TMM-Nets: Transferred Multi- to Mono-Modal Generation for Lupus Retinopathy Diagnosis

Ruhan Liu, Tianqin Wang, Huating Li, Ping Zhang, *Senior Member, IEEE*, Jing Li, Xiaokang Yang, *Fellow, IEEE*, D. Shen, *Fellow, IEEE*, and Bin Sheng, *Member, IEEE*

***Abstract*— Rare diseases, which are severely underrepresented in basic and clinical research, can particularly benefit from machine learning techniques. However, current learning-based approaches usually focus on either mono-modal image data or matched multi-modal data, whereas the diagnosis of rare diseases necessitates the aggregation of unstructured and unmatched multi-modal image data due to their rare and diverse nature. In this study, we therefore propose diagnosis-guided multi-to-mono modal generation networks (TMM-Nets) along with training and testing procedures. TMM-Nets can transfer data from multiple sources to a single modality for diagnostic data structurization. To demonstrate their potential in the context of rare diseases, TMM-Nets were deployed to diagnose the lupus retinopathy (LR-SLE), leveraging unmatched regular and ultra-wide-field fundus images for transfer learning. The TMM-Nets encoded the transfer learning from diabetic retinopathy to LR-SLE based on the similarity of the fundus lesions. In addition, a lesion-aware multi-scale attention mechanism was developed for clinical alerts, enabling TMM-Nets not only to inform patient care, but also to provide insights consistent with those of clinicians. An adversarial strategy was also developed to refine multi- to mono-modal image generation based on diagnostic results and the data distribution to enhance the data augmentation performance. Compared to the baseline model, the TMM-Nets showed 35.19% and 33.56% F1 score improvements on the test and external validation sets, respectively. In addition, the TMM-Nets can be used to develop diagnostic models for other rare diseases.**

***Index Terms*— Lupus retinopathy, generating adversarial training, UWF-FFA, UWF-FP, unmatched multi-modal data.**

I. Introduction

SYSTEMIC lupus erythematosus (SLE) is a potentially serious, chronic autoimmune disorder that particularly affects women of childbearing age [1]. Lupus retinopathy is among the most common vision-threatening complications of systemic lupus erythematosus. In addition, life-table survival estimates have shown decreased survival in patients with LR-SLE compared to those without retinopathy [2]. Diagnosis of LR-SLE, therefore, is of critical importance, both visually and prognostically. In real-world clinical settings, LR-SLE is often diagnosed by the standard examination measurements of UWF-FP and UWF-FFA (examples are shown in Fig. 1), because UWF-FP and UWF-FFA can capture fields of view over 200°, encompassing more than 80% of the retina [3]. Therefore, these approaches can easily visualize the retinal periphery, and retina lesions can be readily detected in an ultra-wide-field image. Although LR-SLE is often empirically clinically detected by identifying vascular occlusions, exudates, and microangiomas that develop at the ends of arteries, the retina phenotype descriptors of LR-SLE remain largely uncertain and less interpretable due to complex of causes, relying heavily on clinician experience.

In the clinical realm, several common barriers remain in the diagnosis of LR-SLE and other rare diseases: 1) a lack of paired and labeled clinic data for rare diseases (the annual incidence of SLE in the United States is 2-7.6 per 100,000 and the prevalence is 19-159 per 100,000, whereas the incidence of LR-SLE is only approximately 10% of that of SLE [1]); 2) insufficient knowledge about rare disease and their identification/diagnosis; 3) different clinical examinations recording different modal data for few and geographically dispersed patients. In clinical practice, many patients with LR-SLE are examined via either UWF-FP or UWF-FFA, but not both. Consequently, the data collected are typically multi-modal but unmatched and unstructured. Although recent advances

Manuscript received 18 August 2022; revised 7 November 2022; accepted 13 November 2022. Date of publication 21 November 2022; date of current version 3 April 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62272298 and Grant 82100879 and in part by the Shanghai Pujiang Program under Grant 2020PJD044. *(Corresponding authors: Bin Sheng; Jing Li.)*

Ruhan Liu and Bin Sheng are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: liurh996@sjtu.edu.cn; shengbin@sjtu.edu.cn).

Tianqin Wang and Jing Li are with the Department of Ophthalmology, Renji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200127, China (e-mail: wang_tianqin@163.com; lijinmpa@163.com).

Huating Li is with the Shanghai Clinical Center for Diabetes, Shanghai Diabetes Institute, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai 200233, China (e-mail: huarting99@sjtu.edu.cn).

