npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02310-w

Enhancing ocular sign detection: AI-based strategic segmentation for improved accuracy and privacy protection

Chaoyu Lei[1,2,8], Chen Zhao[1,8], Jiayu Chen[3,8], Xuran Duan[1], Chudi Xu[3], Chee Chew Yip[4], Sunisa Sintuwong[5], Jianbin Ding[6], P. S. Pandiyan[4], Sunsern Wattanaphanich[5], Yujie Ren[1], Siqi Luo[7], Xuefei Song[1,2], Hong He[3] ✉, Xiaohong Liu[7] ✉ & Huifang Zhou[1,2] ✉

Accurate detection of ocular signs is essential for early diagnosis of eye diseases, but current AI approaches using facial or external ocular images include non-essential information, compromising performance and patient privacy. We conducted a multinational retrospective study of 2360 eyes from 1180 half-face images of thyroid eye disease patients across five racial groups from five hospitals in three countries. We developed a Dense Squeeze-and-Excitation Network (DSE-Net) to segment eyelid, conjunctiva, lacrimal caruncle, and eyeball, minimizing exposure and enhancing privacy. DSE-Net achieved Dice coefficient of 84.7%, 84.8%, 92.7%, and 95.1%, outperforming seven segmentation models. We then built SegmenView, employing LeNet, AlexNet, ResNet50, and VGGNet16 to detect eyelid edema, conjunctival erythema, caruncle or plica edema, and exophthalmos. SegmenView achieved internal Area Under the Curve (AUCs) of 71.09%, 80.81%, 90.07%, and 82.86%; external AUCs ranging 55.58%–84.29% across two test datasets, outperforming half-face and periocular models. We also compared SegmenView with four privacy-preserving methods, showing its superior ability to balance privacy protection with diagnostic accuracy. Additionally, visualizations based on Gradient-weighted Class Activation Mapping (Grad-CAM) further enhanced the model's interpretability. Our approach demonstrates high accuracy, generalizability, and potential for lightweight, privacy-preserving ocular sign detection.

Ophthalmic diseases constitute a significant healthcare burden globally, as approximately 0.5% of the world's population is blind, and over 200 million people experience moderate to severe visual impairment[1]. More than one billion cases of vision loss are potentially reversible with early intervention, highlighting the need for timely early diagnosis[2]. Many ophthalmic diseases—including orbital, eyelid, and ocular surface disorders—manifest early through visible changes in the patient's outer appearance, particularly affecting four regions: the eyelid, conjunctiva, lacrimal caruncle, and eyeball[3–5]. Therefore, accurate detection of signs in these four ocular regions is essential for early intervention. However, detecting these signs is challenging due to their subtle nature and reliance on subjective clinical judgment[6].

Artificial intelligence (AI) offers a promising solution by enhancing diagnostic consistency and reducing reliance on physician experience, with applications in disease diagnosis and sign detection across various medical fields such as ophthalmology[7–9] and neurology[10–12]. However, in ocular sign detection[13–15], existing AI-based approaches often rely on facial or periocular images, which can pose potential privacy risks[16,17]. Regulations like the European Union's General Data Protection Regulation (GDPR)[18] emphasize the importance of protecting facial image data, highlighting these concerns. Detailed segmentation of lesion areas can mitigate this issue by

[1]State Key Laboratory of Eye Health, Department of Ophthalmology, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China. [2]Hainan International Medical Center, Shanghai Jiao Tong University School of Medicine, Hainan, China. [3]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China. [4]Department of Ophthalmology & Visual Sciences, Khoo Teck Puat Hospital, Singapore, Singapore. [5]Department of Ophthalmology, Mettapracharak (Wat Rai Khing) Hospital, Nakhon Pathom, Thailand. [6]Department of Ophthalmology, National University Hospital, Singapore, Singapore. [7]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. [8]These authors contributed equally: Chaoyu Lei, Chen Zhao, Jiayu Chen. ✉e-mail: hehong@usst.edu.cn; xiaohongliu@sjtu.edu.cn; fangzzfang@sjtu.edu.cn

removing identifiable facial features, thereby enhancing privacy protection. This provides a novel approach to ocular sign detection. Therefore, this study aims to explore the impact of precise segmentation of external ocular regions on sign recognition.

This study focuses on thyroid eye disease (TED), a sight-threatening autoimmune condition affecting 25–50% of Graves' disease patients, characterized by early signs such as eyelid edema, conjunctival erythema, caruncle or plica edema, and exophthalmos[19,20]. Due to its diverse clinical features, TED serves as an ideal model for lesion segmentation and sign detection research. To address the limitations of existing methods, we developed SegmenView, powered by our proposed Dense Squeeze-and-Excitation Network (DSE-Net) segmentation method. This approach isolates only the four key ocular regions —eyelid, conjunctiva, lacrimal caruncle, and eyeball, to detect their corresponding clinical signs: eyelid edema, conjunctival erythema, caruncle or plica edema, and exophthalmos. During the development of SegmenView, we used four Convolutional Neural Network (CNN) architectures (LeNet, AlexNet, ResNet50, and VGGNet16) to achieve the best performance of ocular sign detection. Additionally, to ensure robustness, the model's reliability and generalizability were rigorously validated using patient data from five different racial groups, collected across five tertiary institutions in three countries. SegmenView was also compared with four privacy-preserving methods. Visualizations based on Gradient-weighted Class Activation Mapping (Grad-CAM) were used to enhance the model's interpretability.

Results

Baseline data

The study flow chart is shown in Fig. 1. A total of 1180 patient images, representing 2360 eyes, were collected across five hospitals. Based on the facial assessment, the Training dataset's incidence rates for eyelid edema, conjunctival erythema, caruncle or plica edema, and exophthalmos were 32.61%, 38.49%, 38.49%, and 51.28% respectively. For the Internal test dataset, these rates were 38.00%, 30.00%, 32.50%, and 39.50% respectively. In the External test dataset 1, these rates were 54.02%, 59.77%, 42.53%, and 65.52% respectively. In the External test dataset 2, these rates were 46.85%, 61.26%, 44.14%, and 70.27% respectively. The average ages for the Training, Internal test, External test 1, and External test 2 datasets were 49.20 ± 13.64, 46.87 ± 12.96, 45.07 ± 13.10, and 52.23 ± 11.83 years, respectively. Male patients comprised 39.9% of the Training dataset and 34.5% of the Internal test dataset. A comprehensive summary of baseline characteristics can be found in Table 1.

Segmentation performance

The DSE-Net achieved Dice Coefficient (Dice) scores of 84.7%, 84.8%, 92.7%, and 95.1% for the eyelid, conjunctiva, lacrimal caruncle, and eyeball, respectively. The model achieved Accuracy levels of 96.8%, 99.3%, 99.8%, and 99.5% across the four regions. Precision scores were 89.4%, 93.6%, 92.1%, and 94.5%, with Recall rates of 82.4%, 78.0%, 93.6%, and 96.1%. The Intersection over Union (IoU) values were 74.7%, 74.0%, 86.5%, and 90.9% across the respective regions (Table 2). These results represented the highest or near-highest performance compared to all evaluated segmentation models (Table 2).

In addition to these performance metrics, a comparison of the number of trainable parameters was conducted across several segmentation models. The results were displayed in Supplementary Fig. 1. DSE-Net had the lowest number of parameters at 11.02 M, compared to U-Net (31.04 M), SE-Unet (23.59 M), SegNet (29.44 M), DeepLab v3 (39.63 M), DeepLab v3+ (39.73 M), TransUNet (67.08 M), and TransFuse (26.27 M).

The results of the ablation experiment with three configurations: the SE module, DSE-1, and DSE-2, are shown in Supplementary Table 1. The addition of the DSE components progressively improved the performance, with DSE-2 achieving the best results: IoU 81.7%, Dice 89.5%, Precision 92.7%, Recall 87.4%, and Accuracy 98.9%.

