This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3573318

Dual-scale Swin Transformer via Feature Alignment and Adversarial Discrimination for Retinopathy of Prematurity Diagnosis

Shaobin Chen, Xinyu Zhao, Yiyao Liu, Hai Xie, Zhenquan Wu, Yingpeng Xie, Yongtao Zhang, Yafeng Li, Cheng Zhao, Tianfu Wang, Guoming Zhang, Baiying Lei, *Senior Member, IEEE*

***Abstract*— Retinopathy of prematurity (ROP) is a retinal vascular disease that primarily affects premature infants with low birth weight. It is a leading cause of childhood blindness worldwide, but it can often be effectively managed with appropriate and timely diagnosis and treatment. To address the impact of image style on model classification performance, this paper proposes a dual-scale Swin Transformer (DS-Swin-T) network for ROP. The network comprises three components: image synthesis (IS), feature alignment, and advanced adversarial learning. The IS module generates synthesis style images as an intermediate latent space between source and target styles, reducing style difference. The DS-Swin-T serves as the primary framework for image feature extraction. Detail and style encoders extract features in the shallow feature space, with detail and style losses aligning these features to ensure consistency across styles. To extract rich style-invariant features and ensure consistent classification within the same category, adversarial learning is applied in the advanced feature space. Finally, feature fusion units process dual-scale classification representations. Our method achieves an average accuracy of 97.91% on the source style dataset. When transferred to other target style datasets, our method effectively mitigates the performance degradation caused by style difference, reaching a maximum average accuracy of 93.66%. Extensive experiments demonstrate the effectiveness of our method.**

***Index Terms*— Retinopathy of prematurity, Swin Transformer, Image Synthesis, Adversarial Learning, Feature Alignment**

This work was supported partly by Shenzhen Medical Research Funds (Nos. C2301005 and A2403020), National Natural Science Foundation of China (Nos. 62271328, U22A2024, 82301269, 82401315, 62376164 and 62301329), National Natural Science Foundation of Guangdong Province (No. 2024A1515011950), Shenzhen Science and Technology Program (Nos. JCYJ20241202124202004, JCYJ20240813142111015, JCYJ20220818095809021 and JCYJ20240813152703005), Guangdong Basic and Applied Basic Research Foundation (2024A1515011676), Shenzhen Key Medical Discipline Construction Fund (No.SZXK038), Shenzhen Fund for Guangdong Provincial High level Clinical Key Specialties(No.SZGSP014), Sanming Project of Medicine in Shenzhen (No.SZSM202311018) and Medical-Engineering Interdisciplinary Research Foundation of Shenzhen University (Project no.2023YG003). (Corresponding author: Guoming Zhang, Baiying Lei).

Shaobin Chen, Yiyao Liu, Hai Xie, Yingpeng Xie, Yongtao Zhang, Yafeng Li, Cheng Zhao, Tianfu Wang and Baiying Lei are with the National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Marshall Laboratory of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen, China (e-mail:13531194616@163.com, liuyiyao0916@163.com, shehare@szu.edu.cn, xieyingpeng2017@email.szu.edu.cn, zhangyongtao2021@email.szu.edu.cn, 1247969403@qq.com, chengzhao@szu.edu.cn, tfwang@szu.edu.cn, leiby@szu.edu.cn).(Co-first authors :Shaobin Chen and Xinyu Zhao)

Xinyu Zhao, Zhenquan Wu and Guoming Zhang are with Shenzhen Eye Hospital, Shenzhen Eye Medical Center, Southern Medical University, Shenzhen, China (e-mail:zhaoxinyu1@sz-eyes.com, wuzhenquan@sz-eyes.com, zhangguoming@sz-eyes.com)

I. Introduction

Retinopathy of prematurity (ROP) is an abnormal proliferative eye disease occurring in premature infants with low birth weight. Its main manifestations include retinal vascular fibrosis, proliferation, and contraction, potentially leading to retinal detachment, glaucoma, cataracts, and blindness. ROP is a significant cause of childhood blindness and visual impairment [1] [2]. Early ROP studies primarily focus on detecting plus disease, with accuracy heavily reliant on vessel segmentation. Segmentation errors can be amplified in subsequent measurements [3]–[5]. Research on aggressive ROP (A-ROP) is relatively limited [6], and the training processes for automatic diagnosis of ROP and A-ROP are complex, often requiring staged detection or additional large networks for training [7], [8]. In 2021, the International ROP Expert Committee completed the third revision of the international classification of ROP (ICROP3) [6]. According to ICROP3, conventional ROP is classified into five stages based on disease progression, with key pathological features such as a demarcation line between vascular and avascular zones, raised demarcation lines, widening with ridged changes, fibrous tissue proliferation, and partial or complete retinal detachment. Another severe form, A-ROP, is characterized by unusually rapid neovascularization, progressing quickly without following the typical ROP development. As shown in Fig.1, variations in device settings, light intensity, shooting angles, and gestational age can result in fundus images with distinct styles. Despite these difference, the pathological features and semantics remain coupled.

The automatic diagnosis of ROP faces two main challenges:

1) As shown in Fig.1, the style difference phenomenon places high demands on the model's feature extraction capabilities, often leading to decreased classification performance.

Authorized licensed use limited to: Tsinghua University. Downloaded on November 26,2025 at 06:38:37 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3573318

2) The early symptoms of ROP are mild and easily missed, while early A-ROP symptoms resemble certain ROP stages, increasing the risk of misdiagnosis.

These challenges complicate the extraction of key disease features by the model. Several style adaptation techniques have been proposed to address the issue of style distribution shift [9], [10]. The correlation between pathological features and semantics tends to be style-invariant and less affected by style shifts, making it beneficial for reducing these effects.

Convolutional neural networks (CNNs) have made remarkable contributions to the field of computer vision due to their ability to hierarchically extract deep discriminative features through convolutional operations, achieving powerful image representations [11]–[14]. CNNs primarily focus on local image features due to their local connectivity and the sliding window mechanism of convolutional kernels. This makes it challenging for them to capture long-range dependencies and global information. In contrast, vision transformers excel at building long-range dependencies, enabling them to effectively capture global image information [15], [16]. However, vanilla vision transformers [17] may overlook intrinsic pixel-level structural features within patches, resulting in the loss of shallow features like edges and lines. Techniques like the overlapping patch embedding proposed by Xie et al. [18] and the shift window multi-head self-attention mechanism in Swin Transformer (Swin-T) [19], along with patch merging layers, partially address this by enabling information exchange between adjacent non-overlapping windows and introducing positional information. Yet, they still struggle to retain pixel-level details within each patch [18]. To enhance image feature representation, CrossViT [20] introduces a dual-branch vision transformer that combines image patches of varying scales. Similarly, DS-TransUNet by Lin et al. [21] employs a two-scale encoder based on Swin-T, extracting both coarse-grained and fine-grained features at different semantic levels. For addressing performance degradation due to style difference, unsupervised style adaptation is commonly used. This involves image-to-image alignment [22], [23] and adversarial learning techniques for transformation [24] and feature alignment [25].

On the basis of the above researches, we propose a dual-scale Swin Transformer (DS-Swin-T) based on feature alignment and adversarial discrimination for the ROP diagnosis. First, the diversity of styles among ROP images poses a challenge to the model's feature extraction capability. To mitigate the impact of style shifts on diagnostic performance, we use the CycleGAN [22] framework to generate synthesis style images with consistent detail information from the source style images but with different style characteristics. However, using the native CycleGAN can result in some unreasonable structures that may interfere with the diagnosis of disease. Therefore, we make targeted adjustments to the model's input. We observe that the grayscale information in ROP images effectively highlights some clinical diagnostic characteristics, such as dilation and tortuosity of retinal vessels. To alleviate or avoid unreasonable structures in synthetic images that may interfere with diagnosis, we use the grayscale information from the source style images to construct an auxiliary style image as input to the generator. This approach benefits the generator by better preserving the detail information of the source style images, allowing it to focus more on style transformation, thus generating synthesis style images with consistent detail information and different style characteristics. Secondly, to further achieve feature alignment between images from different styles, we design a style encoder and a detail encoder. These encoders aim to align the detail information and style information of images with different styles in the feature space. This alignment reduces the statistical distribution gap between different styles, thereby minimizing styles discrepancy and mitigating the styles shift problem. Finally, to address the last challenge of automatic ROP diagnosis mentioned above, we develop a DS-Swin-T model to extract both coarse-grained and fine-grained feature representations of images. The model effectively enhances its perception of critical lesion features by integrating dual-scale classification representations through a feature fusion unit (FFU). The innovations presented in this paper to address the aforementioned challenges are as follows:

1) **Style Shift Alleviation:** We introduce an image synthesis (IS) module, detail and style feature alignment (DSFA) for shallow-level features, and advanced feature adversarial learning (AFAL). The IS module generates synthesis style images, creating an intermediate latent space between source and target styles. DSFA ensures consistency in style and detail information across different style images using dedicated encoders. AFAL employs adversarial learning to extract style-invariant features, aligning classification results of synthesis and target style images with those of the source.
2) **Enhanced Feature Extraction:** We utilize a DS-Swin-T for feature extraction to explore multi-scale contextual feature representations. The dual-scale features are effectively fused using the feature fusion unit.

II. Related work

A. Automatic diagnosis method for ROP

Currently, some studies have explored predicting the likelihood of ROP occurrence based on biometric information such as an infant's gestational age, birth weight, sex, and mode of delivery, using traditional regression analysis for prediction [26]–[28]. However, these predictive models precisely overlook the retinal condition closely related to the diagnosis of ROP, which presents certain limitations [29]. In recent years, automatic ROP diagnostic models based on color fundus photography have achieved some promising results [30]–[32]. Most research has focused on identifying plus disease [30], [33]. These methods attempt to measure statistical data such as the diameter and curvature of the retinal vessel, as the presence and severity of plus disease are judged based on vascular abnormalities. Therefore, the accuracy of vessel measurements heavily depends on the segmentation results, meaning that errors in segmentation may be magnified in subsequent measurements. In recent years, an increasing number of studies have begun to focus on other clinical features of ROP [7], [8], [34]. To address the issue of current deep learning models being too complex for deployment on detection devices, Xie

Authorized licensed use limited to: Tsinghua University. Downloaded on November 26,2025 at 06:38:37 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3573318

