1) *FCNet* [8] proposing a fully convolutional architecture, (2) *FAZNet* [10] segmenting the FAZ region in projection maps and (3) the heuristic of taking the image center position as fovea center (*ICP*). In general, two main locations of OCT acquisition exist. Depending on the clinical location of interest, OCTs are approximately centered either in the macula/fovea, which is the case in our dataset, or in the optic nerve head. Using the center of the scan as an estimate for the fovea position is therefore a straightforward baseline that does not involve any training and has also been used previously [8].

We conducted a series of ablation experiments to evaluate the importance of each individual component of our proposed method. The one-sided Wilcoxon signed-rank test [49] was used (α=2.5%) to test if the performance of our proposed method is better than the performed ablations. It is important to note that hyper-parameter tuning was based on the performance in the validation set. The test set was only used to demonstrate the impact of the model components on the results.

- **A**- *Input information:* The spatial location prior map was removed from the proposed approach.
- **B**- *Target map:* Three alternative target maps were evaluated separately:

– The normalized distance map D_i.
– V_i with $\zeta_1 = \gamma_1$ as normalization hyper-parameter.
– V_i with pixel distance map D_i $(\zeta_1, \zeta_2, \zeta_3 = 1)$.

Note that Z_i and D_i are always normalized in the same manner.

- **C**- *Architecture:* The U-net based backbone architecture was replaced by the network architecture proposed by Liefers *et al.* [8]. However, the idea of using a spatial location prior as an additional input channel was added. We refer to this adaptation as *centered FCNet*.
- **D**- *Sampling strategy:* Two alternative sampling strategies were tested separately:

– The number of selected non-target B-scans per volume was increased from one to two, three and four.
– The non-target B-scan selection constraint ($d = 3$) was omitted.

- **E**-- *Loss function:* The L1-loss was replaced with the L2-loss function.
- **F**-- *Data augmentation:* A *PRE U-net* instance was trained without applying data augmentation.
- **G**-- *Model:* A DeepLabV3+ [50] based backbone was used without spatial location prior, but with the proposed target map creation process.

IV. Results

A. Quantitative and Qualitative Evaluation

Quantitative results for fovea detection are shown in Table II, including results for our proposed method *PRE U-Net* and the three baseline approaches FCNet [8], FAZNet [10] and *ICP*. The proposed algorithm outperforms the three baseline methods in all metrics on the test set. In particular, the statistical tests reveal that *ICP* ($p = 5.1 \cdot 10^{-10}$), *FAZNet* ($p = 7.2 \cdot 10^{-9}$) and *FCNet* ($p = 2.3 \cdot 10^{-6}$) are significantly outperformed by our method on the test set. Boxplots illustrating the ‘distance to ground truth’-distribution in the test set are shown in Fig. 5. The results demonstrate that for the proposed method the number of outliers was reduced compared to the state-of-the-art deep

TABLE II
QUANTITATIVE PERFORMANCE ON THE TEST SET (SAMPLE SIZE = 400)

Method	Distances		# Outliers			#Significance
	Mean (± Std)	Median	VRC-Criteria	Fovevola	Fovea	p-value
ICP	0.223 (± 0.191)	0.168	252 (63.00%)	190 (47.50%)	9 (2.25%)	$5.1 \cdot 10^{-10}$
FAZNet [10] *	0.244 (± 0.236)	0.175	233 (61.15%)	190 (49.87%)	24 (6.30%)	$7.2 \cdot 10^{-9}$
FCNet [8]	0.261 (± 0.426)	0.128	173 (43.25%)	133 (33.25%)	30 (7.50%)	$2.3 \cdot 10^{-6}$
PRE U-net (ours)	**0.169 (± 0.159)**	**0.122**	**138 (34.50%)**	**102 (25.50%)**	**8 (2.00%)**	-

*19 samples had to be removed due to insufficient segmentation results.

The mean, standard deviation, median and the number of outliers is shown for each method. The distances results (mean, std and median) are given in mm. The number of outliers are provided for the distance-thresholds of 0.135 mm (VRC-criteria), 0.175 mm (fovevola) and 0.750 mm (fovea). The p-values of one-sided wilcoxon signed-rank tests, comparing our proposed approach with the others, are presented.