Detection results for SegmenView

For the detection of eyelid edema, conjunctival erythema, caruncle or plica edema, and exophthalmos, the highest Area Under the Curve (AUC) in the Internal dataset was 0.7109 [0.6555–0.7663], 0.8081 [0.7529–0.8632], 0.9007 [0.8725–0.9289], and 0.8286 [0.7905–0.8667], respectively. In External test dataset 1, the AUCs reached 0.6076 [0.5226–0.6927], 0.8182 [0.7561–0.8803], 0.8429 [0.7881–0.8977], and 0.6564 [0.5719–0.7409] for the same signs. In External test dataset 2, the AUCs were 0.6557 [0.5846–0.7268], 0.6812 [0.6168–0.7456], 0.7021 [0.6313–0.7730], and 0.5558 [0.4738–0.6377], respectively (Fig. 2) (Table 3).

Detection results for the periocular model

In the Internal test dataset, the periocular model achieved AUCs of 0.6855 [0.6284–0.7426], 0.7943 [0.7521–0.8365], 0.8541 [0.8172–0.8910], and 0.7603 [0.7127–0.8079] for detecting eyelid edema, conjunctival erythema, caruncle or plica edema, and exophthalmos, respectively. In External test dataset 1, the AUCs for them were 0.4743 [0.3889–0.5597], 0.6309 [0.5497–0.7121], 0.7129 [0.6400–0.7859], and 0.6494 [0.5646–0.7342]. For External test dataset 2, the model's AUC values were 0.6215 [0.5488–0.6941], 0.6302 [0.5575–0.7029], 0.5578 [0.4829–0.6328], and 0.5370 [0.4593–0.6147] (Fig. 2) (Table 3).

Detection results for the half-face model

In the Internal test dataset, the AUCs of the half-face model for recognizing eyelid edema, conjunctival erythema, caruncle or plica edema, and exophthalmos were 0.5440 [0.4627–0.6253], 0.5148 [0.4273–0.6023], 0.5502 [0.4631–0.6373], and 0.4528 [0.3727–0.5329], respectively. For External test dataset 1, the model achieved AUCs of 0.4016 [0.2785–0.5247], 0.5104 [0.3847–0.6362], 0.2854 [0.1736–0.3972], and 0.4023 [0.2839–0.5208]. In the External test dataset 2, the AUCs for them were 0.3879 [0.2819–0.4938], 0.5219 [0.4183–0.6255], 0.3825 [0.2768–0.4881], and 0.4705 [0.3503–0.5906] (Fig. 2) (Table 3).

Model performance comparison

To investigate the impact of different segmentation levels on ocular sign detection, we compared the optimal performance of three models in detecting four ocular signs. The results show that SegmenView consistently achieved higher AUC values than half-face and periocular models across all signs and in all three test datasets. For example, in recognizing caruncle or plica edema, SegmenView outperformed the others with AUCs of 0.9007 [0.8725–0.9289] vs. 0.8541 [0.8172–0.8910] vs. 0.5502 [0.4631–0.6373] in the Internal test dataset, 0.8429 [0.7881–0.8977] vs. 0.7129 [0.6400–0.7859] vs. 0.2854 [0.1736–0.3972] in External test dataset 1, and 0.7021 [0.6313–0.7730] vs. 0.5578 [0.482–0.6328] vs. 0.3825[0.2768-0.4881] in External test dataset 2 (Table 3). Additionally, when considering other evaluation metrics, the area under the radar plot for the SegmenView was consistently larger than those of the other two models across all conditions (Fig. 3).

Interpretability analysis with Grad-CAM

As shown in Fig. 4, the heatmaps revealed that the model primarily attended to clinically meaningful regions: eyelid margins for eyelid edema, hyperemic conjunctival zones for conjunctival erythema, inflamed lacrimal caruncle tissue for caruncle or plica edema, and nasal and temporal aspects of the eyeball for exophthalmos. These consistent activation patterns across representative samples indicated that SegmenView focused on disease-relevant features rather than irrelevant background information. For half-face and periocular images, the heatmaps showed broader activations that included the periorbital skin. However, these activations also involved irrelevant areas such as the eyebrow, nasal bridge, and cheek (Supplementary Fig. 2).

Detection results for the comparative privacy-preserving methods

The performance of the privacy-preserving methods was generally lower than SegmenView, with AUCs not exceeding 0.65. For pixel scrambling encryption, the AUCs for recognizing eyelid edema, conjunctival erythema, lacrimal caruncle edema, and exophthalmos in the Internal test dataset were

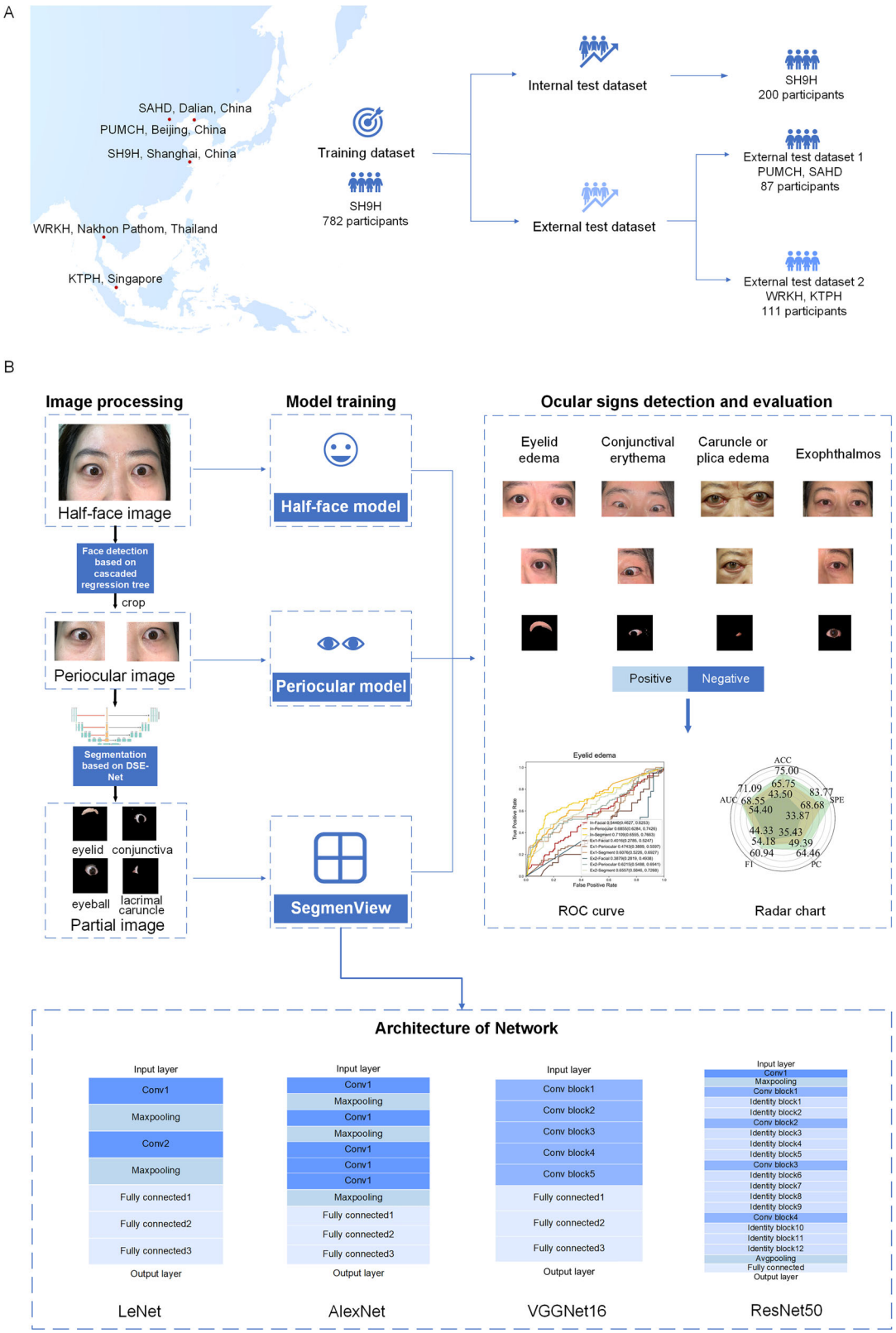

Fig. 1 | Workflow diagram. A Gathering facial images from a total of five hospitals and organizing them into different datasets: a training dataset, an internal test dataset, and two separate external test datasets. **B** Employing three categories of images, each featuring a different level of segmentation, to construct three distinct detection models.