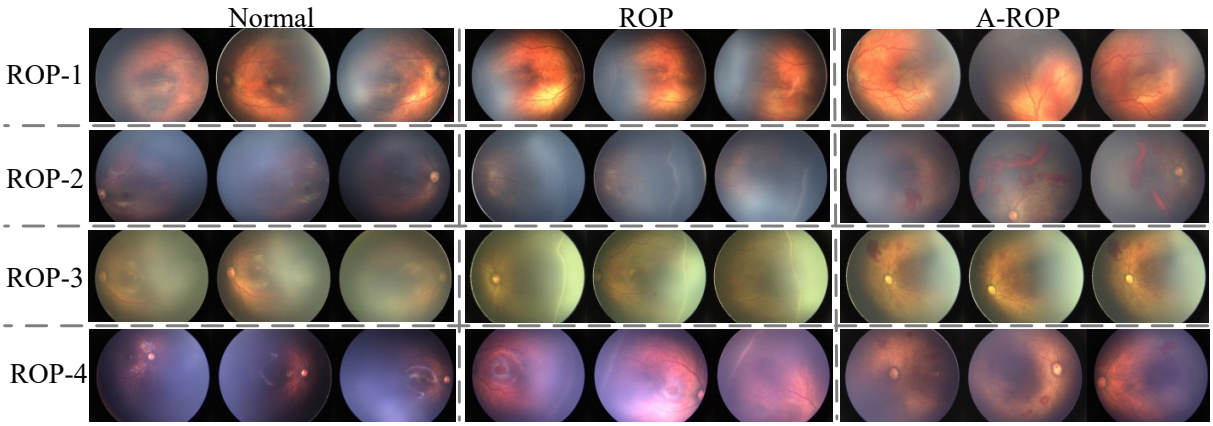


Fig. 1. Style difference phenomenon in ROP RetCam fundus images. According to clinical experience, retinal images captured using the RetCam device from premature infants of varying gestational ages exhibit distinct stylistic appearances, with images from relatively older premature infants tending to have a purplish tint.

et al. [7] proposed a lightweight multi-level dense knowledge distillation method based on adversarial learning for the automatic diagnosis of ROP. Peng et al. [34] proposed a three-streams deep network with feature-level fusion and ordinal classification strategy for automatic ROP staging. Hu et al. [5] proposed two sub-networks connected by a feature aggregation operator to achieve ROP screening and severity grading. Furthermore, Wu et al. [29] developed a deep learning model combining color fundus photography and clinical characteristic to predict ROP occurrence and severity grading. Currently, existing research on automatic ROP diagnosis based on color fundus photography rarely focuses on A-ROP. A-ROP is defined by the rapid development of pathological neovascularization and severe plus disease without progression through the typical stages of ROP, necessitating timely treatment. Moreover, these studies do not address the impact of different image styles on model performance. Therefore, to address the shortcomings of existing research, we have constructed an image synthesis module to generate synthesis style images, reducing distribution difference between different styles and mitigating performance degradation caused by style variations. Additionally, we enhance the model's ability to perceive key lesion features through the DS-Swin-T architecture.

B. Image synthesis and feature alignment

Image-to-image translation involves learning the mapping between input and output images using sufficient training data. Recently, generative models have advanced significantly, evolving from variational autoencoders [35], [36] to generative adversarial networks [37], [38], and more recently, diffusion models [39]. GAN-based methods have been applied to various medical images, including magnetic resonance images [40], [41], computed tomography images [42]–[44], ultrasound images [45], and retinal fundus images [46], [47]. The main challenges faced by generative models include the requirement for large-scale initial data, model operability, and convergence in model training. Model operability has been well-studied, with many solutions like Conditional GAN [48], [49], StyleGAN [50], CycleGAN [22], and ScoreGAN [51]. GAN-based IS methods can map source style images to an intermediate latent space, reducing the distribution gap between source and target style images. However, the requirement for large amounts of initial data remains a significant challenge in the application of GANs. Unlike these studies, we do not use the source style image directly as input. Instead, we construct an auxiliary style image based on the source style to facilitate easier convergence during model training, focusing more on style transformation while preserving source style details.

Previous image transformation techniquess [22], [52] can be seen as pixel-level style distribution alignment. CNN-based style adaptation methods have been extensively studied in image recognition [53]–[56]. For example, Tsai et al. [57] used adversarial learning for output space style adaptation, while Chen et al. [58] proposed GIO-Ada for both input and output levels. Zhang et al. [59] applied style discriminators and adversarial learning across multiple CNN feature extraction blocks. Based on these studies, we use feature alignment and adversarial learning in the dual-scale feature space extracted by the DS-Swin-T to capture style-invariant features, aiming to reduce inter-domain difference and alleviate domain shift issues.

III. METHOD

A. Overall framework of network

To address the issue of model classification performance being impacted by image style difference, we propose a feature alignment and adversarial discriminative method using DS-Swin-T, specifically for ROP classification. Our method includes the image synthesis (IS) module, detail and style feature alignment (DSFA), advanced feature adversarial learning (AFAL), and feature fusion unit (FFU). The overall framework is illustrated in Fig.2. Key symbols used in our method are explained in Table I.

Authorized licensed use limited to: Tsinghua University. Downloaded on November 26,2025 at 06:38:37 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3573318

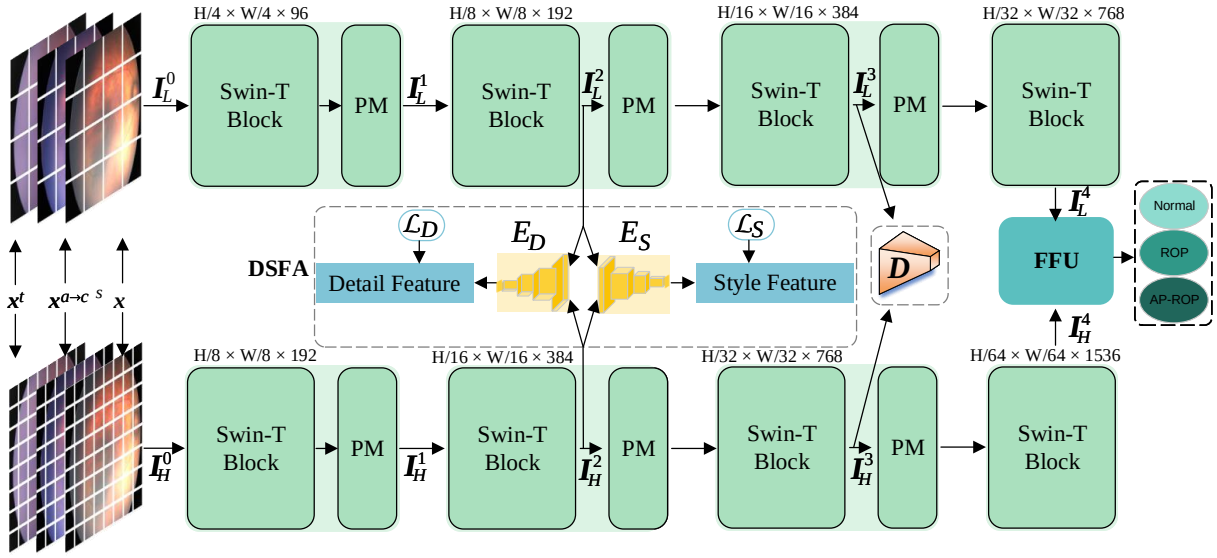


Fig. 2. Overall architecture of our proposed method.PM:Patch Merging, is an operation used in Swin-T for downsampling, enabling the construction of hierarchical feature representations. A DS-Swin-T is employed in the classifier for feature extraction, aiming to explore multi-scale contextual feature representation while preserving local connectivity. In the low-dimensional feature space, DSFA ensures inter-style feature alignment, while in the advanced feature space, adversarial learning enhances the model's capability of extracting style-invariant features, achieving advanced feature alignment across different style images. Finally, effective fusion of dual-scale features is achieved through the FFU.

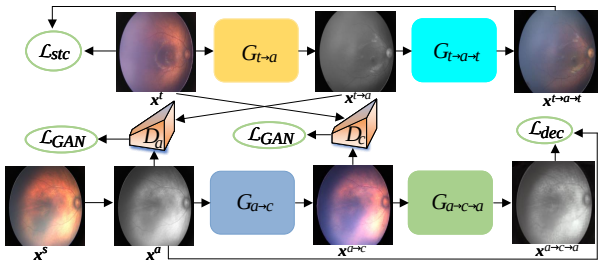


Fig. 3. Illustration of IS module. The IS module is used to generate a synthesis style image that is consistent with the source style image in terms of detail information but with different style features. The synthesis style image serves as an intermediate latent space between the source style image and the target style image, thereby reducing the inter-style difference.

TABLE I
SYMBOLS USED IN THIS STUDY

Symbol	Meaning
$\boldsymbol{I}_S \in \mathcal{R}^{3\times H\times W}$	Source style image set
$\boldsymbol{I}_A \in \mathcal{R}^{3\times H\times W}$	Auxiliary style image set
$\boldsymbol{I}_T \in \mathcal{R}^{3\times H\times W}$	Target style image set
$\boldsymbol{x}^a \in \boldsymbol{I}_A$	Auxiliary style image
$\boldsymbol{x}^t \in \boldsymbol{I}_T$	Target style image
$\boldsymbol{x}^s \in \boldsymbol{I}_S$	Source style image
$\boldsymbol{x}^{a\to c}$	Synthesis style image
$\boldsymbol{x}^{a\to c\to a}$	Auxiliary style reconstructed image
$\boldsymbol{x}^{t\to a}$	Synthesized auxiliary style image
$\boldsymbol{x}^{t\to a\to t}$	Target style reconstructed image
$G_{a\to c}$	Synthesis style generator
$G_{a\to c\to a}$	Auxiliary style reconstruction generator
$G_{t\to a}$	Synthesized auxiliary style generator
$G_{t\to a\to t}$	Target style reconstruction generator
D_a	Auxiliary style image discriminator
D_c	Synthesis style image discriminator
D	Advanced feature discriminator
$\mathcal{L}_{GAN}$	Adversarial loss
$\mathcal{L}_{stc}$	Style consistency loss
$\mathcal{L}_{dec}$	Detail consistency loss
$\mathcal{L}_D$	Detail feature alignment loss
$\mathcal{L}_S$	Style feature alignment loss
$\mathcal{L}_C$	Binary cross-entropy loss

B. IS module

To address style difference in both inter-class and intra-class variations due to the diverse retinal image sources, which often degrade classification performance, we implement a GAN-based IS mechanism. This mechanism generates synthesis images that maintain the detail of source images while having distinct style features, as shown in Fig.3. These synthesis images act as an intermediate latent space, reducing statistical distribution difference between styles and mitigating the style discrepancy issue. The IS module is an improvement upon the CycleGAN [22]. Unlike vanilla style transfer GANs, which use source and target style images as inputs, our approach uses an auxiliary style image derived from the grayscale information of source images. We convert the original RGB images to single-channel grayscale, then replicate and concatenate it to form a three-channel auxiliary style image. This ensures the synthesis images closely resemble the source images in detail. Assume the source style image set is $\boldsymbol{I}_S \in \mathcal{R}^{3\times H\times W}$, the auxiliary style image set is $\boldsymbol{I}_A \in \mathcal{R}^{3\times H\times W}$, and the target style image set is $\boldsymbol{I}_T \in \mathcal{R}^{3\times H\times W}$. For an auxiliary style image $\boldsymbol{x}^a$, the synthesis style generator $G_{a\to c}$ integrates the style of the target style images with the details of the source images to generate the synthesis style image $\boldsymbol{x}^{a\to c}$. The synthesis style image discriminator D_c works against $G_{a\to c}$ to distinguish between $\boldsymbol{x}^{a\to c}$ and the target style image $\boldsymbol{x}^t$. Additionally,