TABLE III
QUANTITATIVE RESULTS OF THE ABLATION EXPERIMENTS, EVALUATED ON THE TEST SET

Method	Distances		# Outliers			# Significance
	Mean (± Std)	Median	VRC-Criteria	Fovevola	Fovea	p-value
PRE U-Net	**0.169(±0.159)**	**0.122**	138 (34.50%)	**102 (25.50%)**	8 (2.00%)	-
A: no spatial prior	0.181(±0.209)	0.122	143 (35.75%)	109 (27.25%)	8 (2.00%)	0.0139
B1: D_i as target map	0.189(±0.210)	0.128	178 (44.50%)	119 (29.75%)	10 (2.50%)	$5.2 \cdot 10^{-7}$
B2: V_i with $\zeta_1 = 0.004$	0.267(±0.500)	0.125	162 (40.5%)	142 (35.50%)	20 (5.00%)	$2.3 \cdot 10^{-5}$
B3: V_i with $\zeta_1, \zeta_2, \zeta_3 = 1$	0.196(±0.240)	0.128	179 (44.75%)	141 (35.25%)	11 (2.75%)	0.0008
C: Architecture: centered FCNet	0.206(±0.289)	0.126	153 (38.25%)	112 (28.00%)	20 (5.00%)	0.0167
D1.1: Sample two non-target B-scans	0.191(±0.216)	0.128	167 (41.75%)	123 (30.75%)	8 (2.00%)	$2.1 \cdot 10^{-5}$
D1.2: Sample three non-target B-scans	0.183(±0.255)	0.122	136 (34.00%)	110 (27.50%)	10 (2.50%)	0.3226
D1.3: Sample four non-target B-scans	0.179(±0.187)	0.125	152 (38.00%)	119 (29.75%)	9 (2.25%)	0.0017
D2: Lift selection limitation	0.170(±0.159)	0.122	**135 (33.75%)**	110 (27.50%)	**6 (1.50%)**	0.0645
E: L2-Loss	0.179(±0.195)	0.122	153 (38.25%)	115 (28.75%)	7 (1.75%)	0.0087
F: w/o augmentation	0.188(±0.246)	0.122	131 (32.75%)	109 (27.25%)	11 (2.75%)	0.0215
G: Architecture: DeepLabV3+	0.182(±0.159)	0.128	164 (41.0%)	133 (33.25%)	**6 (1.50%)**	$3.8 \cdot 10^{-6}$

The distance values are given in mm. The number of outliers are provided for the distance-threshold of 0.135 mm (VRC-criteria), 0.175 mm (fovevola) and 0.750 mm (fovea). The p-values of one-sided wilcoxon signed-rank tests, comparing our proposed approach with other ablations, are presented. With smaller p-values the likeliness of ablations being as good as or better than the proposed approach declines.

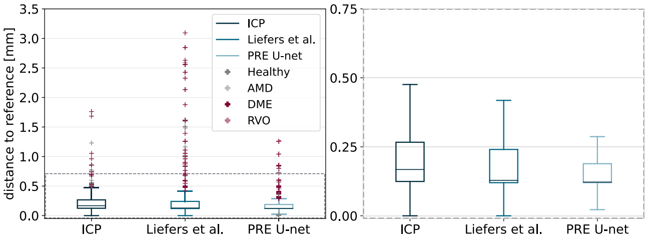

Fig. 5. Boxplots depicting the performance of the three different fovea position estimation methods (Image center position (ICP), FCNet [8]) and *PRE U-net*) on the test set. The boxplot on the right-hand side shows a zoomed-in version.

learning method [8] from 30 to 8, corresponding to a relative decrease of 73.33%. In 392 out of 400 volumes in the test set the distance error of our approach is within the actual fovea size corresponding to a success rate of 98%. For 298 out of 400 (75%) predictions we observed a prediction error lower than 0.175 mm, meaning that the predictions are within the foveola area [48]. For 262 out of 400 (66%) we observed a prediction error lower than 0.135 mm, meaning that the predictions are within the VRC criteria. Notably, the *PRE U-Net* clearly outperforms both baseline approaches also for these more stringent outlier cut-off value criteria. A table containing the quantitative metrics after excluding the outliers can be found in the Appendix I . (Table IV) as well as the training and validation loss curves of the three implemented deep learning approaches (Fig. 8).

Fig. 6 depicts boxplots of the distance between predicted fovea positions and manual fovea annotations, separated by diseases. In line with results described above, the proposed approach *PRE U-Net* outperforms both baselines in all patient groups and shows a reduced number of outliers.