Ping Zhang is with the Department of Computer Science and Engineering, and the Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210 USA (e-mail: zhang.10631@osu.edu).

Xiaokang Yang is with the MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: xiaokang@sjtu.edu.cn).

D. Shen is with the School of Biomedical Engineering, ShanghaiTech University, Shanghai 201210, China, also with Shanghai United Imaging Intelligence Co., Ltd., Shanghai 200230, China, and also with the Shanghai Clinical Research and Trial Center, Shanghai, 201210, China (e-mail: Dinggang.Shen@gmail.com).

Data is available on-line at https://github.com/Liuruhan/TMM-Nets

Digital Object Identifier 10.1109/TMI.2022.3223683

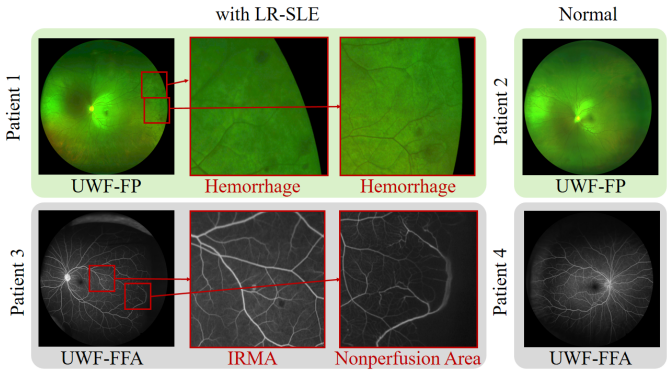

Fig. 1. Examples of ultra-wide-field fundus photography (UWF-FP) images and ultra-wide-field fundus fluorescence angiography (UWF-FFA) images of four patients. Each patient case is enclosed within a rectangular box. Patients with systemic lupus erythematosus retinopathy (LR-SLE) are shown, with the details of their lesions inside the red rectangular boxes. The corresponding lesion types are also labeled below the images. Here, IRMA indicates intraretinal microvascular abnormality.

in medical imaging, particularly in disease diagnosis, have been facilitated by deep learning [4], [5], [6], [7], [8], to the best of our knowledge, very few deep learning approaches have been designed to model unstructured and unmatched multi-modal data that are commonly used as inputs for rare disease diagnosis.

This paper presents learning guided multi- to mono-modal generation networks (TMM-Nets) for LR-SLE diagnosis, which employ unmatched multi-modal data. Two parts of related work are guided our model development: transfer learning, and generative adversarial networks (GANs) provide the basis for this approach [7], [9], [10], [11], [12]. First, we hypothesized that LR-SLE shares some similarity with common retina diseases, such as diabetic retinopathy (DR), in terms of imaging feature extraction, because both LR-SLE and DR are retinopathies with some common lesions (e.g., exudation and hemorrhage). Furthermore, the use of easily accessible DR images to help diagnose LR-SLE is more clinically relevant than distinguishing between LR-SLE and DR for two reasons. On the one hand, there is a low overlap between LR-SLE and DR populations. On the other hand, some fundus lesions, such as hemorrhages and exudates, present similarly in both LR-SLE and DR. Second, the TMM-Nets training framework introduces a structural Cycle- GAN-based UWF FFA-to-FP translator which can make the use of structure information and complete unpair image-to-image generation.

In addition, inspired by the existing transfer-learning models for lesion detection [13], [14], [15], [16] and fundus disease diagnosis [4], [5], [6], we developed a new knowledge transfer model for LR-SLE diagnosis with the assistance of the readily available and well investigated DR data. In addition, by taking advantage of the unmatched UWF-FFA and UWF-FP data, the TMM-Nets achieved remarkable diagnostic performance for LR-SLE by overcoming the challenge of insufficient training data and including the ability to identify characteristics readily that are not typically recognized by human experts. This paper makes several technical contributions towards the goal of developing neural networks to support rare disease diagnosis:

1) We propose a knowledge transfer module learning from DR to LR-SLE that diversifies the distribution of lesion features and a UWF FFA-to-FP translator pre-trained by a structural Cycle-GAN that removes input data constraints for mono-modal data training. These modules broaden the data distribution and improve the model generalization.
2) We develop a multi-scale attention fusion structure that enables the TMM-Nets to capture data at several scales, and a lesion area reminder that directs the network's attention to lesion regions. These modules help the networks extract fine features and perform diagnoses that agree with physician labels.
3) We utilize the UWF FFA-to-FP translator as a generator, and the diagnosis network and fake UWF-FP discriminator are employed as two discriminators to guarantee that the generated images are inside the UWF-FP distribution and that the diagnostic results are consistent with the original UWF-FFA.