42.78, 43.81, 46.89, and 62.89, respectively. For encryption based on chaotic logistic maps, the AUCs in the Internal test dataset were 60.81, 51.88, 57.91, and 52.73. For XOR encryption, the AUCs were 51.48, 50.47, 46.59, and 52.58. Finally, for homomorphic encryption, the AUCs in the Internal test dataset were 50.67, 54.81, 49.27, and 53.24. The results for the two external test datasets are shown in Supplementary Table 2.

Table 1 | Baseline table for patients in the training dataset and test dataset

Variable	Training dataset (SH9H, China)	Internal test dataset (SH9H, China)	External test dataset 1 (PUMCH, China; SAHD, China)	External test dataset 2 (KTPH, Singapore; WRKH, Thailand)
Number of participants	782	200	87	111
Number of eyes	1564	400	174	222
Age, mean (SD), y	49.20 ± 13.64	46.87 ± 12.96	45.07 ± 13.10	52.23 ± 11.83
Sex				
Male	312(39.9)	69(34.5)	30(34.48)	54(48.65)
Female	470(60.1)	131(65.5)	57(65.52)	57(51.35)
Race				
Chinese	782(100)	200(100)	87 (100)	33(29.73)
Thai	0	0	0	68(61.26)
Malay	0	0	0	7(6.31)
Indian	0	0	0	2(1.8)
Burmese	0	0	0	1(0.9)
TED activity				
Active	343(43.86)	60(30)	56(64.37)	39(35.14)
Inactive	439(56.14)	140(70)	31(35.63)	72(64.86)
TED severity				
Mild	197(25.19)	57(28.5)	20(22.99)	21(18.92)
Moderate-to-severe	498(63.68)	125(62.5)	50(57.47)	77(69.37)
Sight-threatening	87(11.13)	18(9)	17(19.54)	13(11.71)
Ocular signs (Based on face)				
Eyelid edema	255(32.61)	76(38)	47(54.02)	52(46.85)
Conjunctival erythema	301(38.49)	60(30)	52(59.77)	68(61.26)
Caruncle or plica edema	301(38.49)	65(32.5)	37(42.53)	49(44.14)
Exophthalmos	401(51.28)	79(39.5)	57(65.52)	78(70.27)
Ocular signs (Based on eyes)				
Eyelid edema	475(30.37)	135(33.75)	99(56.90)	96(43.24)
Conjunctival erythema	481(31.14)	99(24.75)	90(51.72)	117(52.70)
Caruncle or plica edema	515(32.93)	109(27.25)	67(38.51)	85(38.29)
Exophthalmos	627(40.09)	125(31.25)	105(60.34)	129(58.11)

Discussion

In this study, we have developed and validated SegmenView for detecting ocular signs of TED. Among eight segmentation networks, DSE-Net has the lowest parameter count but the best performance. Utilizing lesion areas segmented by DSE-Net, this model further employed four CNN frameworks for detecting ocular signs. Notably, the model exhibited robust generalizability across five different races and five imaging devices. Comparing the performance of this model with traditional models that used half-face photos and periocular images, SegmenView consistently outperformed these approaches. This highlights the benefits of precise lesion segmentation, which not only enhances detection accuracy but also reduces the ocular area analyzed, thereby offering better privacy protection for patients.

In recent years, image segmentation techniques have seen increasingly broad applications in medical imaging. These methods efficiently and precisely delineate pathological areas in images captured[21,22]. However, there has been little research focused on the precise segmentation of lesion areas for ophthalmic diseases. This may inadvertently include excessive irrelevant regions, limiting model performance and introducing potential privacy concerns. To address these challenges, we employed the DSE-Net segmentation technique, which showed exceptional performance in precisely segmenting ophthalmic lesion areas and capturing subtle features. We also compared DSE-Net with seven other segmentation models, proving its superiority. For instance, in detecting the lacrimal caruncle, the algorithm achieved a Dice score of 92.7%, surpassing all other models. This is attributed to the DSE module, which combines multiple convolutional blocks and Squeeze-and-Excitation modules through dense connections. The results of the ablation study also show that adding the DSE components progressively improves performance across all metrics, including IoU, Dice, Precision, Recall, and Accuracy. This approach enhances the effective feature information while preventing input information loss, thereby overcoming the challenges posed by the complex and varied morphology of patient facial images on the segmentation module. Furthermore, compared to other segmentation models, this model has the smallest number of parameters (11.02 M), providing a foundation for embedding it into lightweight devices.

Based on segmented lesion regions, we selected four CNN frameworks to build SegmenView for their effectiveness in extracting local features and their well-established reliability in medical imaging tasks. This improvement is due to precise segmentation, which isolates small lesion areas that are often overlooked in full-face or periocular images. For example, the lacrimal caruncle, a small anatomical structure, can easily be obscured by surrounding tissues in broader images. By focusing on these specific regions, SegmenView enhances the model's ability to detect subtle changes that might otherwise be missed. Additionally, different CNN architectures perform better for different classification tasks, and combining their

Table 2 | Segmentation results of each model (%)

Part	Method	IoU	Dice	Precision	Recall	Accuracy
Eyelid	U-Net	74.6	**84.7**	**92.2**	80.1	**96.9**
	SE-UNet	73.7	83.9	90.8	80.2	96.7
	SegNet	73.7	84.0	90.0	80.8	96.7
	DeepLab v3	74.0	84.5	90.4	81.2	96.8
	DeepLab v3+	73.3	83.7	88.7	81.1	96.6
	TransUNet	73.5	84.1	88.2	82.3	96.7
	TransFuse	74.5	84.5	89.8	81.5	96.7
	DSE-Net	**74.7**	**84.7**	89.4	**82.4**	96.8
Conjunctiva	U-Net	71.7	83.0	**95.0**	74.5	99.2
	SE-UNet	73.4	84.3	93.9	77.2	**99.3**
	SegNet	71.8	83.1	92.7	76.4	99.2
	DeepLab v3	71.3	82.9	91.1	76.8	99.2
	DeepLab v3+	73.7	84.5	93.1	77.9	**99.3**
	TransUNet	72.7	83.9	92.4	77.4	**99.3**
	TransFuse	61.3	74.4	94.8	63.5	99.0
	DSE-Net	**74.0**	**84.8**	93.6	**78.0**	**99.3**
Lacrimal Caruncle	U-Net	86.3	92.5	**92.8**	92.7	**99.8**
	SE-UNet	85.0	91.7	91.2	92.8	**99.8**
	SegNet	83.6	90.8	90.7	91.7	**99.8**
	DeepLab v3	83.7	91.0	90.0	92.5	**99.8**
	DeepLab v3+	83.8	90.9	91.7	91.2	**99.8**
	TransUNet	84.0	91.1	89.6	93.3	**99.8**
	TransFuse	80.0	87.1	90.6	86.6	**99.8**
	DSE-Net	**86.5**	**92.7**	92.1	**93.6**	**99.8**
Eyeball	U-Net	**90.9**	**95.1**	93.2	**97.5**	**99.5**
	SE-UNet	90.0	94.6	92.9	96.9	**99.5**
	SegNet	88.8	93.4	92.0	96.3	99.4
	DeepLab v3	89.2	94.2	91.9	97.0	**99.5**
	DeepLab v3+	89.1	94.2	91.9	97.0	**99.5**
	TransUNet	89.5	94.4	92.1	97.1	**99.5**
	TransFuse	88.4	93.4	92.0	96.2	99.4
	DSE-Net	**90.9**	**95.1**	**94.5**	96.1	**99.5**

The bold values represent the best performance for each segmentation metric.

complementary strengths contributes to the overall improvement in model performance. Notably, to ensure a fair comparison, all baselines and SegmenView were trained under an identical protocol and evaluated at the best validation checkpoint. Accordingly, the observed gains reflect the effect of input scope and architecture rather than differences in training quality or tuning. Thus, compared to traditional models, SegmenView enables accurate detection of ocular signs and may be used for allergic conjunctivitis[23], and orbital tumors[24], which have similar clinical signs. Furthermore, the lightweight CNN architectures increase the potential for the model to be easily deployed on mobile devices, making it more suitable for smartphone-based applications in the future.