Analyzing the outliers, of the 8 volumes which were not detected within 0.750 mm by our method, most occurred in the RVO subset (Fig. 6): 0 Healthy, 1 wet AMD, 2 DME, 5 RVO. Similarly, for the two baseline approaches, the RVO subset produced the most outliers (ICP: 0 Healthy, 1 wet AMD, 2 DME, 6 RVO; FCNet: 1 Healthy, 7 wet AMD, 8 DME, 14 RVO). Since ICP reflects the distance distribution of manual

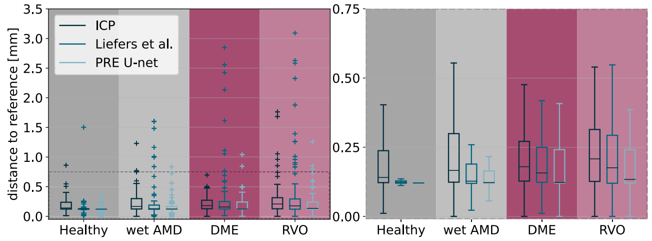

Fig. 6. Boxplots illustrating the results of all three methods on the test set, separated by each data class (Healthy, AMD, DME and RVO). The boxplot on the right-hand side shows a zoomed-in version.

fovea annotations compared to the center of the scan, Fig. 6 illustrates that the fovea was the most off-centered in the RVO dataset.

Qualitative results of the *PRE U-net* on the test set are shown in Fig. 7. We empirically observed particularly smooth transverse plane prediction heatmaps for healthy cases. Moreover, results with low prediction errors (<0.025 mm) were not limited to the healthy patient group. While eight outliers (>0.750 mm) occurred in subjects with non-healthy morphologies, in most cases the algorithm can achieve localizations close to the labelled positions, even in those with highly altered appearance (Fig. 7).

B. Ablation Study

The conducted experiments revealed that all ablations resulted in a decline of performance (Table III). The mean performance value drops more than the median value. Isotropic normalization of the target map (axial, lateral) is an important measure, indicated by the performance drop in B2 (Table III). Besides the normalization of the target map, the architecture has the highest impact on the landmark detection performance. Even the 'centered FCNet' with a spatial location prior added to the method of Liefers *et al.* [8], shows a reduced performance with an increased number of outliers (20) (C). Lifting the normalization of D_i and Z_i, which is the same as setting all corresponding hyper-parameters to one, results in a 4.9% higher median distance (B3). Without the transformation of target map distance values, described in section II-C, the mean distance increased by 11.8% and the std by 32.1% (B1). The non-target B-scan selection limitation has the smallest effect, resulting in a mean value only 0.001 mm lower (D2).

V. Discussion

Automated detection of the fovea in OCT images is an essential task for quantifying, analyzing and monitoring retinal diseases. Automated detection has the potential to overcome several limitations of the manual annotation process such as the dependence on human experts and reader variability due to subjectivity. At the same time, however, the performance and robustness of current approaches is affected by various challenges. For instance, the wide variety of appearances in retinal OCTs due to retinal disease make identification of the fovea sometime very hard, even for trained humans with many years of experience.

In this work, we propose a new approach for fovea detection in retinal OCT images. Our method *PRE U-net* clearly outperforms previous methods, particularly in terms of robustness. Our method is based on a U-net as backbone and takes B-scans as input to predict a distance value to the landmark for each pixel of an OCT volume. We believe that the achieved robustness gain can be explained by two main contributions. First, our approach was trained on full B-scans using an encoder-decoder architecture, allowing to use more information for each distance prediction than e.g. patch-based algorithms. In addition, using B-scans instead of 3D volumes as input mitigates problems when dealing with a low or varying number of B-scans, which is a common situation in clinical practice. Second, a spatial location prior was added as additional input to the network, providing three-dimensional spatial context without the need for a 3D model with all its expenses. Based on the image information the model might identify more than one landmark position candidate (i.e. pixels with a distance prediction close to zero). In this case, the spatial location prior can be used by the network as additional guidance, putting a candidate in context to its location in the volume. In general, we hypothesize that the model learns how to profit the most from the spatial location prior, depending on the relation between input image characteristics and specific spatial locations. The positioning of the spatial location prior is of importance. We experimented with feeding the prior channel together with the image information to the network which was not leading to reasonable landmark predictions. However, the proposed placement was working satisfyingly. We hypothesize that considering the abstraction level is important for combining image and prior information in deep learning models. We believe that using the spatial location prior in an optimal way requires the network to already have a high level of abstraction and context as is the case for the features in the last convolution. Even though out of scope for this work, future studies might focus on a more detailed investigation of combining different input types such as priors or constraints with image data.