Authorized licensed use limited to: Tsinghua University. Downloaded on November 26,2025 at 06:38:37 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3573318

we use an auxiliary style reconstruction generator $G_{a\to c\to a}$ to closely match the auxiliary style reconstructed image $\boldsymbol{x}^{a\to c\to a}$ with $\boldsymbol{x}^{a}$, ensuring that the synthesis image retains the source details through the detail consistency loss $\mathcal{L}_{dec}$. For a target style image $\boldsymbol{x}^{t}$, we employ the synthesized auxiliary style generator $G_{t\to a}$ to create the synthesized auxiliary style image $\boldsymbol{x}^{t\to a}$. The auxiliary style image discriminator D_a competes with $G_{t\to a}$ to differentiate between $\boldsymbol{x}^{t\to a}$ and $\boldsymbol{x}^{a}$. Finally, the target style reconstruction generator $G_{t\to a\to t}$ takes $\boldsymbol{x}^{t\to a}$ as input to produce the target style reconstructed image $\boldsymbol{x}^{t\to a\to t}$, ensuring the synthesis image retains the target style through the style consistency loss $\mathcal{L}_{stc}$. The loss function of the IS module includes two adversarial losses, detail consistency loss, and style consistency loss, as shown in (1), (2), (3), and (4).

$$\mathcal{L}_{dec} = \mathbb{E}_{\boldsymbol{x}^{a}\sim \boldsymbol{I}_{A}}\left[\| \boldsymbol{x}^{a\to c\to a} - \boldsymbol{x}^{a} \|_{1}\right] \tag{3}$$

$$\mathcal{L}_{stc} = \mathbb{E}_{\boldsymbol{x}^{t}\sim \boldsymbol{I}_{T}}\left[\| \boldsymbol{x}^{t\to a\to t} - \boldsymbol{x}^{t} \|_{1}\right] \tag{4}$$

Equation (5) is the total loss function of the IS module, where λ_1 is the balancing coefficient, and it is set to 10 in this study.

C. Dual-scale Swin-T

Image patches of different scales can complement each other in feature extraction. Larger patches capture coarse-grained features, while smaller patches identify fine-grained details. Inspired by this concept, we use a DS-Swin-T for feature extraction to explore multi-scale contextual feature representation, maintaining local connectivity to enhance model robustness and improve classification performance. Although patches of different scales may lose some local continuity, fusing the two encoder scales provides implicit overlap, promoting feature interactions and preserving local continuity around patches. Specifically, this study employs two independent Swin-T branches. The initial patch sizes in the linear projection embedding layer are 4 and 8, with embedding feature dimensions of 96 and 192, respectively. These parallel Swin-T branches perform fine-grained and coarse-grained feature extraction on the patch representation vectors $\boldsymbol{I}_L^0$ and $\boldsymbol{I}_H^0$.

D. Detail and style feature alignment

The source style images possess unique style features and share detail information with the synthesis images. In contrast, the target style images have unique detail information and share style features with the synthesis images. This study employs a detail encoder E_D and a style encoder E_S to extract detail and style features from the low-dimensional features $\boldsymbol{I}_L^2$ and $\boldsymbol{I}_H^2$ of the DS-Swin-T. The E_D extracts detail features from the synthesis and source style images, while the E_S extracts style features from the synthesis and target style images. To ensure alignment, the L1 loss is used to directly align detail and style features in the feature space, making the style features of the synthetic images similar to those of the target style, and the detail features similar to those of the source style. The detail feature alignment loss function $\mathcal{L}_D$ and the style feature alignment loss function $\mathcal{L}_S$ are defined in (6) and (7), where E refers to the DS-Swin-T. Fig.4 illustrates the structure of the detail and style encoders.

$$\mathcal{L}_D = \mathbb{E}_{\boldsymbol{x}^{s}\sim \boldsymbol{I}_{S},\ \boldsymbol{x}^{a\to c}\sim \boldsymbol{I}_{A\to C}}\left[\| E_D\left(E\left(\boldsymbol{x}^{s}\right)\right) - E_D\left(E\left(\boldsymbol{x}^{a\to c}\right)\right) \|_{1}\right] \tag{6}$$

$$\mathcal{L}_S = \mathbb{E}_{\boldsymbol{x}^{t}\sim \boldsymbol{I}_{T},\ \boldsymbol{x}^{a\to c}\sim \boldsymbol{I}_{A\to C}}\left[\| E_S\left(E\left(\boldsymbol{x}^{a\to c}\right)\right) - E_S\left(E\left(\boldsymbol{x}^{t}\right)\right) \|_{1}\right] \tag{7}$$

E. Advanced feature adversarial learning

To enhance the model's ability to extract style-invariant features and ensure that the classification results of synthesis style images align with those of source style images, this study employs unsupervised adversarial learning. This approach extracts style-invariant features in the advanced feature space. Specifically, the synthesis style images are treated as unlabeled samples, and an advanced feature discriminator *D* is constructed for style discrimination on advanced features. Fig.5 illustrates the structure of the discriminator. If the advanced features from the synthesis style images are style-invariant, meaning they contain similar detail information as those from source style images, the discriminator cannot correctly distinguish between the features of source and synthesis style images. The discriminator takes paired high-dimensional features $\boldsymbol{I}_L^3$ and $\boldsymbol{I}_H^3$ as input, and the loss function is shown in (8). To ensure that $\boldsymbol{x}^{a\to c}$ has consistent classification results with $\boldsymbol{x}^{s}$, the adversarial loss in (9) is used to confuse the discriminator, making it can classify $\boldsymbol{x}^{a\to c}$ as real images. This aligns the advanced feature distributions of $\boldsymbol{x}^{a\to c}$ with those of $\boldsymbol{x}^{s}$.

$$\mathcal{L}_2 = \frac{1}{N}\left(\sum_{\boldsymbol{x}^{s}}\mathcal{L}_C\left(\boldsymbol{d}_s,1\right) + \sum_{\boldsymbol{x}^{a\to c}}\mathcal{L}_C\left(\boldsymbol{d}_c,0\right) + \sum_{\boldsymbol{x}^{t}}\mathcal{L}_C\left(\boldsymbol{d}_t,0\right)\right), \tag{8}$$

$$\mathcal{L}_3 = \frac{1}{N}\left(\sum_{\boldsymbol{x}^{a\to c}}\mathcal{L}_C\left(\boldsymbol{d}_c,1\right) + \sum_{\boldsymbol{x}^{t}}\mathcal{L}_C\left(\boldsymbol{d}_t,1\right)\right), \tag{9}$$

where *N* denotes the total number of samples, $\mathcal{L}_C$ is the binary cross-entropy loss function, $\boldsymbol{d}_s$, $\boldsymbol{d}_t$, and $\boldsymbol{d}_c$ are the outputs of the discriminator for $\boldsymbol{x}^{s}$, $\boldsymbol{x}^{t}$, and $\boldsymbol{x}^{a\to c}$, respectively.

F. Feature fusion Unit

To capture long-term dependencies and global context connections between different scale features, a novel FFU is proposed. It employs a transformer encoder for effective interaction between dual-scale features, rather than directly concatenating the dual-scale features $\boldsymbol{I}_L^4$ and $\boldsymbol{I}_H^4$. Unlike pure vision transformers, which downsample at a fixed rate from the start and keep feature map sizes constant, Swin-T uses a hierarchical approach to gradually downsample image features without classification token. As shown in Fig.6, since $\boldsymbol{I}_L^4$ and $\boldsymbol{I}_H^4$ are based on rectangular three-dimensional features, they first undergo average pooling and are then flattened to obtain global abstract information $\boldsymbol{l}_0$ and $\boldsymbol{h}_0$, serving as classification representations. The features $\boldsymbol{I}_L^4$ and $\boldsymbol{I}_H^4$ can be represented as patch representation vector sequences $\boldsymbol{F}_l$ and $\boldsymbol{F}_h$, respectively.

$$\boldsymbol{F}_l = \left[\boldsymbol{l}_1;\boldsymbol{l}_2;\cdots;\boldsymbol{l}_{\left(\frac{H}{32}\times\frac{W}{32}\right)}\right] \in \mathcal{R}^{\left(\frac{H}{32}\times\frac{W}{32}\right)\times 768} \tag{10}$$

Authorized licensed use limited to: Tsinghua University. Downloaded on November 26,2025 at 06:38:37 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3573318

$$\mathcal{L}_{GAN}\left(G_{a\to c}, D_c, \boldsymbol{x}^a, \boldsymbol{x}^t\right) = \mathbb{E}_{\boldsymbol{x}^t\sim \boldsymbol{I}_T}\left[\log D_c\left(\boldsymbol{x}^t\right)\right] + \mathbb{E}_{\boldsymbol{x}^a\sim \boldsymbol{I}_A}\left[\log\left(1 - D_c\left(G_{a\to c}\left(\boldsymbol{x}^a\right)\right)\right)\right] \tag{1}$$

$$\mathcal{L}_{GAN}\left(G_{t\to a}, D_a, \boldsymbol{x}^a, \boldsymbol{x}^t\right) = \mathbb{E}_{\boldsymbol{x}^a\sim \boldsymbol{I}_A}\left[\log D_a\left(\boldsymbol{x}^a\right)\right] + \mathbb{E}_{\boldsymbol{x}^t\sim \boldsymbol{I}_T}\left[\log\left(1 - D_a\left(G_{t\to a}\left(\boldsymbol{x}^t\right)\right)\right)\right] \tag{2}$$

$$\mathcal{L}_1 = \mathcal{L}_{GAN}\left(G_{a\to c}, D_c, \boldsymbol{x}^a, \boldsymbol{x}^t\right) + \mathcal{L}_{GAN}\left(G_{t\to a}, D_a, \boldsymbol{x}^a, \boldsymbol{x}^t\right) + \lambda_1\left(\mathcal{L}_{dec} + \mathcal{L}_{stc}\right) \tag{5}$$

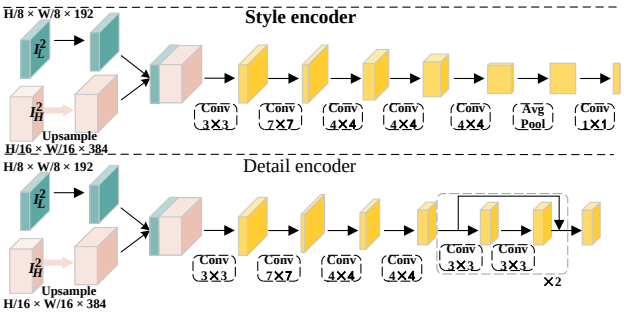


Fig. 4. Illustration of style encoder and detail encoder. Both encoders take two low-dimensional features from the DS-Swin-T as input, where the style encoder is used to extract style features from the synthesis image and the target style image, while the detail encoder is used to extract detail features from the synthesis image and the source style image.