We further validated model interpretability using Grad-CAM visualizations. For the segmented lesion areas, the heatmaps concentrated on clinically relevant structures such as eyelid margins and hyperemic conjunctival zones. This supports that SegmenView uses anatomically meaningful structures for detection. By contrast, the periocular and half-face heatmaps showed broader activations that included periorbital skin but also many unrelated regions such as the eyebrow, nasal bridge, cheek, and hairline. These extra activations dilute feature focus and can reduce sensitivity, particularly for small localized signs such as lacrimal caruncle edema. Thus, while unsegmented blind zones may contain clinically relevant structures, the accompanying unrelated regions likely introduce noise and reduce detection accuracy, which helps explain the superior performance of SegmenView compared with periocular and half-face models.

However, it is worth noting that we did not include signs such as eyelid retraction and eyelid erythema in our model, as these signs require comparison with surrounding structures for accurate identification. The precise segmentation approach removes the surrounding anatomical structures, which may impact the ability to distinguish these signs effectively. The signs selected, including eyelid edema, conjunctival erythema, and caruncle or plica edema, are relatively independent of surrounding structures for accurate identification. Although exophthalmos requires comparison with surrounding structures for differentiation, it causes relative retraction of the upper and lower eyelids, which in turn enlarges the segmented eyeball area, aiding in classification. Importantly, these signs often emerge in the early stages of TED, when the disease is still in its subclinical or mildly active form and therefore more difficult to detect using conventional approaches.

Table 3 | Performances in the convolutional neural networks across different data sources (%)

		Internal test dataset					External test dataset 1					External test dataset 2				
		Acc	Spe	P_c	F1	AUC	Acc	Spe	P_c	F1	AUC	Acc	Spe	P_c	F1	AUC
Half-face model	Eyelid edema	43.50	33.87	35.43	44.33	54.40	41.38	67.50	40.91	26.09	40.16	36.04	57.63	19.35	14.46	38.79
	Conjunctival erythema	61.50	77.14	31.91	28.04	51.48	48.28	62.86	60.61	47.06	51.04	47.75	76.74	66.67	40.82	52.19
	Caruncle or plica edema	44.00	40.00	29.57	37.78	55.02	32.18	36.00	23.81	25.32	28.54	42.34	24.19	40.51	50.00	38.25
	Exophthalmos	48.00	25.62	41.94	55.56	45.28	43.68	26.67	57.69	55.05	40.23	57.66	21.21	68.67	70.81	47.05
Periocular model	Eyelid edema	65.75	68.68	49.39	54.18	68.55	52.30	26.67	56.35	63.11	47.43	56.76	26.19	50.00	65.96	62.15
	Conjunctival erythema	77.25	74.09	52.44	65.40	79.43	58.05	52.38	58.76	60.96	63.09	53.15	60.95	56.84	50.94	63.02
	Caruncle or plica edema	75.00	68.38	52.33	66.89	85.41	63.79	42.06	51.56	67.69	71.29	52.25	47.45	41.46	49.04	55.78
	Exophthalmos	68.00	55.64	49.38	65.03	76.03	59.77	40.58	64.96	68.47	64.94	63.51	30.11	63.48	73.62	53.70
SegmenView	Eyelid edema	75.00	83.77	64.46	60.94	71.09	63.22	40.00	64.00	71.43	60.76	62.16	43.65	53.90	66.40	65.57
	Conjunctival erythema	84.25	92.36	71.95	65.19	80.81	72.99	73.81	74.71	73.45	81.82	61.26	86.67	76.27	51.14	68.12
	Caruncle or plica edema	82.25	75.60	60.56	75.43	90.07	74.71	59.81	60.55	75.00	84.29	63.96	81.02	54.39	43.66	70.21
	Exophthalmos	75.75	64.73	56.31	72.05	82.86	65.52	62.32	73.20	70.30	65.64	63.06	50.54	66.91	69.40	55.58

Acc Accuracy, *Spe* Specificity, *Pc* Precision, *F1* F1 score, *AUC* Area Under the Curve.

Notably, this study's dataset is comprised of images captured using two types of professional cameras and three types of smartphones, with significant differences in image quality. Despite these variations, SegmenView demonstrated robust generalizability across different devices. This highlights the model's applicability in various devices and underscores its potential to be integrated into different mobile devices, increasing its accessibility for remote or primary care facilities without advanced imaging equipment. However, embedding the model in mobile devices inevitably raises privacy concerns. The use of partial images in SegmenView allows the device to capture only the lesion areas, minimizing the risk of privacy breaches by avoiding the collection of irrelevant regions. Furthermore, our dataset includes individuals of various races, such as Chinese, Malay, Thai, Indian, and Burmese. There are significant structural differences in ocular features across these groups—for instance, Chinese individuals often exhibit more pronounced epicanthic folds[25], and different Asian populations may have significantly different orbital structures[26]. Skin color may also affect the detection of these ocular signs. To enhance data diversity, radiographic contrast adjustment techniques are applied. These augmentations simulate variations in image characteristics, helping to mitigate the impact of dataset distribution shifts. It turns out that despite these variations, SegmenView demonstrated strong robustness across these multiracial datasets, laying a foundation for future international promotions.

Our approach offers two key advantages for privacy protection. First, we minimize exposed facial information by applying segmentation, processing only the disease-related ocular regions and excluding all other facial areas, reducing the risk of identifiable data exposure. Second, the method could operate on edge devices, performing inference offline without uploading patient data to cloud servers, thereby preventing privacy breaches during data transmission or storage. With greater emphasis on the protection of facial image privacy globally, stringent laws to safeguard personal facial image privacy information[27]. For example, China's Personal Information Protection Law strictly regulates the use of facial images and other biometric information, mandating the prevention of information leaks during the handling and storage of such data[28]. Private entities are also required to use reasonable standards for storage and transmission to ensure the security of biometric information[29]. However, current research on ocular sign detection often directly uses half-face and periocular images without special privacy protection, increasing the risk of data exposure. As for the four privacy-preserving baselines, sign-detection performance was consistently lower than SegmenView. This indicates that while these techniques preserve privacy, they also degrade disease-relevant structure and make detection unreliable. Pixel scrambling disrupts spatial arrangement, which is essential for position-dependent signs such as exophthalmos. Logistic chaotic mapping introduces broad random perturbations that blur morphology. XOR encryption destroys pixel-value distributions and local neighborhood continuity, randomizing color gradients, edges, and texture. Homomorphic encryption typically requires dimensionality reduction and quantization that discards fine detail.

All in all, SegmenView can be integrated into clinical workflows as a powerful screening tool, enabling specialists to detect subtle ocular signs, such as conjunctival erythema and caruncle or plica edema, which are often missed in routine exams. SegmenView's lightweight, on-device design supports deployment in primary care, telemedicine, and community screening without the need for additional hardware or cloud data transfer. This reduces privacy risks and makes the method feasible in resource-limited settings, including mobile applications for real-time use. In clinical practice, for example, in TED patients, a practical end-to-end workflow could unfold as follows: at home, a patient takes a photo using a smartphone app that highlights potential ocular signs and encourages further medical evaluation. Many patients' first visit primary care or endocrinology, where subtle eye changes might be overlooked. By using the same screening tool in these settings, early signs can be recognized, the differential diagnosis narrowed, and timely referral to ophthalmology ensured. After the specialist's assessment, periodic photos can track the progression of signs, enabling timely therapeutic adjustments to prevent irreversible vision loss.