Many studies have been published on landmark detection problems in medical imaging [20]. However, the automated fovea detection task in diseased cases is particularly difficult. While basic anatomical structures in the surrounding area of a landmark usually look similar, retinal diseases cause a big variety of fovea shapes and appearances, making rule-based approaches and handcrafted features tough to find. For learning based algorithms, the idea of using additional spatial information is not new. Multiple works have already shown performance improvements for landmark predictions by combining image and spatial information [20], [22], [34], [51]. However, to the best of our knowledge this is the first work applying this idea to the task of fovea detection in OCT images. Moreover, most approaches depend on the availability of multiple landmarks that are related to each other, which is not transferable to our use case. Other deep learning based approaches such as the atlas location autocontext algorithm by O'Neil *et al.* [51] use a two-stage algorithm to identify landmark locations. In contrast, our proposed method adds the spatial location prior directly as a feature map to the architecture, simplifying the model complexity.

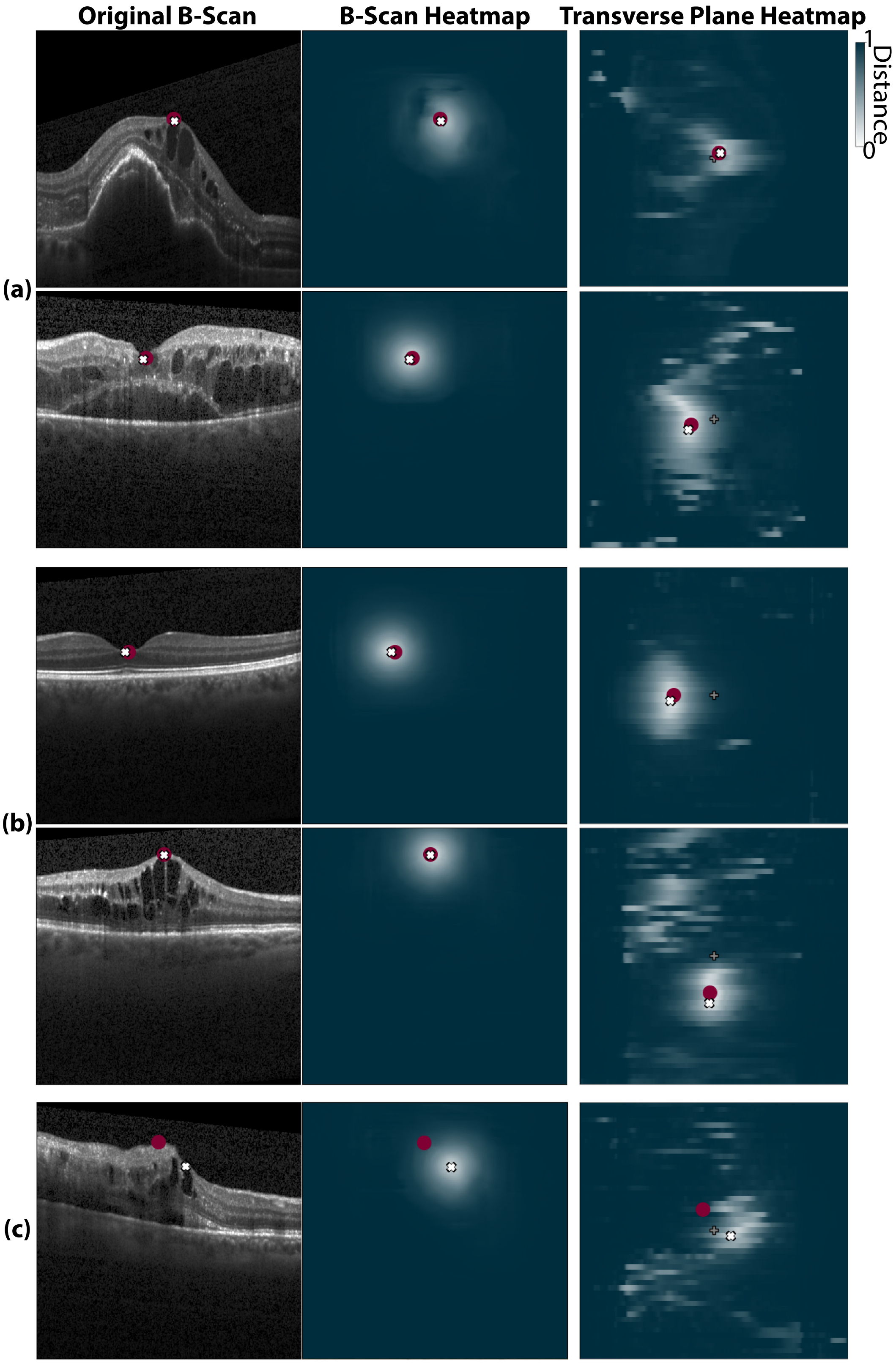

Fig. 7. Qualitative results of the *PRE U-net* on the test set. Two exemplary samples with (a) low (<0.025 mm) and (b) medium (0.120 mm - 0.189 mm) error as well as an (c) outlier (>0.750 mm) are illustrated. The red circles display the location of the ground truth labeling, the white crosses show the automated detected positions (*PRE U-net*) and the gray pluses mark the image center position (ICP). The original B-scans, their heatmaps and the transverse plane heatmaps are shown on the left, middle and right, respectively. The detected landmark position determines the presented B-scan and transverse planes.

The baselines used were all significantly outperformed by our approach. We think that using the center-point of the image as estimation of the fovea position should not be used in clinical practice nor for placing the ETDRS grids on retinal images. Nevertheless, it has often been used in large-scale retrospective, population-level analysis of ETDRS-grid derived parameters due to an absence of automated fovea localization method, and it helps to put the experiments and quantitative results in context. Our implementation of Liefers' *et al.* [8] method is in line with their reported results. However, RVO cases have not been part of their evaluation, which seem to be particularly challenging. While outliers were reported, but excluded for the quantitative statistics in their study, we included them. In clinical practice, a detection method without the need of an additional outlier monitoring process is desirable. Including outliers in the calculation, our results show that the mean value of 0.261 mm of the FCNet is even higher than the ICP baseline (0.223 mm), while this is not the case for our proposed approach (0.169 mm). The issue with the FAZNet approach by Li *et al.* [10] for our data can be explained by the lower projection map quality resulting from the lower B-scan resolution (49 B-Scans instead of 200). Other factors might be the distribution of diseased cases and the vendor type.