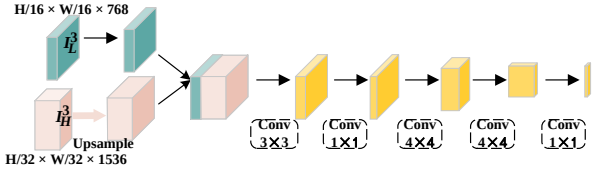


Fig. 5. Illustration of advanced feature alignment discriminator. The discriminator utilizes two advanced features from the DS-Swin-T as input to discriminate between different style images, thereby enhancing the model's ability to extract style-invariant features.

$$\boldsymbol{F}_h = \left[\boldsymbol{h}_1; \boldsymbol{h}_2; \cdots; \boldsymbol{h}_{\left(\frac{H}{64}\times\frac{W}{64}\right)}\right] \in \mathcal{R}^{\left(\frac{H}{64}\times\frac{W}{64}\right)\times 1536} \tag{11}$$

The $\boldsymbol{l}_0$ and $\boldsymbol{h}_0$ are then transformed into $\boldsymbol{l}_0^*$ and $\boldsymbol{h}_0^*$ by a linear transformation layer, respectively.

$$\boldsymbol{l}_0^* = LT\left(Flatten\left(Avgpool\left(\boldsymbol{I}_L^4\right)\right)\right) \tag{12}$$

$$\boldsymbol{h}_0^* = LT\left(Flatten\left(Avgpool\left(\boldsymbol{I}_H^4\right)\right)\right) \tag{13}$$

where LT represents a linear transformation operation.

To achieve dual-scale feature fusion, the global abstract information $\boldsymbol{l}_0$ and $\boldsymbol{h}_0$ serve as mediators for information exchange, interacting with the input patch representation vector sequences $\boldsymbol{F}_h$ and $\boldsymbol{F}_l$, respectively. The refined vectors $\boldsymbol{l}_0^*$ and $\boldsymbol{h}_0^*$ are combined with $\boldsymbol{F}_h$ and $\boldsymbol{F}_l$ to form new representation vector sequences, which are then input into the transformer encoder to calculate global self-attention.

$$\boldsymbol{F}_h^* = Transformer\left(\left[\boldsymbol{l}_0^*; \boldsymbol{h}_1; \cdots; \boldsymbol{h}_{\frac{H}{64}\times\frac{W}{64}}\right]\right) \in \mathcal{R}^{\left(\frac{H}{64}\times\frac{W}{64}+1\right)\times 1536} \tag{14}$$

$$\boldsymbol{F}_l^* = Transformer\left(\left[\boldsymbol{h}_0^*; \boldsymbol{l}_1; \cdots; \boldsymbol{l}_{\frac{H}{32}\times\frac{W}{32}}\right]\right) \in \mathcal{R}^{\left(\frac{H}{32}\times\frac{W}{32}+1\right)\times 768} \tag{15}$$

The interaction between $\boldsymbol{I}_L^4$ and $\boldsymbol{I}_H^4$ enables the fine-grained features of the small-scale branch to integrate coarse-grained information from the large-scale branch. Next, classification representations $\boldsymbol{l}_0^*$ and $\boldsymbol{h}_0^*$ are extracted from the processed representation sequences $\boldsymbol{F}_h^*$ and $\boldsymbol{F}_l^*$, respectively. These are dimensionally restored via a linear transformation layer and combined with their original global abstract information $\boldsymbol{l}_0$ and $\boldsymbol{h}_0$. The final FFU outputs the concatenation of these two classification representations:

$$Z_{out} = Concat\left(\left(LT\left(\boldsymbol{h}_0^*\right) + \boldsymbol{h}_0\right), \left(LT\left(\boldsymbol{l}_0^*\right) + \boldsymbol{l}_0\right)\right), \tag{16}$$

where *Concat* indicates the concatenation operation. At this stage, the classification representations retain their original branch information while integrating information from the other branch. The FFU output, which is the concatenation of these two representations, effectively fuses dual-scale features. Finally, the total loss function of the model is:

$$\mathcal{L} = \mathcal{L}_C\left(Z_{out}, y\right) + \mathcal{L}_D + \mathcal{L}_S + \mathcal{L}_2 + \mathcal{L}_3, \tag{17}$$

where y represents the label of labeled samples.

IV. Experiments and Results

A. Datasets and experimental settings

This study utilizes wide-field retinal images of premature infants captured by the RetCam system. The dataset contains a total of 14,433 images, including 7,494 normal, 1,698 A-ROP, and 5,241 ROP images. The specific distribution of our ROP dataset is shown in Table II. Based on the variations in image styles, we partition the data into four subsets, which are named ROP-1, ROP-2, ROP-3, and ROP-4. The specific styles of these four subsets are illustrated in Fig.1. This study is reviewed and approved by the Medical Ethics Committee of Shenzhen Eye Hospital (Ethical approval reference No.2023KYPJ091). In addition, we collect two publicly available fundus image datasets, iChallenge-GON (Refuge, Refuge2) [60] and iChallenge-GON2 (GAMMA) [61]. Each dataset consists of two categories: glaucoma and healthy, and the specific distribution of the datasets is shown in Table III.

TABLE II
Detailed distribution of the four subsets in ROP dataset

Dataset	ROP	A-ROP	Normal	Total
ROP-1	1992	568	3381	5941
ROP-2	1247	341	669	2257
ROP-3	331	549	1493	2373
ROP-4	1671	240	1951	3862

The proposed model and comparative methods are implemented on the PyTorch platform and trained on an NVIDIA TITAN RTX GPU (24GB) using a 5-fold cross-validation strategy. The evaluation metrics for classification tasks in this study include: accuracy (ACC), area under the ROC curve (AUC), precision (Pre), sensitivity/recall (Sen/Recall),

Authorized licensed use limited to: Tsinghua University. Downloaded on November 26,2025 at 06:38:37 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3573318

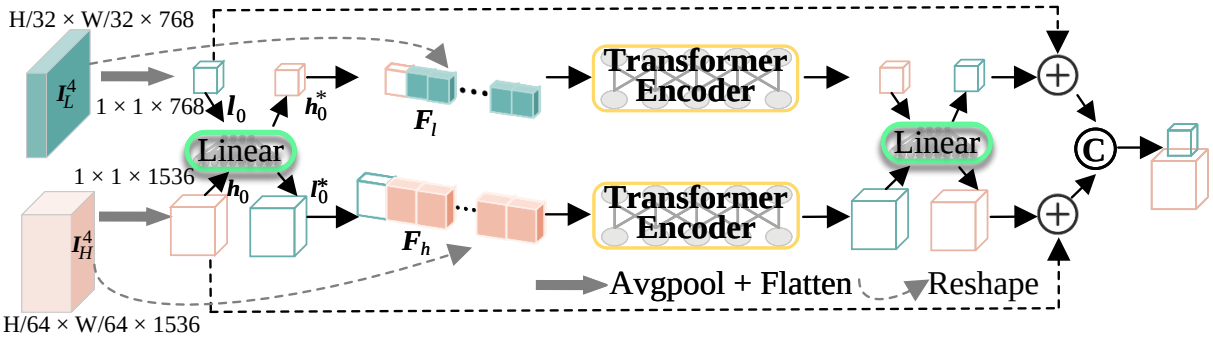


Fig. 6. Illustration of FFU. The dual-scale output features of the labeled data in the source style are fused through FFU to capture the long-term dependencies and global contextual connections between different scales.

TABLE III
ICHALLENGE-GON DATASET.

Source	Label	Amount	Total
iChallenge-GON	Healthy	1080	1200
	Glaucoma	120	
iChallenge-GON2	Healthy	50	100
	Glaucoma	50	

specificity (Spec), F1 score (F1), and Kappa. For multi-classification tasks, seven evaluation metrics are calculated for each category, and the model's overall performance is assessed using the unweighted average of these metrics across all categories. Specifically, for class i, TP_i, FP_i, TN_i, and FN_i represent the number of true positives, false positives, true negatives, and false negatives, respectively.

$$\mathrm{Pre}_i = \frac{TP_i}{TP_i + FP_i}, \mathrm{Sen}_i = \frac{TP_i}{TP_i + FN_i} \tag{18}$$

$$\mathrm{F1}_i = \frac{2 \times Pre_i \times Recall_i}{Sen_i + Recall_i}, \mathrm{Spec}_i = \frac{TN_i}{TN_i + FP_i} \tag{19}$$

Finally, the unweighted average of each evaluation metric in the multi-classification task is $M \in Pre, Sen, F1, Spec, Kappa, ACC, AUC$, which is defined as follows:

$$M_{macro} = \sum_{i=1}^{k} M_i / k, \tag{20}$$

where k represents the number of categories.

B. Image synthesis results

To ensure high-quality synthesis style images, which are crucial for classification performance, this study train the IS module separately. The synthesis style images generated by the vanilla CycleGAN and our IS module are shown in Fig.7. The vanilla CycleGAN is used, which takes images in the source and target styles as input during training to transfer the target style onto the source style images, creating synthesis style images. From the results in Fig.7, it can be seen that although the vanilla CycleGAN can achieve style transfer to some extent, it has significant shortcomings in preserving detailed information. For example, in the Normal category, the synthesis images generated by the vanilla CycleGAN perform poorly in retaining vascular details, failing to preserve the detail information of the source style images adequately. This issue is more pronounced in A-ROP, as A-ROP involves more pathological neovascularization. In some cases, the vanilla CycleGAN also performs poorly in style transfer, such as in the synthesis of A-ROP images. The analysis of the results indicates that for the ROP dataset, the synthesis style images generated by the vanilla CycleGAN differ significantly in detail from the source style images. These discrepancies can affect the model's ability to extract domain-invariant features, interfering with performance. Therefore, we improve the input to the generative model by constructing an auxiliary style image based on the source style image. The auxiliary image removes interference from the source style while preserving key clinical diagnostic features, allowing the generator to focus more on style transformation and better retain details from the source style image. Experimental results show that our IS module effectively maintains the detail of the source style images while successfully transferring the target style. Thus, the GAN-based IS module can serve as an intermediate latent space between source and target style images, reducing statistical distribution difference among intra-class style images, alleviating style shift issues, and enhancing classification performance.

C. Classification performance evaluation

To validate the effectiveness of the proposed method in this study, seven popular classification network are selected as comparison methods, including three CNN models: EfficientNet-B0 [62] (E-Net), ResNet50 [63], and ConvNeXt-T [64], and four transformer-related models: ViT [17], CrossViT-Ti [20], DeiT [65], and Swin-T [19]. ViT used the ViT-Base/16 architecture, DeiT used the ViT-Base/16 architecture as the student model and ResNet50 as the teacher model. When training on the ROP-1 dataset, all models can only access the ROP-1 dataset, while the other three subsets are inaccessible. Using ROP-1 as the source domain data,

Authorized licensed use limited to: Tsinghua University. Downloaded on November 26,2025 at 06:38:37 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3573318

TABLE IV
EXPERIMENTAL RESULTS OF ROP-1 DATASET.