II. Related Work

Previous studies, including those on knowledge transfer in fundus disease, image generation techniques, and attention mechanisms for feature enhancement, have contributed to the development of the LR-SLE diagnosis framework with unpaired UWF-FP and UWF-FFA inputs.

A. Automatic Diagnosis in Fundus Disease

After Yosinski et al. [17] explored the transferability of deep neural networks, a large body of work started to be performed using deep transfer learning [18], [19], [20], [21], [22]. Tamaazousti et al. [21] proposed methods relying on human classification knowledge and retraining using fine-tuning to evaluate the generality in transfer learning schemes and explored the generality expressions used for transfer learning. As a representative deep transfer learning, Long et al. [19] proposed a deep adaptation network (DAN) that made full use of the "transfer" property of deep networks and then introduced the multi-kernel maximum mean discrepancy distance (MK-MMD) in statistical learning to achieve good results. Additionally, Li et al. [22] proposed an efficient implementation of a deep residual correction network that effectively enhances source-to-target adaptation by inserting a residual block into the source network along with a task-specific feature layer and explicitly weakens the influence of irrelevant sources.

Furthermore, multi-task learning is also a strategy for fundus disease automatic diagnosis. Many works use multiple fundus lesion labels to enable different disease determinations on fundus images [23], [24]. Li et al. [23] proposed a cross-disease attention network (CANet) to jointly grade DR and DME by exploring the internal relationship between the diseases with only image-level supervision. Moreover, in [14], the authors introduced a cross-attention multi-branch network to diagnose three fundus diseases (Coats, retinitis pigmentosa, DR). In addition, transfer learning is frequently used to diagnose fundus diseases. Transfer learning has achieved outstanding results in DR diagnosis, which is currently widely studied [4].

Furthermore, as described in [13], different fundus diseases may have similar fundus lesions, so a transfer learning method is proposed to improve the multi-disease identification performance using its dataset.

Due to the rarity of LR-SLE and its specific population (SLE patients), accumulating fundus data is difficult for LR-SLE. Thus, using easily accessible DR images to aid in diagnosing LR-SLE is more clinically relevant than differentiating LR-SLE from DR. Moreover, because some fundus lesions, such as hemorrhages and exudates, have similar appearances in both LR-SLE and DR, the transfer learning model was used as the basis for knowledge mapping from DR to LR-SLE in our work.

B. Image Generation Techniques

Due to the difficulty of obtaining medical data, image generation techniques are also used extensively to improve model performance. GANs [25] are neural network models in which two networks are trained concurrently, one for image generation and one for discrimination. Currently, methods based on GANs are making significant progress in various fields related to image generation and transformation [26]. The utility of adversarial domain transfer and the efficacy of generating new image samples also make this technique promising for a wide variety of medical imaging applications [27]. In addition, prior to the introduction of Cycle-GAN [12], using one style of image to generate another style of image was extremely resource intensive, such as in the case of pixel2pixel [28], which required pairs of data for training. In order to complete our task, we required UWF-FP and UWF-FFA data collected at the same location, which is an impossible task in data collection. Cycle-GAN enables unpaired data training and lays the groundwork for generating UWF-FP data with similar fundus features from corresponding UWF-FFA data.

Moreover, in medical image generation, a very large amount of work has been conducted in recent years on GAN-based image generation, including 3D image generation and the production of super-resolution medical images [29], [30], [31]. In [30], the authors proposed a novel GAN-based unsupervised deformable image registration method. The work was based on Cycle-GAN for the translation of computed tomography images into magnetic resonance images. This work provided the basis for the design of our UWF-FP-to-UWF-FFA image transformation network.

C. Attention Mechanism for Feature Enhancement

Attention mechanisms have been proposed to improve model performance by enhancing important features and suppressing irrelevant ones by focusing on important regions [10], [32]. Furthermore, our previous work demonstrated that attention models can produce excellent results and exceptional stability for medical image tasks involving specific problems [33]. Xie et al. proposed a cross-attention multi-branch network for diagnosing three fundus diseases, employing ultra-wide-field scanning laser ophthalmoscopy images [14]. Ouyang et al. presented a dual-sampling attention network, a diagnostic network using a dual-sampling strategy combined with an attention mechanism, to classify COVID-19 and community-acquired pneumonia infections [7]. In addition, He at al. [8] developed a new category attention block, which explores more distinguishing regional features for each DR class and treats each category equally. All of these scholars have had the objective of designing a model more focused on the regions of interest to physicians while improving the classification performance of the model. Following the attention-enhanced feature map can generate a heat map representing the importance of each original input pixel.