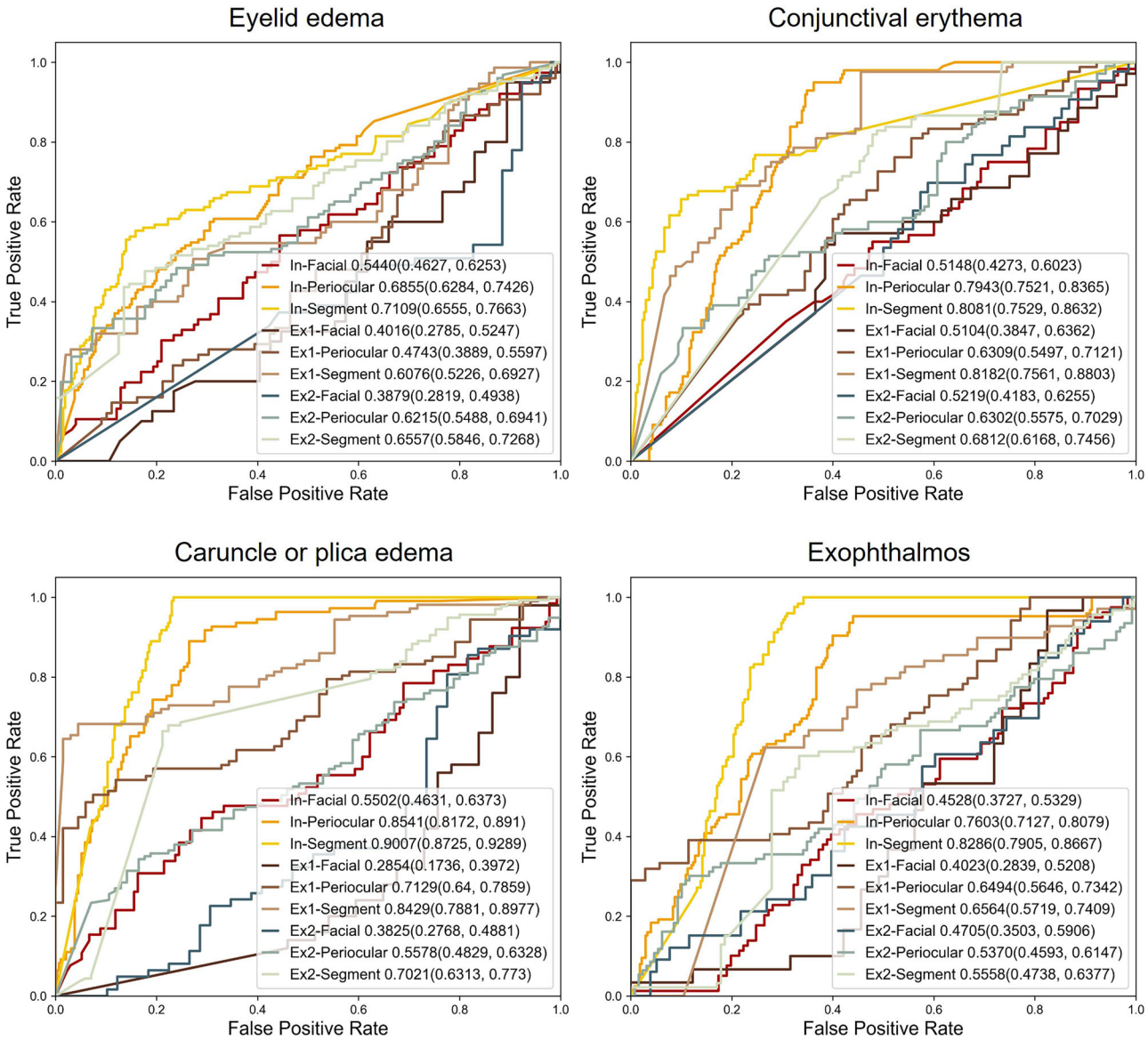

Fig. 2 | Receiver Operating Characteristic curves and Area Under the Curve values for four ocular signs. The results for eyelid edema, conjunctival erythema, caruncle or plica edema, and exophthalmos across internal and external test datasets based on different models (half-face model, periocular model, SegmenView) were displayed. The legends represent the following: In-Facial: Results for the Internal test dataset using the half-face model; In-Periocular: Results for the Internal test dataset using the periocular model; In-Segment: Results for the Internal test dataset using SegmenView; Ex1-Facial: Results for External test dataset 1 using the half-face model; Ex1-Periocular: Results for External test dataset 1 using the periocular model; Ex1-Segment: Results for External test dataset 1 using SegmenView; Ex2-Facial: Results for External test dataset 2 using the half-face model; Ex2-Periocular: Results for External test dataset 2 using the periocular model; Ex2-Segment: Results for External test dataset 2 using SegmenView.

This study has several limitations. First, the study population was predominantly Chinese, with limited representation from other racial groups. Future studies need to validate the model with data from multiple countries in real-world settings to further enhance the model's robustness and universality. Second, this study only focused on four ocular signs and features well-defined within segmented regions. Future studies should focus on developing advanced detection and segmentation strategies for diverse ocular signs to enhance model performance. Third, the images collected in this study were limited to frontal views. Future studies could incorporate 9 gaze photos, which are particularly valuable in neuro-ophthalmology. These images are essential for documenting and diagnosing neurological conditions affecting ocular motility, as well as strabismus. Besides, gaze photos could serve as a useful tool for the long-term follow-up of such disorders. Additionally, future research could explore self-supervised learning for tasks with limited labeled data, and multimodal large language models combining images and text to develop a lightweight multimodal screening tool, improving prediction accuracy in complex clinical scenarios[30].

Methods

Study design and dataset

This multinational, multicenter, retrospective study was led by Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine (SH9H, Shanghai, China). The four local and international collaborating institutions were Peking Union Medical College Hospital (PUMCH, Beijing, China), Second Affiliated Hospital of Dalian Medical University (SAHD, Liaoning, China), Khoo Teck Puat Hospital (KTPH, Singapore), and Mettapracharak (Wat Rai Khing) Hospital (WRKH, Nakhon Pathom, Thailand).

Ethical approval for the study was obtained from the Ethics Committee of Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School