Method	ACC	AUC	Pre	Sen	F1	Spec	Kappa
ResNet50	96.53±0.99	99.52±0.33	96.44±0.68	94.78±1.97	95.48±1.31	97.97±0.42	93.74±1.76
E-Net	93.00±1.60	99.13±0.19	90.16±2.42	93.17±1.66	91.03±2.56	96.00±1.03	87.42±2.92
ConvNeXt-T	93.60±4.05	98.76±1.16	93.48±3.89	89.53±6.52	90.86±5.72	96.42±2.22	88.46±7.27
ViT	95.77±0.67	99.45±0.32	95.20±0.93	93.21±2.40	94.03±1.45	97.54±0.45	92.35±1.26
CrossViT-Ti	96.55±1.30	99.58±0.34	96.46±0.99	94.49±3.39	95.26±2.45	97.95±0.71	93.75±2.38
DeiT	96.80±1.20	99.63±0.34	97.00±1.33	93.28±4.21	94.80±3.01	98.00±0.72	94.16±2.24
Swin-T	96.88±1.52	99.69±0.36	96.11±1.78	95.17±3.22	95.55±2.54	98.20±0.94	94.37±2.76
Proposed	**97.91±0.93**	**99.89±0.11**	**98.01±0.85**	**96.09±2.18**	**96.93±1.55**	**98.80±0.49**	**96.22±1.67**

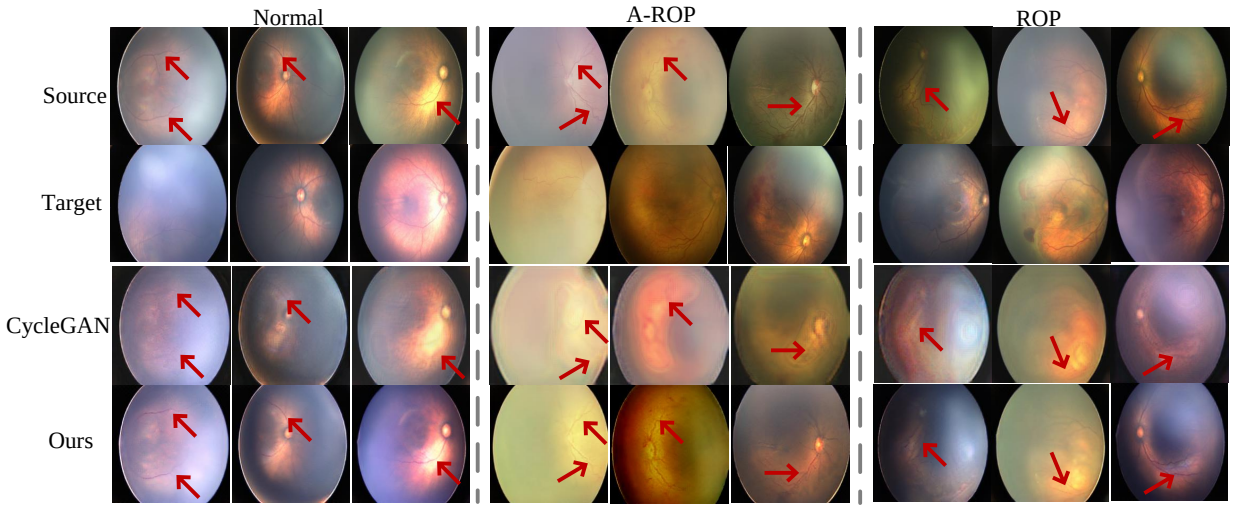


Fig. 7. Comparative results of ROP retinal images generated by CycleGAN and our IS module.

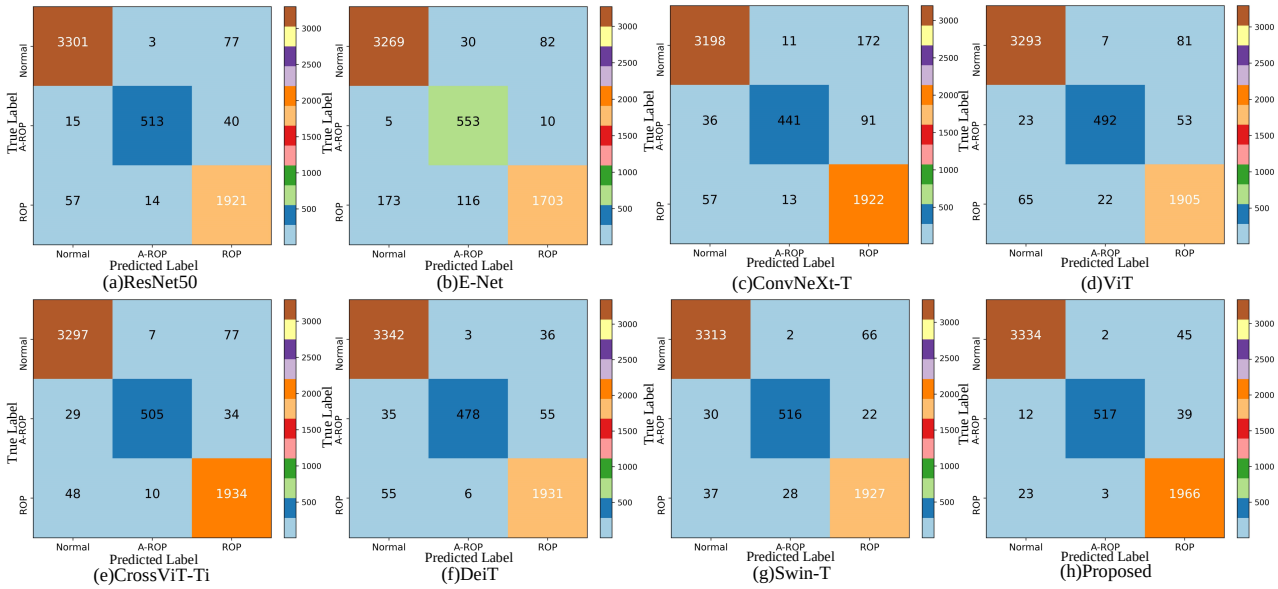


Fig. 8. Confusion matrixs of different models in the ROP classification task.

the other three subsets serve as corresponding target domain data, allowing the IS module to obtain three image style transformation generators. During training iterations, we use a strategy similar to random data augmentation, randomly selecting one of the three generators to create synthesis style images, enabling models to indirectly access the styles of other subsets. For a fair comparison with other models, synthesis style images are incorporated as labeled data directly into the supervised training of each comparison model.

Authorized licensed use limited to: Tsinghua University. Downloaded on November 26,2025 at 06:38:37 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3573318

1) Experimental results of ROP-1 dataset: We first conduct model training and validation on the ROP-1 dataset. The experimental results are shown in Table IV. To visually compare model performance across various metrics, we present the results as bar charts in Fig.9. The proposed method achieves the best performance in multiple metrics. Among the seven comparison methods, Swin-T performed the best, but our method shows improvements of 1.03% in Acc, 1.90% in Pre, and 1.85% in Kappa compared to Swin-T. This demonstrates that our model effectively extracts discriminative features, achieving superior ROP diagnosis performance.

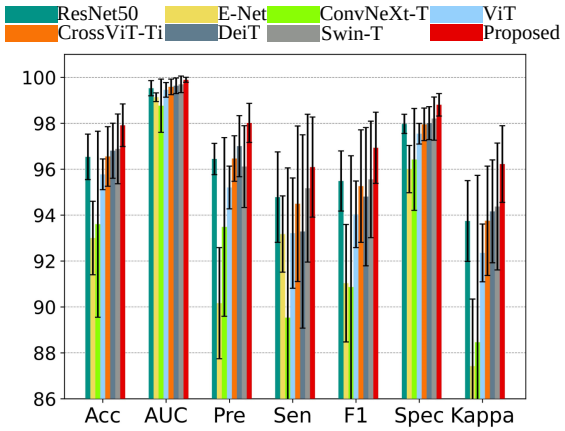


Fig. 9. Bar charts of various evaluation indicators under different methods.

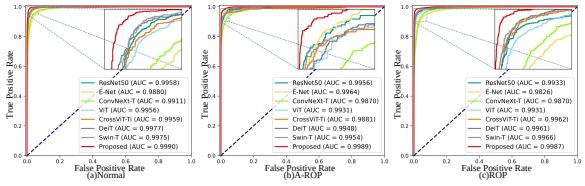


Fig. 10. ROC curves of different methods in different categories of the ROP classification task.

To further explore the performance of different categories in the ROP diagnosis task, we plot the receiver operating characteristic (ROC) curves for normal, ROP, and A-ROP. The results are shown in Fig.10, where the red curve represents our proposed method. Overall, our method performs better across all three categories. Although the difference in AUC are not particularly pronounced, our method's ROC curve is closer to the upper left corner, indicating superior performance. The confusion matrix in Fig.8 visually compares the classification results of our method and the seven comparison models. It shows that our method more effectively detects the ROP and A-ROP categories.

Additionally, we further analyze the experimental results by calculating the intra-class and inter-class distances of the features extracted by each model. Specifically, we first aggregate the features from the final feature extraction layer of the models. Before calculating the intra-class and inter-class distances between the sample features, we perform normalization on the extracted features. The comparison of inter-class and intra-class distances for features extracted by each model is shown in Table V. As can be observed, among the intra-class distances of features for the three categories, our proposed model exhibits the smallest intra-class distances compared to the comparison models, with A-ROP demonstrating the lowest intra-class distance. Regarding the inter-class distances of features among the three categories, our proposed model shows the largest inter-class distances compared to the comparison models, with the greatest inter-class distance observed between A-ROP and ROP. Our proposed method demonstrates superior capability in extracting both coarse-grained and fine-grained feature representations from images. This enhances ability to capture crucial lesion characteristics results in more compact intra-class feature clusters and greater inter-class feature discrimination.

Finally, to demonstrate the excellent feature extraction ability of the proposed method, we conducted a class activation mapping experiment on the model's output features. This verifies that the extracted features effectively distinguish prediction categories. The Grad-CAM visualizations for A-ROP and ROP are shown in Fig.11. Our model successfully learns the characteristics of the focal region, accurately predicting the target category.

2) Model Performance on subsets with varying styles: We further explore the model's performance across different dataset styles. The model parameters trained on the ROP-1 dataset are tested on ROP-2, ROP-3, and ROP-4 datasets. The results are shown in Table VI. As seen in Table IV, although the models perform well on ROP-1, there is a significant drop in performance when applied to the other subsets. For example, Acc decreases by 4.24% to 33.98% on ROP-2, 12.15% to 25.73% on ROP-3, and 12.46% to 37.15% on ROP-4. Our proposed method shows the smallest performance drop, with decreases of 4.24%, 12.15%, and 12.46%, respectively. The results in Table VI demonstrate that our method achieves the best performance across various classification metrics. The synthesis style images generated by the IS module serve as an intermediate latent space between source and target style images. Detail and style features are aligned in the shallow feature space, while adversarial learning in the high-level feature space further extracts style-invariant features. This helps reduce statistical distribution difference between images of different styles, mitigating performance decline due to style variations to a certain extent.