The ablation experiments confirmed that all evaluated parts of our approach contributed to the landmark detection success. The ablations revealed higher impact on the mean value than on the median value, indicating the positive effect of the proposed concepts on robustness. The results revealed the importance of using target maps with symmetrical distance expansion in axial and lateral direction during the training procedure (B2). We hypothesize that isotropically normalized target maps lead to a simpler representation of the landmark detection problem and therefore help to improve the performance. Increasing the number of sampled non-target B-scans (D1) per volume does not improve the results and even requires more resources during training than only sampling a single one. Not considering the anatomical pixel size (B3) for the generation of Z_i leads to a performance drop, supporting the significance of our proposed spatial location prior creation procedure. We implemented the DeepLabV3+ as a backbone for our approach (Table III, ablation experiment G). The spatial location prior was not added to the DeepLabV3+ as the positioning of the prior is an open question for this architecture. Nevertheless, DeepLabV3+ and the U-Net architecture can be compared without spatial location prior (ablation experiments: G vs. A) which we think is better suited to evaluate the choice of the backbone architecture. Both mean and median performance dropped for G, supporting the usage of the U-Net architecture.

We also noticed that even though we used different amount of data for the four data groups (Healthy, wet AMD, DME, RVO) during training time, the performance of the model was not affected by this class imbalance. From a clinical perspective, DME and RVO are very similar retinal disease which could explain the results on DME although the dataset was small. Another explanation might be that the network finds features that are independent from the diseases. This might be supported by the sampling strategy which aims at teaching the network to differentiate between target and non-target B-scans.

A potential limitation of the PRE U-Net is that the network operates on two-dimensional B-scan images while solving a three-dimensional localization task. While this means that our model is to some extent independent of the number of B-scans and their anatomical distance, it may miss potentially important context in the third dimension. However, we proposed the spatial location prior to address the missing context issue in the third dimension. Moreover, having a B-scan-based approach reduces the memory-need of the deep learning model.

The idea of the spatial location prior is not limited to the proposed task, but might be of interest for every task where a landmark is expected to be in a certain image area. In those cases, the center of the landmark expectancy range can be used to create the spatial location prior, not limited to the image center coordinates. Theoretically, the prior can be created for arbitrary dimension size, even though we only evaluated it on three dimensional data.