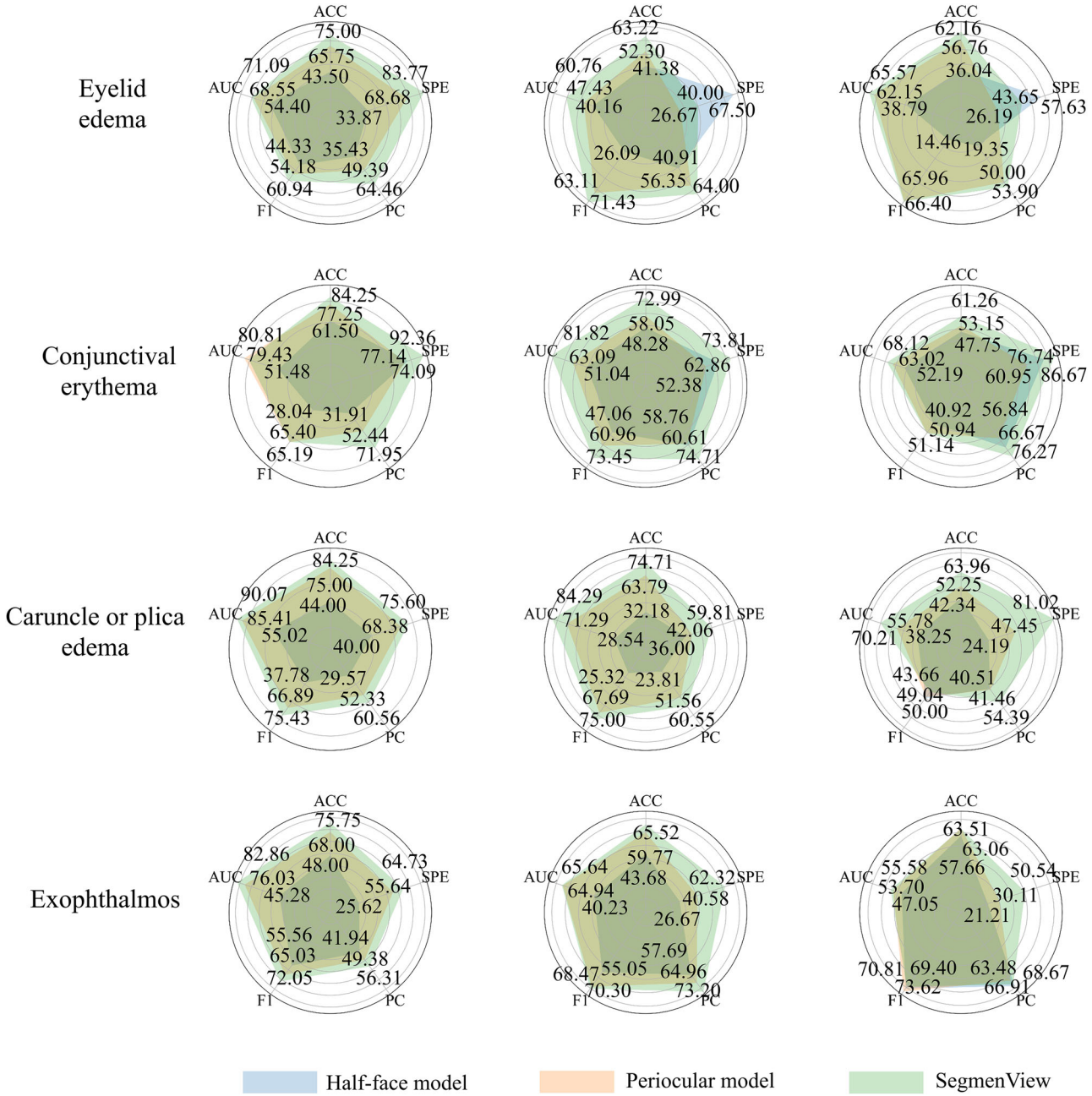

Fig. 3 | Radar chart of the performance of three detection models on different evaluation metrics. Each axis represents an evaluation metric, with the enclosed area indicating overall performance. A larger area signifies better performance. AUC Area Under the Curve, ACC accuracy, SPE specificity, PC precision, F1 F1-score.

of Medicine (Approval No. SH9H-2019-T8-1), the Ethics Committee of Peking Union Medical College Hospital, the Ethics Committee of Second Affiliated Hospital of Dalian Medical University, the Ethics Committee of Khoo Teck Puat Hospital, and the Ethics Committee of Mettapracharak (Wat Rai Khing) Hospital. This study was conducted in accordance with the principles of the Declaration of Helsinki. All participating individuals provided written informed consent.

Demographic and facial image data were collected from TED patients from January 2018 to December 2022. Inclusion criteria were: (1) diagnosis of TED according to the Bartley criteria[31]; (2) availability of complete clinical data; and (3) photographs without significant obstructions such as scars, birthmarks, tattoos, masks, excessive makeup, or hair.

Data from the Primary Center (SH9H) constituted the Training dataset with facial images from 782 TED patients (1564 eyes), and the Internal test dataset with data from 200 TED patients (400 eyes). The remaining four hospitals provided data for two external test datasets. External test dataset 1 included data from PUMCH and SAHD, with images from 87 TED patients (174 eyes). External test dataset 2 comprised data from KTPH and WRKH, with 111 TED patients (222 eyes) (Fig. 1A).

Data collection and annotation

Physicians captured clinical photographs of inpatient and outpatient evaluations. Patients were instructed to maintain primary gaze (looking straight ahead) with neutral facial expressions for image consistency. Multiple photographs per patient were taken under varying conditions, including differences in shooting distance, lighting, and background, to ensure a diverse dataset. A variety of imaging devices, including professional cameras (Canon, Sony) and smartphones (iPhone, Samsung, Huawei), were employed across different hospitals to achieve comprehensive image data collection.

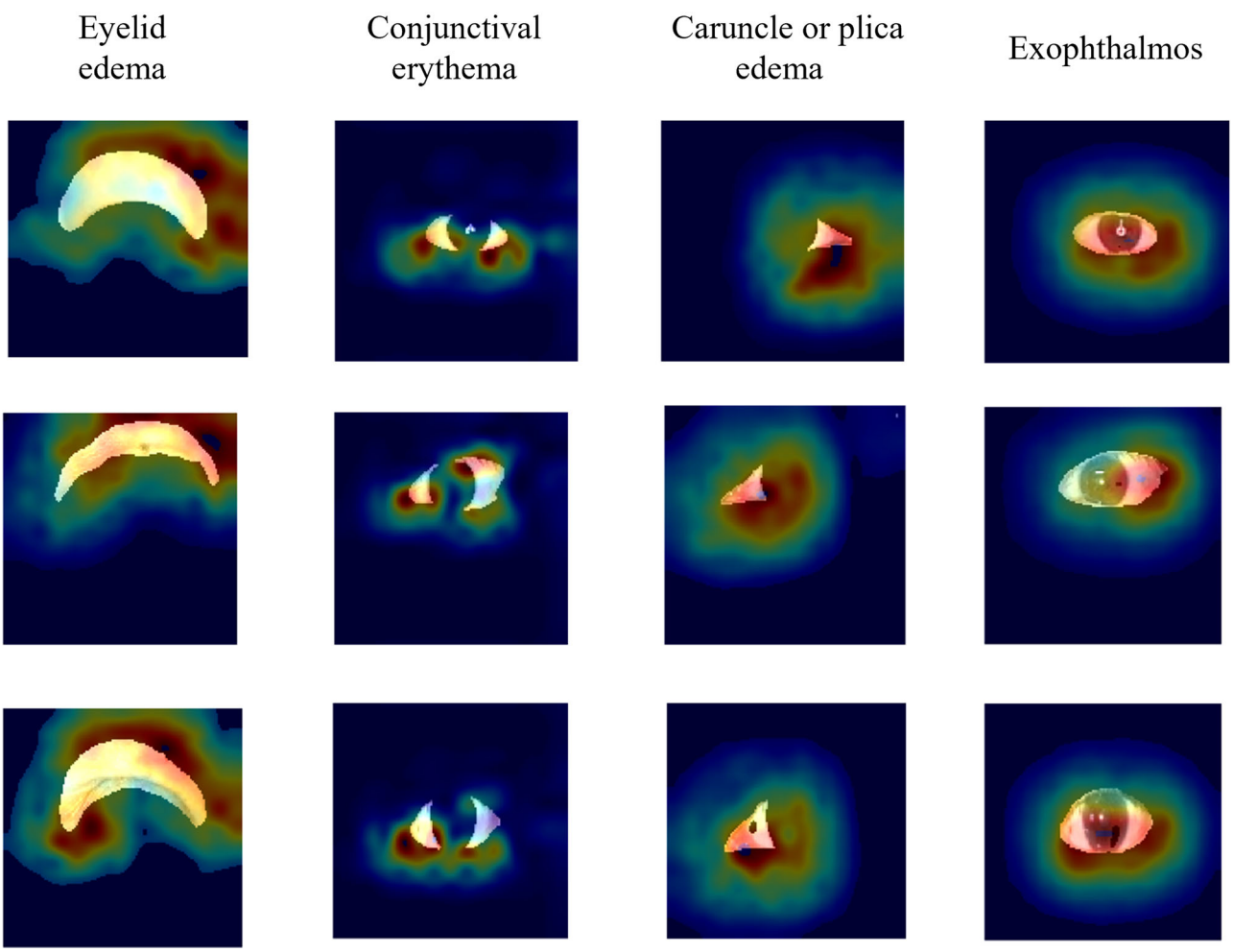

Fig. 4 | Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations of the contribution of segmented regions to classification. Attention concentrated on eyelid margins, hyperemic conjunctiva, inflamed caruncle, and the nasal and temporal aspects of the eyeball.

The annotation process focused on four key areas commonly affected by eye diseases: the eyelid, conjunctiva, lacrimal caruncle, and eyeball. Images were examined and annotated with signs of TED using the clinical activity score, including eyelid edema, conjunctival erythema, and caruncle or plica edema[32]. A measurement of 16 mm or greater on the Hertel exophthalmometer was considered significant for exophthalmos. The initial data annotation involved two ophthalmologists with at least five years of professional experience, who first assessed each image to determine the presence or absence of ocular signs in collected images. For images identified as positive, the lesion areas were meticulously outlined using Photoshop. A senior ophthalmologist with more than twenty years of experience reviewed and finalized these annotations.