D. Ablation study

The network architecture proposed in this study includes the DS-Swin-T block (DSTB), IS, DSFA, AFAL, and FFU. To verify the effectiveness of these modules in enhancing classification performance, we conduct ablation experiments, with results shown in Table VII. Compared to using only single-scale Swin-T as the baseline, DS-Swin-T increases Acc by 0.12%. Embedding the FFU in DS-Swin-T further boosts Acc by 0.51%. Incorporating synthesis style images generated by the IS module into supervised training improves Acc by 0.41% over DS-Swin-T. However, simply using synthesis

Authorized licensed use limited to: Tsinghua University. Downloaded on November 26,2025 at 06:38:37 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3573318

TABLE V
COMPARISON OF INTRA-CLASS AND INTER-CLASS DISTANCES OF FEATURES EXTRACTED BY MODELS

Method	Intra-class			Inter-class			Average	
	Normal↓	A-ROP↓	ROP↓	Normal & A-ROP ↑	Normal & ROP ↑	A-ROP & ROP ↑	Intra Avg. ↓	Inter Avg. ↑
ResNet50	0.5593	0.2898	0.6282	0.4933	0.7834	0.7086	0.4924	0.6618
E-Net	0.4756	0.4697	0.4663	0.4727	0.4710	0.4681	0.4705	0.4706
ConvNeXt-T	0.6744	0.4876	0.7033	0.7585	0.7826	0.6493	0.6218	0.7301
ViT	0.6769	0.5587	0.5248	0.7028	0.9370	0.5039	0.5868	0.7146
CrossViT-Ti	0.6420	0.1991	0.6116	0.5776	0.7345	0.5149	0.4842	0.6090
DeiT	0.5838	0.2293	0.6014	0.4285	0.5559	0.4354	0.4715	0.4733
Swin-T	0.5622	0.0916	0.6052	0.5310	0.7276	0.5181	0.4197	0.5922
Proposed	**0.3696**	**0.0736**	**0.3752**	**0.8368**	**0.8559**	**0.9168**	**0.2728**	**0.8698**

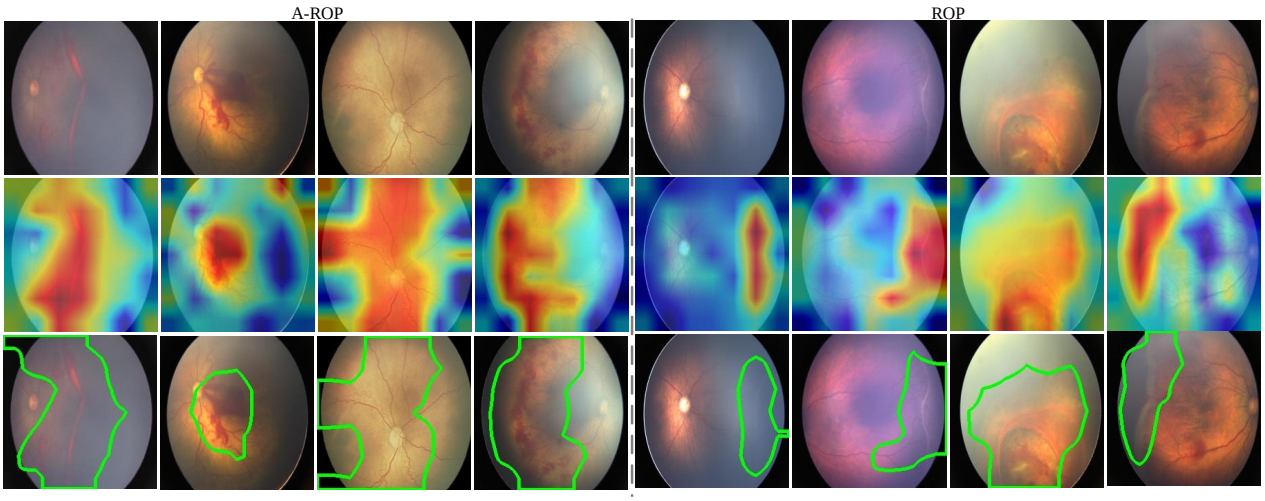


Fig. 11. GradCam visualization results of A-ROP and ROP under the proposed method

TABLE VI
EXPERIMENTAL RESULTS OF STYLE TRANSFER FROM ROP-1 DATASET TO OTHER DIFFERENTLY STYLED SUBSETS.

Test dataset	Method	ACC	AUC	Pre	Sen	F1	Spec	Kappa
ROP-2	ResNet50	86.04	97.43	83.91	89.46	85.92	93.31	77.22
	E-Net	59.02	84.43	61.38	67.65	58.63	81.87	39.39
	ConvNeXt-T	86.18	97.19	82.36	86.72	84.07	93.57	77.19
	ViT	85.33	96.43	83.99	85.44	84.41	92.39	75.50
	CrossViT-Ti	84.98	96.98	81.26	87.63	83.44	93.08	75.50
	DeiT	84.58	97.08	80.08	86.63	81.89	93.30	74.96
	Swin-T	90.21	97.83	87.50	88.87	88.14	94.87	83.35
	Proposed	**93.66**	**99.46**	**91.67**	**94.06**	**92.72**	**97.08**	**89.36**
ROP-3	ResNet50	79.69	88.24	69.60	72.92	70.86	90.25	63.83
	E-Net	68.44	81.65	60.07	63.31	59.87	86.29	48.11
	ConvNeXt-T	75.26	87.69	66.34	67.34	65.79	88.35	56.26
	ViT	70.04	82.99	64.16	65.73	63.17	85.72	49.16
	CrossViT-Ti	78.72	91.45	73.46	78.92	74.24	90.19	63.74
	DeiT	73.83	88.53	68.15	71.19	67.99	87.41	55.32
	Swin-T	76.70	85.89	68.44	67.60	67.76	85.81	55.73
	Proposed	**85.76**	**95.97**	**79.48**	**82.14**	**79.91**	**92.85**	**73.98**
ROP-4	ResNet50	77.68	90.97	81.49	73.23	76.51	85.96	59.03
	E-Net	55.85	75.54	51.67	61.44	54.26	73.75	22.61
	ConvNeXt-T	73.43	87.78	80.46	63.03	66.38	84.00	51.64
	ViT	70.74	80.80	73.32	67.02	69.62	81.73	46.43
	CrossViT-Ti	75.22	89.37	72.42	70.53	71.19	85.11	55.38
	DeiT	76.46	90.31	82.29	67.92	72.16	85.40	56.76
	Swin-T	79.93	93.90	81.78	75.50	77.55	87.96	63.81
	Proposed	**85.45**	**94.97**	**85.50**	**75.99**	**79.47**	**90.86**	**73.17**

images as supplementary data doesn't significantly enhance performance. We then used these images as an unlabeled dataset to evaluate DSFA and AFAL. Adding DSFA improved Acc by 0.54%, and adding AFAL improved it by 0.75%, compared to the baseline. Ultimately, utilizing all modules, our method improved Acc by 1.03% over the baseline.

E. Robustness to public datasets

To validate the effectiveness of our proposed method, we conduct experiments on two publicly available datasets related to the glaucoma classification task. As shown in Fig.12, a style shift phenomenon exists in these datasets. Consistent with previous experiments, we first use these datasets to train the IS module separately to obtain synthesis style images. The results are shown in Fig.13. The synthesis style images for normal and glaucoma fully retain the details of the source images while successfully transferring the style information of the target images. This demonstrates the effectiveness of the IS module, where the synthesis style images act as an intermediate latent space between source and target images, reducing statistical distribution difference and alleviating style shift issues.

For the classification task on the iChallenge datasets, we compare the proposed method with seven popular classification networks, using 10-fold cross-validation. We independently train on iChallenge-GON and iChallenge-GON2, with results shown in Table VIII. Using iChallenge-GON2 as the source style dataset, Swin-T performs best among the seven comparison methods. However, our method improves Pre by 2.86% and Spec by 4% compared to Swin-T. With iChallenge-GON as the source, DeiT achieves the top performance. Our

Authorized licensed use limited to: Tsinghua University. Downloaded on November 26,2025 at 06:38:37 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3573318

TABLE VII
MODULE ABLATION EXPERIMENTAL RESULTS.

Baseline	DSTB	FFU	IS	DSFA	AFAL	ACC	AUC	Pre	Sen	F1	Spec	Kappa
✓						96.88±1.52	99.69±0.36	96.11±1.78	95.17±3.22	95.55±2.54	98.20±0.94	94.37±2.76
✓	✓					97.00±0.70	99.58±0.33	95.90±1.15	95.91±1.67	95.84±1.02	98.34±0.49	94.61±1.26
✓	✓	✓				97.39±1.69	99.70±0.37	97.74±1.28	93.79±5.26	95.23±4.04	98.38±1.01	95.22±3.14
✓	✓	✓	✓			97.41±1.52	99.90±0.05	97.48±1.17	94.59±3.55	95.74±2.70	98.56±0.77	95.30±2.75
✓	✓	✓	✓	✓		97.42±1.56	99.84±0.10	97.65±1.23	93.70±4.99	95.15±3.89	98.48±0.94	95.29±2.90
✓	✓	✓	✓		✓	97.63±0.84	99.77±0.16	97.84±0.57	94.66±2.83	96.01±1.88	98.48±0.53	95.67±1.56
✓	✓	✓	✓	✓	✓	**97.91±0.93**	**99.89±0.11**	**98.01±0.85**	**96.09±2.18**	**96.93±1.55**	**98.80±0.49**	**96.22±1.67**



Fig. 12. Style shift phenomenon in iChallenge dataset.

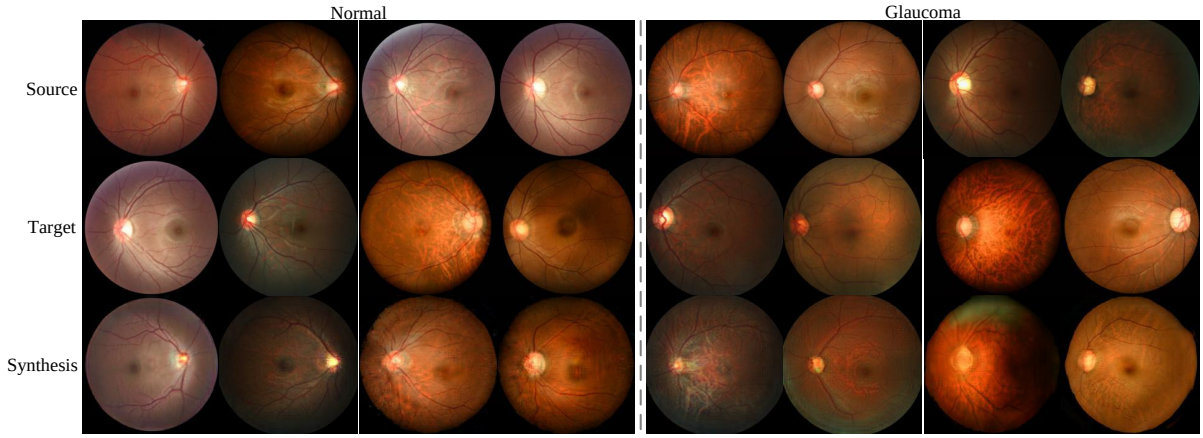


Fig. 13. Image synthesis results of iChallenge

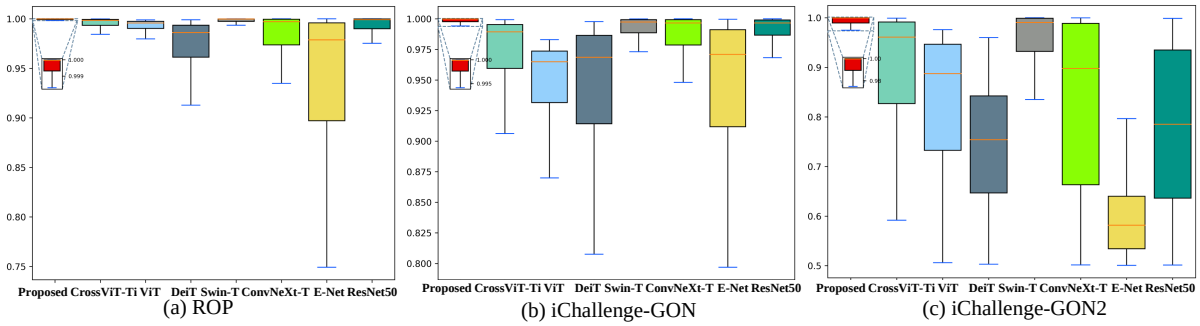


Fig. 14. Box plots of predicted confidence score statistics for each comparative method in making decisions on corresponding categories.

method improves Acc by 0.66%, Sen by 6.67%, F1 by 5.02%, and Kappa by 5.3% over DeiT. We further investigate the style shift between iChallenge-GON and iChallenge-GON2, with results in Table IX. The findings reveal a noticeable disparity between the results of the target and source style datasets.