The most common gradient-based heat map generation methods are class activation mapping (CAM) [9] and gradient-weighted class activation mapping (Grad-CAM) [11]. The heat maps they generate can reflect regions of network concern, thus enhancing the model interpretability, which could help elucidate diagnosis mechanisms. Some recent studies have also explored online CAM or Grad-CAM training to help models locate regions of interest more effectively [34], [35].

III. Method

Our proposed TMM-Net architecture is divided into two components: a structural Cycle-GAN-based UWF FFA-to-FP translator and an LR-SLE DN. The two components are firstly trained for image generation and disease diagnosis, and then the translator and DNs are refined for diagnosis-guide generation. The framework is depicted in Fig. 2.

A. Structural Cycle-GAN-Based UWF FFA-to-FP Translator

We trained a structural Cycle-GAN model [12] on our unpaired UWF-FFA and UWF-FP datasets to translate UWF-FFA images into UWF-FP images based on similar features, because Cycle-GAN performs exceptionally well in image generation and is widely used in medical image translation [36], [37], [38]. The structural Cycle-GAN model consists of two generators and two discriminators that are trained iteratively through an adversarial process to achieve mapping between unpaired data. The generation functions G_{fa-fp} and G_{fp-fa} attempt to capture the underlying data density to achieve UWF FFA-to-FP and UWF FP-to-FFA translation, respectively, and confuse the two discriminant functions D_{fa} and D_{fp} used to distinguish between fake UWF-FFA (f_{fa}) and real UWF-FFA (r_{fa}), as well as fake UWF-FP (f_{fp}) and real UWF-FP (r_{fp}). The optimization procedures D_{fa} and D_{fp} aim to achieve distinguishability and distinguish the real UWF-FFA image and the real UWF-FP images from the generated UWF-FFA and UWF-FP images, respectively. The risk function for optimizing this two-maximum-minimum, two-player game can be written as

$$\mathcal{V} = arg \min_{G_{fp-fa}, G_{fa-fp}} \max_{D_{fp}, D_{fa}} \mathcal{L}_{total}(G_{fp-fa}, G_{fa-fp}, D_{fp}, D_{fa}) \quad (1)$$