Segmentation algorithm

The segmentation model (Fig. 5) employed a lightweight automatic segmentation algorithm for TED lesion areas based on DSE-Net[33]. DSE-Net used U-Net as its backbone model, progressively extracting low-level features from the encoding structure and integrating them with high-level features from the decoding structure, enhancing the model's feature extraction capability. A DSE channel attention module was designed to densely connect multiple feature maps, emphasizing valuable feature information while avoiding input information loss. This improved sensitivity in the feature channel dimension, thereby enhancing segmentation performance. For the rough segmentation of the left and right eye regions, we utilized a facial landmark detection technique based on regression trees, which divides the facial image into left and right eye regions. To enhance model performance, data augmentation methods including translation, horizontal flipping, vertical flipping, and random adjustment of brightness were employed. Besides, to address the issue of the limited pixel count of the lacrimal caruncle region, adjacent conjunctival regions were added during label generation to enhance feature learning and mitigate the impact of small regions. Before the classification task, post-processing algorithms of the segmentation model will be used to remove excessive sclera regions from the lacrimal caruncle area.

The segmentation model's training data were sourced from 369 TED patients at SH9H, and exclusively consisted of samples from TED patients. The data were preprocessed and cropped into 256×256 images, with each eye treated as an individual sample. The dataset was divided into training, validation, and test datasets in a ratio of 8:1:1.

To evaluate the segmentation results of DSE-Net across the four ocular regions, the Dice coefficient was adopted. Dice is a commonly used metric for evaluating the accuracy of image segmentation. It measures the overlap between the predicted segmentation and the ground truth, providing a value between 0 and 1, where higher values indicate better alignment.

In addition to the Dice, a comparison of DSE-Net with several widely used segmentation models, including U-Net, Squeeze-and-Excitation-Unet (SE-Unet), SegNet, DeeplabV3, DeeplabV3+, TransUNet, and TransFuse, was conducted. The comparison was based on performance metrics such as IoU, Dice, Precision, Recall, and Accuracy. The number of trainable parameters for each model was also compared to highlight the efficiency of DSE-Net.

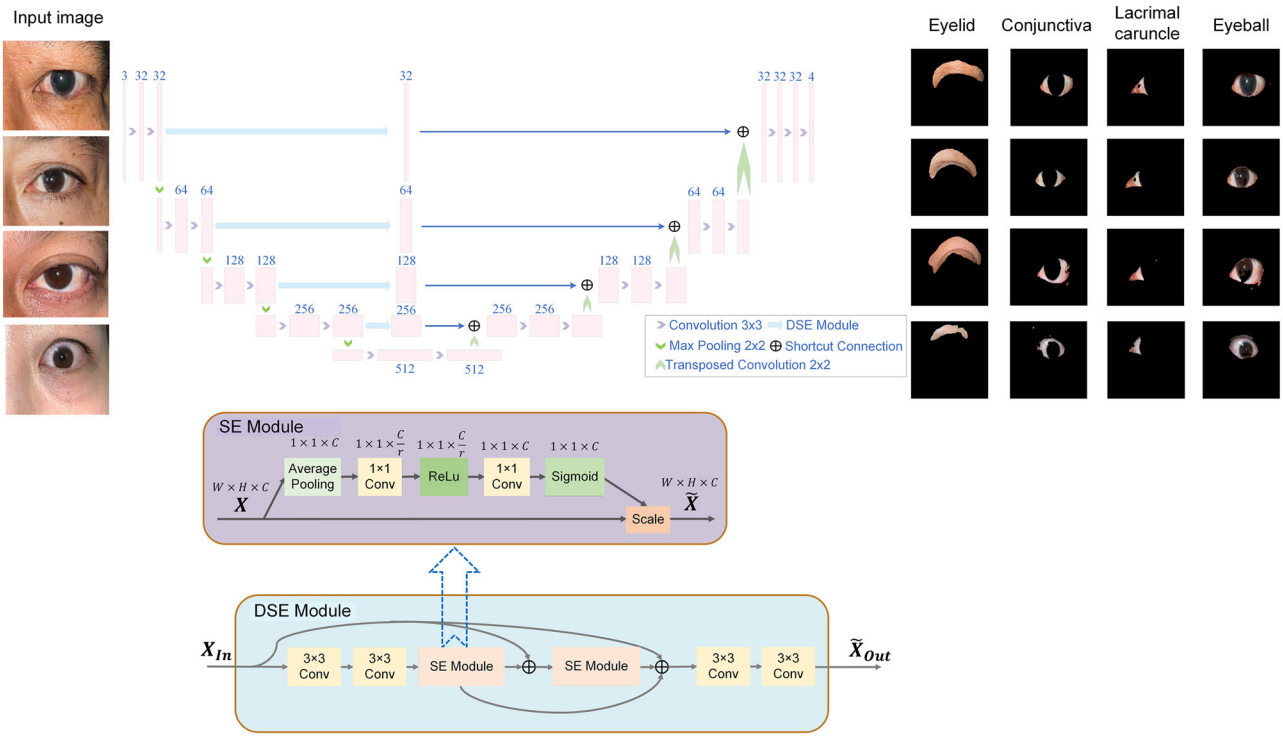

Fig. 5 | DSE-Net uses U-Net as its backbone model. DSE-Net incorporated a Dense Squeeze-and-Excitation (DSE) channel attention module that progressively extracted low-level features from the encoding structure and integrated them with high-level features from the decoding structure.

Additionally, to verify the effectiveness of the DSE module, we conducted an ablation study. The SE module, DSE with one SE module (DSE-1), and DSE with two SE modules (DSE-2) were applied to the encoder-decoder structure, and the segmentation performance across the four ocular regions was tested.

Detection algorithm

Based on the partial images, SegmenView was developed for sign detection. Additionally, a half-face model was built using half-face images, and a periocular model was constructed based on the periocular images. For each model, four frameworks: LeNet, AlexNet, ResNet50, and VGGNet16 were applied to perform binary classification on four ocular signs (eyelid edema, conjunctival erythema, caruncle or plica edema, and exophthalmos). For the half-face model, a positive result is recorded if either eye exhibits a positive manifestation. In contrast, the periocular model and SegmenView only focus on specific eyes.

The data were preprocessed and cropped into image samples of a specific size of 128 × 128, with each eye being treated as a separate sample. To ensure fairness and avoid undertraining or suboptimal tuning, we applied the same training methods and hyperparameters across all compared models, including Adam optimizer, learning rate of 0.001, cross-entropy loss function, batch size of 32, and 100 epochs. We also used early stopping if the validation performance showed no improvement for 20 consecutive epochs. The dataset was divided into training, validation, and test datasets in a ratio of 8:1:1. For final evaluation, we selected the version of each model that achieved the best validation performance before testing on the test set. Each experiment was set with the same hyperparameters for the binary classification task of disease signs. All experiments were conducted on an NVIDIA GeForce RTX 3090 graphics workstation with RAM of 24GB at the Medical Artificial Intelligence Laboratory, University of Shanghai for Science and Technology.

To evaluate the performance of the detection algorithms, the Receiver Operating Characteristic (ROC) curves were plotted, and the AUC was calculated. Additionally, the average values of Accuracy, Precision, Sensitivity, and F1-Score across the ten experiments on the test set were also calculated (Fig. 1B).

Grad-CAM visualization

To enhance the model's interpretability, we applied Grad-CAM to SegmenView, as well as to the periocular and half-face models. This analysis was performed to visualize and compare the differences in model attention across these three input types (segmented lesion areas, as well as the periocular and half-face images). By doing so, we were able to evaluate how each model focused on disease-relevant features and better understand the contributions of the segmented lesion areas, periocular, and half-face models to the detection process.

Privacy-preserving comparative methods

To compare SegmenView with traditional privacy-preserving methods in balancing privacy protection and diagnostic performance, we implemented four classical approaches: pixel scrambling encryption, encryption based on chaotic logistic maps, XOR encryption, and homomorphic encryption. We first expanded the dataset using data augmentation, including random rotations from −15° to 15°, horizontal flipping, and brightness adjustment. Then, we applied a wavelet transform to extract texture features and converted patient images into the HSV color space to better align with human visual perception. For each channel, we computed histograms with a preset number of bins (default: eight for H, S, and V) and performed L1 normalization to remove scale effects, creating a color feature vector. These wavelet and color features were flattened and fused into a joint feature vector, which was then encrypted using each of the four methods. Finally, a support vector machine with an RBF kernel was trained on the encrypted feature vectors to evaluate performance.