We analyze the disparities in style shift results from the

Authorized licensed use limited to: Tsinghua University. Downloaded on November 26,2025 at 06:38:37 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3573318

TABLE VIII
EXPERIMENTAL RESULTS ON THE ICHALLENGE GON AND GON2 DATASETS.

Dataset	Method	ACC	AUC	Pre	Sen	F1	Spec	Kappa
GON2	ResNet50	94.00±6.63	96.40±6.31	98.00±6.00	90.00±10.00	93.56±6.92	98.00±6.00	88.0±13.27
	E-Net	87.00±11.00	88.40±8.66	92.67±11.72	82.00±16.61	85.65±13.01	92.00±13.27	74.00±22.00
	ConvNeXt-T	86.00±9.17	92.00±9.80	90.95±11.70	84.00±21.54	84.45±12.30	88.00±16.0	72.00±18.33
	ViT	82.00±7.48	96.00±5.37	97.50±7.50	66.00±12.81	77.94±10.36	98.00±6.00	64.00±14.97
	CrossViT-Ti	93.00±6.40	96.80±4.66	91.33±8.72	96.00±8.00	93.25±6.31	90.00±10.00	86.00±12.81
	DeiT	93.00±7.81	96.40±7.26	92.14±10.16	96.00±12.00	93.11±8.22	90.00±13.42	86.00±15.62
	Swin-T	**98.00±6.00**	**99.60±1.20**	97.14±8.57	**100.00±0.00**	**98.33±5.00**	96.00±12.00	**96.00±12.00**
	Proposed	**98.00±6.00**	98.00±6.00	**100.00±0.00**	96.00±12.00	97.50±7.50	**100.00±0.00**	**96.00±12.00**
GON	ResNet50	95.58±1.75	93.88±5.54	93.76±8.43	60.83±17.89	71.85±13.76	99.44±0.74	69.71±14.36
	E-Net	92.83±2.21	86.88±11.11	74.64±38.74	31.67±23.21	41.63±28.77	99.63±0.74	39.72±27.83
	ConvNeXt-T	95.42±1.50	92.77±5.22	90.19±10.47	62.50±15.92	72.07±11.78	99.07±1.01	69.77±12.30
	ViT	93.83±1.45	93.63±3.78	**94.86±8.15**	41.67±16.24	55.34±17.37	**99.63±0.61**	52.88±17.11
	CrossViT-Ti	95.50±1.87	93.26±5.38	85.71±11.41	67.50±16.44	74.20±12.00	98.61±1.39	71.84±12.88
	DeiT	96.42±1.71	94.92±3.16	92.93±9.42	70.00±15.46	78.85±11.27	99.35±0.93	76.99±12.09
	Swin-T	96.25±1.64	95.30±3.34	91.50±9.19	69.17±12.94	78.17±10.30	99.26±0.81	76.19±11.11
	Proposed	**97.08±1.46**	**95.73±4.14**	93.49±9.07	**76.67±9.72**	**83.87±8.14**	99.35±0.93	**82.29±8.91**

TABLE IX
EXPERIMENTAL RESULTS OF MODEL TRANSFER BETWEEN ICHALLENGE GON AND GON2.

Test dataset	Method	ACC	AUC	Pre	Sen	F1	Spec	Kappa
GON2	ResNet50	63.00	83.88	**100.00**	26.00	41.27	**100.00**	26.00
	E-Net	55.00	89.40	**100.00**	10.00	18.18	**100.00**	10.00
	ConvNeXt-T	66.00	88.12	94.44	34.00	50.00	98.00	32.00
	ViT	60.00	94.44	**100.00**	20.00	33.33	**100.00**	20.00
	CrossViT-Ti	60.00	88.64	**100.00**	20.00	33.33	**100.00**	20.00
	DeiT	70.00	92.32	91.67	44.00	59.46	96.00	40.00
	Swin-T	68.00	**94.72**	**100.00**	36.00	52.94	**100.00**	36.00
	Proposed	**77.00**	90.92	90.91	**60.00**	**72.29**	94.00	**54.00**
GON	ResNet50	81.33	81.97	30.88	**70.00**	**42.86**	82.59	33.65
	E-Net	78.83	75.31	25.19	56.67	34.87	81.30	24.40
	ConvNeXt-T	84.17	76.84	31.77	50.83	39.10	87.87	30.56
	ViT	88.33	77.71	26.19	9.17	13.58	**97.13**	8.85
	CrossViT-Ti	78.50	80.98	26.69	65.83	37.98	79.91	27.69
	DeiT	84.75	73.50	24.80	25.83	25.31	91.30	16.82
	Swin-T	82.75	74.96	26.98	42.50	33.01	87.22	23.67
	Proposed	**89.83**	**84.44**	**48.81**	34.17	40.20	96.02	**34.83**

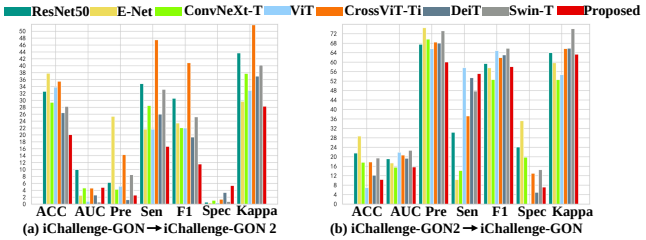


Fig. 15. Disparity analysis between iChallenge GON and GON2.

source dataset to the target dataset,depicted in Fig.15. When applying model parameters trained on iChallenge-GON to iChallenge-GON2, classification metrics such as ACC, Sen, F1, and Kappa show performance difference of over 10%, with the maximum reaching nearly 52%. In contrast, Spec and AUC difference are smaller, under 10%. For all seven comparision methods, ACC difference exceeded 20%. Our method demonstrates improvements across multiple metrics: ACC increased by 7%, Sen by 16%, F1 by 12.83%, and Kappa by 14%. When applying model parameters from iChallenge-GON2 to iChallenge-GON,there is a significant decrease in Pre, F1, and Kappa, with reductions over 50%, and sometimes exceeding 70%. Metrics like ACC, AUC, and Spec showed smaller decreases, mostly above 4%. Our method shows improvements: ACC increased by 1.5%, AUC by 2.47%, Pre by 17.04%, and Kappa by 1.18%. Fig.12 illustrates style difference between datasets, and style shift issues lead to classification performance declines, as validated by metric difference in Fig.15. These results confirm the effectiveness of our method on the two publicly available datasets.

F. Statistical analysis of results

This study conducts statistical analysis on the experimental results. For the ROP classification task, the model's last layer outputs predicted probabilities P_i (i=1, 2, $\cdots$,3) for each input sample. The predicted category j with the highest probability is selected, with the confidence score denoted as $P_j(j \leq 3)$. In each cross-validation experiment, we analyze and plot the predicted confidence scores P_j for all test samples using box plots, as shown in Fig.14. The height of the box plot, from the first to third quartiles, reflects data fluctuation. A lower height indicates concentrated decision confidence scores. The proposed method's confidence scores fluctuate less and are concentrated at higher values. The orange line represents the median confidence score, which is higher for our method. The blue lines show the upper and lower limits of confidence scores, with our method having a lower limit. Overall, our method outperforms the seven comparison methods in the ROP classification task. Similarly, in the iChallenge-GON dataset, our method shows less fluctuation, higher concentration of predicted confidence, larger median values, and lower limits of decision confidence.

TABLE X
SIGNIFICANT DIFFERENCE T-TEST ANALYSIS BETWEEN THE MODEL PROPOSED IN THIS STUDY AND SEVEN COMPARATIVE METHODS.

Dataset	Method	ResNet50	E-Net	ConvNeXt-T	ViT	CrossViT-Ti	DeiT	Swin-T
Ours	p-value	<0.001	<0.001	<0.0001	<0.0001	<0.0001	<0.001	<0.0001
iChallenge-GON	p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
iChallenge-GON2	p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.01

Finally, we conduct a t-test analysis on the predicted confidence of our proposed model and seven other comparison

Authorized licensed use limited to: Tsinghua University. Downloaded on November 26,2025 at 06:38:37 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3573318

methods to determine if there is a statistically significant difference in classification performance. Here, the null hypothesis is given as follows: there is no statistically significant difference in classification performance between each comparison method and the proposed model. Let X be the set of confidence scores P_j for each test sample predicted as class j by our model, and X_i ($i \leq 7$) be the set for each comparison method. An independent samples t-test is performed on X and X_i. The p-value, calculated based on the distribution of the test statistics, indicates the probability of obtaining a result as extreme as the observed one, assuming the null hypothesis is true. A significance level of 0.05 is generally used, and the null hypothesis is rejected if the p-value is less than or equal to 0.05. The t-test results, shown in Table X, reveal that in the ROP, iChallenge-GON, and iChallenge-GON2 datasets, most p-values are below 0.05. This suggests a very low probability for the null hypothesis, providing sufficient reason to reject it. Therefore, the performance difference between our model and the seven comparison methods are statistically significant.

V. Conclusion

This paper introduces a feature alignment and adversarial discrimination approach based on DS-Swin-T for ROP classification. Two independent Swin-T branches extract fine-grained and coarse-grained features, which are effectively fused using the FFU. To mitigate classification performance degradation due to style transfer, an IS module generates synthesis style images as an intermediate latent space between source and target style images, aligning detail and style features in the shallow low-level feature space. Adversarial learning is then applied in the advanced feature space to enable the network to learn style-invariant features. The proposed method's effectiveness is validated through extensive experiments on a private ROP dataset and two public datasets. While our method reduces performance degradation caused by style shifts compared to other approaches and improves various evaluation metrics, there is still room for further improvement in cross-style testing performance. This work has some limitations. First, for one source style and three target styles, three style transfer generators need to be trained to produce synthesis style data. As the number of target styles increases, the required training cost also rises. Second, the automated ROP diagnosis in this study only involves a three-level classification task of normal, A-ROP, and ROP, while in the real world, clinical features related to ROP are more complex. Therefore, future work will expand in the following two directions: (1) explore a new generative model that, given a specified source style image, can randomly generate synthesis style images of any existing target style through a control factor; (2) build upon the existing dataset to construct a larger ROP cohort dataset that encompasses more ROP-related clinical features, such as ROP plus disease, ROP staging and zoning, etc. This will facilitate the development of a more comprehensive automated ROP diagnostic model to meet broader clinical needs to the extent possible.