Taking the high variability between different manual annotations into account, where two trained humans, that result in comparable central subfield thickness values in standardized reading center settings, in mean show a distance between two foveas of 45 μm (unpublished results, VRC), having 66% within the VRC Criteria (which is the most stringent and standardized criteria available) is comparable to human annotations. These criteria are most likely more stringent than clinicians in clinical routine and as such useful. The algorithm is a good example how to improve the visualization for ophthalmologist in routine. Many OCT evaluation reports used in the clinic are dependent on the fovea position (e.g. central retinal thickness in the foveal and parafoveal area) that is set in most devices just to the center of the OCT image. The clinician has the possibility to manully re-set the foveal position at his own discretion, which many clinicians do not use. Thus, if the PRE U-net approach would pre-align the OCT analyses for the clinical reports better than the center of the scan, this means already a large improvement to the current clinical standard.

VI. Conclusion

We proposed the *PRE U-Net* for fully automated fovea detection in OCT volumes, addressing the task as a pixel-wise regression problem. The spatial location prior and a novel target map creation procedure were presented as measures to improve the performance and robustness. The results demonstrate that the *PRE U-net* significantly outperforms the current state-of-the-art for fovea detection in OCT volumes. Moreover, the proposed approach clearly reduced the number of outliers, which is of particular relevance in clinical practice. We are convinced that using such a model in clinical practice would be more than useful compared to the current clinical standards. Future fovea detection studies might also focus on three-dimensional approaches to include all relevant image context information.

APPENDIX I
SUPPLEMENTARY MATERIAL

TABLE IV
QUANTITATIVE PERFORMANCE ON THE TEST SET (SAMPLE SIZE = 400) WITHOUT OUTLIERS (> 0.750 MM)

Method	Distances	
	Mean (± Std)	Median
ICP	0.204 (± 0.128)	0.164
FAZNet [10] *	0.195 (± 0.139)	0.159
FCNet [8]	0.157 (± 0.108)	0.127
PRE U-net (ours)	**0.153 (± 0.108)**	**0.122**

*19 samples had to be removed due to insufficient segmentation results.
The mean, standard deviation and median is given in mm for each method.

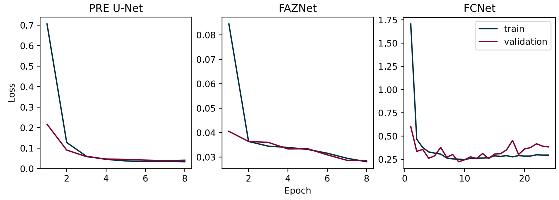

Fig. 8. Training curves of the three deep learning approaches.

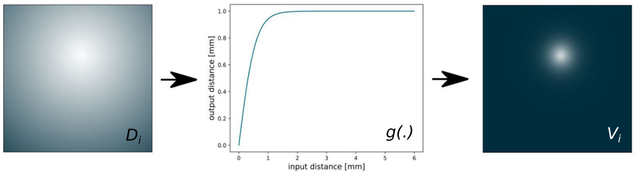

Fig. 9. Visualization of target map distance transformation.

REFERENCES

[1] G. D. Hildebrand and A. R. Fielder, "Anatomy and physiology of the retina," in *Pediatric Retina*, Berlin, Heidelberg: Springer, 2011, pp. 39–65.

[2] L. S. Lim, P. Mitchell, J. M. Seddon, F. G. Holz, and T. Y. Wong, "Age-related macular degeneration," *Lancet*, vol. 379, no. 9827, pp. 1728–1738, 2012.

[3] N. Bhagat, R. A. Grigorian, A. Tutela, and M. A. Zarbin, "Diabetic macular edema: Pathogenesis and treatment," *Surv. Ophthalmol.*, vol. 54, no. 1, pp. 1–32, 2009.

[4] N. Karia, "Retinal vein occlusion: Pathophysiology and treatment options," *Clin. Ophthalmol.*, Auckland, New Zealand, vol. 4, pp. 809–816, 2010.

[5] F. Wang, G. Gregori, P. J. Rosenfeld, B. J. Lujan, M. K. Durbin, and H. Bagherinia, "Automated detection of the foveal center improves SD-OCT measurements of central retinal thickness," *Ophthalmic Surg., Lasers Imag. Retina*, vol. 43, no. 6, pp. S32–S37, 2012.

[6] J. Wu, S. M. Waldstein, A. Montuoro, B. S. Gerendas, G. Langs, and U. Schmidt-Erfurth, "Automated fovea detection in spectral domain optical coherence tomography scans of exudative macular disease," *Int. J. Biomed. Imag.*, vol. 2016, 2016, Art. no. 7468953.