Data analysis

Descriptive statistics are used to describe the basic information of the collected data, such as age, gender, etc. The Mann–Whitney U-tests/t-tests were employed to evaluate the differences in the distributions of continuous variables, and the Pearson's Chi-squared test was applied to compare the

distributions of categorical variables. To compare the performance of different algorithms, we utilized DeLong's test to analyze the differences in AUC between various models. Statistical significance was determined by a p-value of less than 0.05. All statistical analyses and plotting were done in Python 3.8 (Python Software Foundation, Wilmington, Delaware, USA).

Data availability

The de-identified data supporting our findings are available from the corresponding author upon reasonable request.

Code availability

Code for data processing and model training is available from the corresponding author upon reasonable request.

Received: 6 December 2024; Accepted: 21 December 2025;
Published online: 06 January 2026

References

1. Burton, M. J. et al. The Lancet Global Health Commission on global eye health: vision beyond 2020. *Lancet Glob. Health* **9**, e489–e551 (2021).
2. Stern, J., Yasmin, S., Qureshi, M. B. & Bourne, R. 2030 in sight: the future of global eye health. *Eye* **38**, 1979–1980 (2024).
3. Fonollosa, A. et al. Update on ocular manifestations of the main monogenic and polygenic autoinflammatory diseases. *Front. Ophthalmol.* **4**, 1337329 (2024).
4. Chennakesavalu, M. et al. Corneal lymphangiogenesis as a potential target in dry eye disease - a systematic review. *Surv. Ophthalmol.* **66**, 960–976 (2021).
5. Clare, G., Kempen, J. H. & Pavésio, C. Infectious eye disease in the 21st century—an overview. *Eye* **38**, 2014–2027 (2024).
6. Liang, S. et al. A structure-aware convolutional neural network for automatic diagnosis of fungal keratitis with in vivo confocal microscopy images. *J. Digit. Imaging* **36**, 1624–1632 (2023).
7. Qiu, J. et al. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI* **1**, AIoa2300221 (2024).
8. Li, J. et al. Integrated image-based deep learning and language models for primary diabetes care. *Nat. Med.* **30**, 2886–2896 (2024).
9. Lei, C. et al. Facial expression of patients with Graves' orbitopathy. *J. Endocrinol. Investig.* **46**, 2055–2066 (2023).
10. Shi, X. H. et al. Deep learning models for the screening of cognitive impairment using multimodal fundus images. *Ophthalmol. Retina* **8**, 666–677 (2024).
11. Huang, W. et al. Auto diagnosis of Parkinson's disease via a deep learning model based on mixed emotional facial expressions. *IEEE J. Biomed. Health Inform.* **28**, 2547–2557 (2024).
12. Xue, C. et al. AI-based differential diagnosis of dementia etiologies on multimodal data. *Nat. Med.* **30**, 2977–2989 (2024).
13. Huang, X. et al. An intelligent diagnostic system for thyroid-associated ophthalmopathy based on facial images. *Front. Med.* **9**, 920716 (2022).
14. Moon, J. H. et al. Machine learning-assisted system using digital facial images to predict the clinical activity score in thyroid-associated orbitopathy. *Sci. Rep.* **12**, 22085 (2022).
15. Yan, C. et al. An ensemble deep learning diagnostic system for determining Clinical Activity Scores in thyroid-associated ophthalmopathy: integrating multi-view multimodal images from anterior segment slit-lamp photographs and facial images. *Front. Endocrinol.* **15**, 1365350 (2024).
16. Lei, C. et al. AI-assisted facial analysis in healthcare: from disease detection to comprehensive management. *Patterns* **6**, 101175 (2025).
17. Jin, B. et al. Simulated multimodal deep facial diagnosis. *Expert Syst. Appl.* **252**, 123881 (2024).
18. Hussein, R. et al. General data protection regulation (GDPR) toolkit for digital health. *Stud. Health Technol. Inform.* **290**, 222–226 (2022).
19. Moledina, M., Damato, E. M. & Lee, V. The changing landscape of thyroid eye disease: current clinical advances and future outlook. *Eye* **38**, 1425–1437 (2024).
20. Gontarz-Nowak, K., Szychlińska, M., Matuszewski, W., Stefanowicz-Rutkowska, M. & Bandurska-Stankiewicz, E. Current knowledge on graves' orbitopathy. *J. Clin. Med.* **10**, https://doi.org/10.3390/jcm10010016 (2020).
21. Zhou, Y. et al. Development and validation of a deep learning-based framework for automated lung CT segmentation and acute respiratory distress syndrome prediction: a multicenter cohort study. *EClinicalMedicine* **75**, 102772 (2024).
22. Shen, T., Gou, C., Wang, J. & Wang, F. Y. Simultaneous segmentation and classification of mass region from mammograms using a mixed-supervision guided deep model. *IEEE Signal Process. Lett.* **27**, 196–200 (2020).
23. Tariq, F. Allergic conjunctivitis: review of current types, treatments, and trends. *Life* **14**, https://doi.org/10.3390/life14060650 (2024).
24. Vogele, D. et al. Orbital Tumors-clinical, radiologic and histopathologic correlation. *Diagnostics* **12**, https://doi.org/10.3390/diagnostics12102376 (2022).
25. Mao, R., Zhou, L., Yu, L., Wang, T. & Wang, J. Dual-plane epicanthoplasty in Chinese blepharoplasty. *J. Cosmet. Dermatol.* **19**, 3323–3330 (2020).
26. Kongkasuriyachai, N. P., Palee, P., Prasitwattanaseree, S. & Mahakkanukrauh, P. Ancestry estimation using image analysis of orbital shapes from Thai and Japanese skulls. *Anthropol. Sci.* **128**, 19–26 (2020).
27. Sarabdeen, J. Protection of the rights of the individual when using facial recognition technology. *Heliyon* **8**, e09086 (2022).
28. Yin, D., Li, X., Liu, R., Zhang, L. & Zhan, Q. M. China's personal information protection law. *Bmj* **379**, e072619 (2022).
29. Merkley, J. & Sanders, B. National Biometric Information Privacy Act of 2020. GovInfo https://www.govinfo.gov/app/details/BILLS-116s4400is (2020).
30. Kim, B. R., Choi, J. Y. & Yoo, T. K. Multimodal large language models as assistance for evaluation of thyroid-associated ophthalmopathy. *Comput. Biol. Med.* **192**, 110301 (2025).
31. Bartley, G. B. & Gorman, C. A. Diagnostic criteria for Graves' ophthalmopathy. *Am. J. Ophthalmol.* **119**, 792–795 (1995).
32. Bartalena, L. et al. The 2021 European Group on Graves' orbitopathy (EUGOGO) clinical practice guidelines for the medical management of Graves' orbitopathy. *Eur. J. Endocrinol.* **185**, G43–g67 (2021).
33. Jiayu, C., Hong, H., Haipeng, Z. & Xuefei, S. Lightweight segmentation algorithm for TAO diseased areas based on DSE-Net. *J. Data Acquis. Process.* **38**, 915–925 (2023).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (82388101, 82271122), National Key R&D Program of China (2024YFB4710200, 2024YFB4710205), Science and Technology Commission of Shanghai Municipality (20DZ2270800), Shanghai Key Clinical Specialty, Shanghai Eye Disease Research Center (2022ZZ01003), Shanghai Municipal Commission of Health and Family Planning Project (2022XD006), Shanghai Jiao Tong University 2030 Initiative (WH510272301), Shanghai Three-Year Plan for the Inheritance and Innovative Development of Traditional Chinese Medicine (2-5-1), Hainan Province Science and Technology Special Fund (ZDYF2024LCLH004), Shanghai Jiao Tong University Hainan Institute's Independent Research Initiative (HRSJ-ZSZX-009). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

C.L., X.L., H.H., and H.Z. contributed to the conception and design of the study. C.L., J.D., P.S.P., S.W., and Y.R. contributed to data acquisition and interpretation. H.H., J.C., and C.X. built the deep learning model. C.Z., X.D., and S.L. performed the data analysis. C.L., C.Z., and J.C. drafted the manuscript. C.C.Y., S.S., J.D., H.H., X.S., X.L., and H.Z. jointly supervised the project and provided critical revisions to the manuscript. All authors had access to the study data, critically read and reviewed the manuscript, and approved the final version for publication.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-02310-w.

Correspondence and requests for materials should be addressed to Hong He, Xiaohong Liu or Huifang Zhou.