References

[1] J.-A. Quinn, F. M. Munoz *et al.*, "Preterm birth: Case definition & guidelines for data collection, analysis, and presentation of immunisation safety data," *Vaccine*, vol. 34, no. 49, pp. 6047–6056, 2016.

[2] A. Hellström, L. E. Smith, and O. Dammann, "Retinopathy of prematurity," *The lancet*, vol. 382, no. 9902, pp. 1445–1457, 2013.

[3] Y. Zhang, L. Wang, Z. Wu *et al.*, "Development of an automated screening system for retinopathy of prematurity using a deep neural network for wide-angle retinal images," *Ieee Access*, vol. 7, pp. 10 232–10 241, 2018.

[4] J. Wang, R. Ju *et al.*, "Automated retinopathy of prematurity screening using deep neural networks," *EBioMedicine*, vol. 35, pp. 361–368, 2018.

[5] J. Hu, Y. Chen *et al.*, "Automated analysis for retinopathy of prematurity by deep neural networks," *IEEE Transactions on Medical Imaging*, vol. 38, no. 1, pp. 269–279, 2019.

[6] M. F. Chiang, G. E. Quinn, A. R. Fielder *et al.*, "International classification of retinopathy of prematurity, third edition," *Ophthalmology*, vol. 128, no. 10, pp. e51–e68, 2021.

[7] H. Xie, B. Lei *et al.*, "Adversarial learning-based multi-level dense-transmission knowledge distillation for ap-rop detection," *Medical Image Analysis*, vol. 84, p. 102725, 2023.

[8] R. Zhang, B. Lei *et al.*, "Automatic diagnosis for aggressive posterior retinopathy of prematurity via deep attentive convolutional neural network," *Expert Systems with Applications*, vol. 187, p. 115843, 2022.

[9] Y. Zhang, S. Miao *et al.*, "Task driven generative modeling for unsupervised domain adaptation: Application to x-ray image segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Springer International Publishing, 2018, pp. 599–607.

[10] P. Costa, A. Galdran, M. I. Meyer *et al.*, "End-to-end adversarial retinal image synthesis," *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 781–791, 2017.

[11] F. Qin, N. Gao *et al.*, "Fine-grained leukocyte classification with deep residual learning for microscopic images," *Computer Methods and Programs in Biomedicine*, vol. 162, pp. 243–252, 2018.

[12] X. Wu, Y.-T. Zhang *et al.*, "A novel centralized federated deep fuzzy neural network with multi-objectives neural architecture search for epistatic detection," *IEEE Transactions on Fuzzy Systems*, pp. 1–13, 2024.

[13] X. Wu, Y. Wei *et al.*, "A micro-aggregation algorithm based on density partition method for anonymizing biomedical data," *Current Bioinformatics*, vol. 14, no. 7, pp. 667–675, 2019.

[14] M. Thakur, H. Kuresan *et al.*, "Soft attention based densenet model for parkinson's disease classification using spect images," *Frontiers in Aging Neuroscience*, vol. 14, p. 908143, 2022.

[15] K. Han, Y. Wang, H. Chen *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[16] S. Chen, Z. Wu, B. Lei *et al.*, "Fit-net: Feature interaction transformer network for pathologic myopia diagnosis," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2023.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[18] E. Xie, W. Wang, Z. Yu *et al.*, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[19] Z. Liu, Y. Lin, Y. Cao *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[20] C.-F. R. Chen, Q. Fan *et al.*, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 357–366.

[21] A. Lin, B. Chen, J. Xu *et al.*, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2022.

[22] J.-Y. Zhu, T. Park *et al.*, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[23] P. Russo, F. M. Carlucci, T. Tommasi, and B. Caputo, "From source to target and back: symmetric bi-directional adaptive gan," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8099–8108.

Authorized licensed use limited to: Tsinghua University. Downloaded on November 26,2025 at 06:38:37 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3573318

[24] Y. Ganin, E. Ustinova, H. Ajakan *et al.*, "Domain-adversarial training of neural networks," *JOURNAL OF MACHINE LEARNING RESEARCH*, vol. 17, no. 1, pp. 2096–2030, 2016.

[25] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[26] T. M. Hundscheid, S. Gulden, and OTHERS, "Sex differences in the risk of retinopathy of prematurity: a systematic review, frequentist and bayesian meta-analysis, and meta-regression," *World Journal of Pediatrics*, vol. 20, no. 4, pp. 340–352, 2024.

[27] A. Pivodic, A.-L. Hård *et al.*, "Individual Risk Prediction for Sight-Threatening Retinopathy of Prematurity Using Birth Characteristics," *JAMA Ophthalmology*, vol. 138, no. 1, pp. 21–29, 01 2020.

[28] S. Chen, X. Zhao *et al.*, "Multi-risk factors joint prediction model for risk prediction of retinopathy of prematurity," *EPMA Journal*, vol. 15, no. 2, pp. 261–274, Jun 2024.

[29] Q. Wu, Y. Hu *et al.*, "Development and Validation of a Deep Learning Model to Predict the Occurrence and Severity of Retinopathy of Prematurity," *JAMA Network Open*, vol. 5, no. 6, pp. e2 217 447–e2 217 447, 06 2022.

[30] S. M. Sharafi, N. Ebrahimiadib *et al.*, "Automated diagnosis of plus disease in retinopathy of prematurity using quantification of vessels characteristics," *Scientific Reports*, vol. 14, no. 1, p. 6375, March 16 2024.

[31] A. S. Coyner, T. Murickan *et al.*, "Multinational External Validation of Autonomous Retinopathy of Prematurity Screening," *JAMA Ophthalmology*, vol. 142, no. 4, pp. 327–335, 04 2024.

[32] T.-T. P. Nguyen, B. K. Young *et al.*, "Discrepancies in diagnosis of treatment-requiring retinopathy of prematurity," *Ophthalmology Retina*, vol. 8, no. 1, pp. 88–91, 2024.

[33] J. M. Brown, J. P. Campbell *et al.*, "Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks," *JAMA Ophthalmology*, vol. 136, no. 7, pp. 803–810, 07 2018.

[34] Y. Peng, W. Zhu *et al.*, "Automatic staging for retinopathy of prematurity with deep feature fusion and ordinal classification strategy," *IEEE Transactions on Medical Imaging*, vol. 40, no. 7, pp. 1750–1762, 2021.

[35] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2014.

[36] S. Ramchandran, G. Tikhonov *et al.*, "Learning conditional variational autoencoders with missing covariates," *Pattern Recognition*, vol. 147, p. 110113, 2024.

[37] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza *et al.*, "Generative adversarial nets," in *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 27 (NIPS 2014)*, ser. Advances in Neural Information Processing Systems, vol. 27, 2014, pp. 2672–2680.

[38] T. Chakraborty, U. R. K. S *et al.*, "Ten years of generative adversarial nets (gans): a survey of the state-of-the-art," *Machine Learning: Science and Technology*, vol. 5, no. 1, p. 011001, jan 2024.

[39] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851.

[40] T. Zhou, H. Fu, G. Chen *et al.*, "Hi-net: hybrid-fusion network for multi-modal mr image synthesis," *IEEE transactions on medical imaging*, vol. 39, no. 9, pp. 2772–2781, 2020.

[41] S. U. Dar, M. Yurt, L. Karacan *et al.*, "Image synthesis in multi-contrast mri with conditional generative adversarial networks," *IEEE transactions on medical imaging*, vol. 38, no. 10, pp. 2375–2388, 2019.

[42] D. Nie, R. Trullo, J. Lian *et al.*, "Medical image synthesis with context-aware generative adversarial networks," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*. Springer International Publishing, 2017, pp. 417–425.

[43] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern Recognition*, 2018, pp. 9242–9251.

[44] K. Wang, Z. Chen *et al.*, "Score-based counterfactual generation for interpretable medical image classification and lesion localization," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2024.

[45] X. Yi, E. Walia *et al.*, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, vol. 58, p. 101552, 2019.

[46] H. Li, H. Liu *et al.*, "A generic fundus image enhancement network boosted by frequency self-supervised representation learning," *Medical Image Analysis*, vol. 90, p. 102945, 2023.

[47] T. Zhang, H. Fu, Y. Zhao *et al.*, "Skrgan: Sketching-rendering unconditional generative adversarial networks for medical image synthesis," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Springer International Publishing, 2019, pp. 777–785.

[48] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[49] S. P. H. Boroujeni and A. Razi, "Ic-gan: An improved conditional generative adversarial network for rgb-to-ir image translation with applications to forest fire monitoring," *Expert Systems with Applications*, vol. 238, p. 121962, 2024.

[50] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[51] M. Daniels, T. Maunu, and P. Hand, "Score-based generative neural networks for large-scale optimal transport," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 12 955–12 965.

[52] P. Isola, J.-Y. Zhu *et al.*, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[53] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[54] M. Ghifary, W. B. Kleijn, M. Zhang *et al.*, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *Computer Vision – ECCV 2016*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2016, pp. 597–613.

[55] S. Motiian, M. Piccirilli *et al.*, "Unified deep supervised domain adaptation and generalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5715–5725.

[56] H. Lei, W. Liu, B. Lei *et al.*, "Unsupervised domain adaptation based image synthesis and feature alignment for joint optic disc and cup segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 1, pp. 90–102, 2021.

[57] Y.-H. Tsai, W.-C. Hung *et al.*, "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7472–7481.

[58] Y. Chen, W. Li, X. Chen, and L. V. Gool, "Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1841–1850.

[59] W. Zhang, W. Ouyang *et al.*, "Collaborative and adversarial network for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3801–3809.

[60] J. I. Orlando, H. Fu *et al.*, "Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Medical image analysis*, vol. 59, p. 101570, 2020.

[61] H. Fang, H. Fu *et al.*, "Multi-modality images analysis: A baseline for glaucoma grading via deep learning," in *Ophthalmic Medical Image Analysis*. Springer International Publishing, 2021, pp. 139–147.

[62] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[64] Z. Liu, H. Mao, C.-Y. Wu *et al.*, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[65] H. Touvron, M. Cord, M. Douze *et al.*, "Training data-efficient image transformers & distillation through attention," in *INTERNATIONAL CONFERENCE ON MACHINE LEARNING, VOL 139*, ser. Proceedings of Machine Learning Research, vol. 139, 2021, pp. 7358–7367.

Authorized licensed use limited to: Tsinghua University. Downloaded on November 26,2025 at 06:38:37 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.