[7] A. Montuoro, S. M. Waldstein, B. S. Gerendas, U. Schmidt-Erfurth, and H. Bogunović, "Joint retinal layer and fluid segmentation in OCT scans of eyes with severe macular edema using unsupervised representation and auto-context," *Biomed. Opt. Exp.*, vol. 8, no. 3, pp. 1874–1888, 2017.

[8] B. Liefers *et al.*, "Automatic detection of the foveal center in optical coherence tomography," *Biomed. Opt. Exp.*, vol. 8, no. 11, pp. 5160–5178, 2017.

[9] B. Liefers, F. G. Venhuizen, T. Theelen, C. Hoyng, B. van Ginneken, and C. I. Sánchez, "Fovea detection in optical coherence tomography using convolutional neural networks," *Proc. SPIE*, vol. 10133, 2017, Art. no. 1013302.

[10] M. Li, Y. Wang, Z. Ji, W. Fan, S. Yuan, and Q. Chen, "Fast and robust fovea detection framework for OCT images based on foveal avascular zone segmentation," *OSA Continuum*, vol. 3, no. 3, pp. 528–541, 2020.

[11] S. Vujosevic *et al.*, "Diabetic macular edema with neuroretinal detachment: OCT and OCT-angiography biomarkers of treatment response to anti-VEGF and steroids," *Acta Diabetologica*, vol. 57, no. 3, pp. 287–296, 2020.

[12] C. Simader *et al.*, "Retinal thickness measurements with spectral domain optical coherence devices from different manufacturers in a reading center environment," *Invest. Ophthalmol. Vis. Sci.*, vol. 53, no. 14, pp. 4067–4067, 2012.

[13] D. Odell, A. M. Dubis, J. F. Lever, K. E. Stepien, and J. Carroll, "Assessing errors inherent in OCT-derived macular thickness maps," *J. Ophthalmol.*, vol. 2011, 2011, Art no. 692574.

[14] P. Seeböck *et al.*, "Unsupervised identification of disease marker candidates in retinal OCT imaging data," *IEEE Trans. Med. Imag.*, vol. 38, no. 4, pp. 1037–1047, Apr. 2019.

[15] J. Fujimoto and E. Swanson, "The development, commercialization, and impact of optical coherence tomography," *Invest. Ophthalmol. Vis. Sci.*, vol. 57, no. 9, pp. OCT1–OCT13, 2016.

[16] E. Swanson and D. Huang, "Ophthalmic OCT reaches $1 billion per year," *Retinal Physician*, vol. 8, no. 4, pp. 58–59, 2011.

[17] U. Schmidt-Erfurth, A. Sadeghipour, B. S. Gerendas, S. M. Waldstein, and H. Bogunović, "Artificial intelligence in retina," *Prog. Retinal Eye Res.*, vol. 67, pp. 1–29, 2018.

[18] E. Garcia-Melendo and E. J. Delp, "The use of image processing techniques for the analysis of echocardiographic images," School Elect. Eng., Purdue Univ., West Lafayette, Ind., Tech. Rep. TR-EE 88-29, Jul. 1988.

[19] I. Wächter *et al.*, "Patient specific models for planning and guidance of minimally invasive aortic valve implantation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, Berlin, Heidelberg: Springer, 2010, pp. 526–533.

[20] D. Zhang, J. Wang, J. H. Noble, and B. M. Dawant, "HeadLocNet: Deep convolutional neural networks for accurate classification and multi-landmark localization of head CTs," *Med. Image Anal.*, vol. 61, 2020, Art. no. 101659.

[21] J. M. Wolterink, R. W. van Hamersvelt, M. A. Viergever, T. Leiner, and I. Išgum, "Coronary artery centerline extraction in cardiac CT angiography using a CNN-based orientation classifier," *Med. Image Anal.*, vol. 51, pp. 46–60, 2019.

[22] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Integrating spatial configuration into heatmap regression based CNNs for landmark localization," *Med. Image Anal.*, vol. 54, pp. 207–219, 2019.

[23] M. I. Meyer, A. Galdran, A. M. Mendonça, and A. Campilho, "A pixel-wise distance regression approach for joint retinal optical disc and fovea detection," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, Cham, Switzerland: Springer, 2018, pp. 39–47.

[24] A. Alansary *et al.*, "Evaluating reinforcement learning agents for anatomical landmark detection," *Med. Image Anal.*, vol. 53, pp. 156–164, 2019.

[25] W. A. Al and I. D. Yun, "Partial policy-based reinforcement learning for anatomical landmark localization in 3D medical images," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 1245–1255, Apr. 2020.

[26] J. Zhang, M. Liu, and D. Shen, "Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4753–4764, Oct. 2017.

[27] D. Štern, T. Ebner, and M. Urschler, "From local to global random regression forests: Exploring anatomical landmark localization," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, Cham, Switzerland: Springer, 2016, pp. 221–229.

[28] C. Lindner and T. F. Cootes, "Fully automatic cephalometric evaluation using random forest regression-voting," in *Proc. IEEE Int. Symp. Biomed. Imag.*, Citeseer, 2015, pp. 1–8.

[29] O. Oktay *et al.*, "Stratified decision forests for accurate anatomical landmark localization in cardiac images," *IEEE Trans. Med. Imag.*, vol. 36, no. 1, pp. 332–342, Jan. 2017.

[30] R. I. Ionasec *et al.*, "Dynamic model-driven quantitative and visual evaluation of the aortic valve from 4D CT," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, Berlin, Heidelberg: Springer, 2008, pp. 686–694.

[31] B. Ibragimov, B. Likar, F. Pernus, and T. Vrtovec, "Computerized cephalometry by game theory with shape-and appearance-based landmark refinement," in *Proc. Int. Symp. Biomed. Imag.*, 2015, pp. 1–8.

[32] M. A. Dabbah *et al.*, "Detection and location of 127 anatomical landmarks in diverse CT datasets," *Proc. SPIE*, vol. 9034, 2014, Art. no. 903415.

[33] D. Mahapatra, "Landmark detection in cardiac MRI using learned local image statistics," in *Proc. Int. Workshop Stat. Atlases Comput. Models Heart*, Berlin, Heidelberg: Springer, 2012, pp. 115–124.

[34] M. Urschler, T. Ebner, and D. Štern, "Integrating geometric configuration and appearance information into a unified framework for anatomical landmark localization," *Med. Image Anal.*, vol. 43, pp. 23–36, 2018.

[35] D. Han, Y. Gao, G. Wu, P.-T. Yap, and D. Shen, "Robust anatomical landmark detection for MR brain image registration," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, Cham, Switzerland: Springer, 2014, pp. 186–193.

[36] W. A. Al, H. Y. Jung, I. D. Yun, Y. Jang, H.-B. Park, and H.-J. Chang, "Automatic aortic valve landmark localization in coronary CT angiography using colonial walk," *PLoS One*, vol. 13, no. 7, 2018, Art. no. e0200317.

[37] Y. Gao and D. Shen, "Collaborative regression-based anatomical landmark detection," *Phys. Med. Biol.*, vol. 60, no. 24, 2015, Art. no. 9377.

[38] H. Lee, M. Park, and J. Kim, "Cephalometric landmark detection in dental x-ray images using convolutional neural networks," *Proc. SPIE*, vol. 10134, 2017, Art. no. 101341W.

[39] M. Bhende, S. Shetty, M. K. Parthasarathy, and S. Ramya, "Optical coherence tomography: A guide to interpretation of common macular diseases," *Indian J. Ophthalmol.*, vol. 66, no. 1, pp. 20–35, 2018.

[40] N. Cuenca, I. Ortuño-Lizarán, and I. Pinilla, "Cellular characterization of OCT and outer retinal bands using specific immunohistochemistry markers and clinical implications," *Ophthalmology*, vol. 125, no. 3, pp. 407–422, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0161642017316743

[41] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, Cham, Switzerland: Springer, 2015, pp. 234–241.

[42] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.* 2015.

[43] A. L. Maas *et al.*, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int. Conf. Mach. Learn.*, vol. 30, no. 1. Citeseer, 2013, pp. 1–6.

[44] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," in *Proc. Int. Conf. Mach. Learn. Deep Learn.*, 2015, pp. 1–5.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Rrepresentations*, 2015.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[47] A. Paszke *et al.*, "PYTORCH: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach *et al.*, Eds., New York, NY, USA: Curran Associates, 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[48] H. Kolb, R. F. Nelson, P. K. Ahnelt, I. Ortuño-Lizarán, and N. Cuenca, "The architecture of the human fovea," *Webvision: The Organization of the Retina and Visual System*, Moran Eye Center, 2020. [Online]. Available: https://webvision.med.utah.edu/book/part-ii-anatomy-and-physiology-of-the-retina/the-architecture-of-the-human-fovea/

[49] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.

[50] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[51] A. Q. O'Neil *et al.*, "Attaining human-level performance with atlas location autocontext for anatomical landmark detection in 3D CT data," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 470–484.