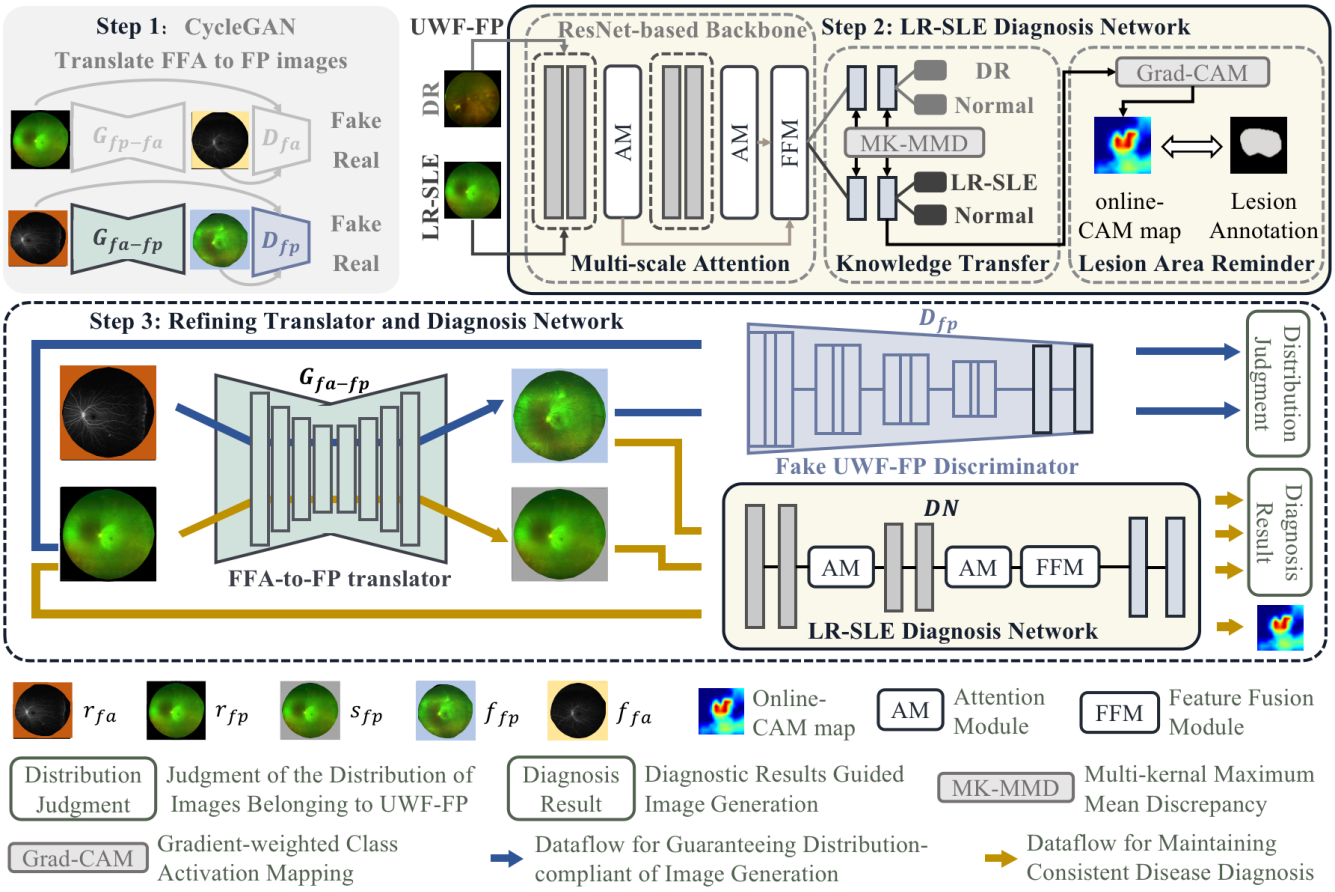

Fig. 2. Overview figure of the guided multi- to mono-modal generation and diagnosis networks (TMM-Nets) and their training strategy. In step 1, the UWF FFA-to-FP translator is trained through a structural Cycle-GAN-based strategy. In step 2, we train the LR-SLE diagnosis network (DN) is trained by encoding transfer learning-based diagnosis module training with the UWF-FP images in DR patients with lesion annotations. In step 3, the UWF FFA-to-FP translator as a generator is trained through an adversarial strategy with two discriminators to maintain both image realism and lesion similarity. Here, r_{fa} and r_{fp} are the real UWF-FFA and UWF-FP images, respectively; f_{fa} and f_{fp} are the fake UWF-FFA and UWF-FP images generated by generators, respectively; and s_{fp} represents the UWF-FP images obtained by feeding UWF-FP image into the UWF FFA-to-FP translator. The online-CAM map is the heat map of UWF-FP images with LR-SLE generated by the DN through online CAM.

where G_{fp-fa} and G_{fa-fp} aim to minimize this objective against adversaries D_{fp} and D_{fa} that attempt to maximize it, respectively. The total loss $\mathcal{L}_{total}$ in our structural Cycle-GAN based model is

$$\mathcal{L}_{total} = \mathcal{L}_{GAN} + \alpha \cdot \mathcal{L}_{cyc} + \beta \cdot \mathcal{L}_{idt} + \gamma \cdot \mathcal{L}_{SSIM} \quad (2)$$

where α, β, and γ are the relative weights of each term; $\mathcal{L}_{GAN}$ is the adversarial loss; $\mathcal{L}_{cyc}$ is the cycle consistency loss; $\mathcal{L}_{idt}$ is the identity loss; and $\mathcal{L}_{SSIM}$ is the SSIM loss. The detailed calculations of $\mathcal{L}_{GAN}$, $\mathcal{L}_{cyc}$, and $\mathcal{L}_{idt}$ can be found in [12], and the detailed calculation of $\mathcal{L}_{SSIM}$ can be found in [39].

B. LR-SLE Diagnosis Network

The LR-SLE diagnosis network is composed of three modules: an attention-based feature extraction module (AFEM), a knowledge transfer module based on deep adaptation networks (DKTM), and an online-CAM-based lesion area reminder module (OLRM). These modules enable the network to extract significant features at multiple scales, to transfer knowledge about diabetic fundus lesions as if it were a physician, and to diagnose lesions based on their location.

Multi-scale feature extraction is critical for the final diagnostic results because LR-SLE patients may present with lesion features such as cotton wool spots that cover large areas or hemorrhages and exudates that are more widely distributed but cover a smaller proportion of lesions. Inspired by our previous work [33], the proposed attention-based feature extractor uses a multi-scale attention structure to capture features at both high and low levels, which enables it to perform exceptionally well in diagnosing very small lesions, as depicted in Fig. 3 (AFEM).

In addition, we used an online attention algorithm to learn the critical features involved in the diagnosis process and to generate the relative attention map. Unlike Grad-CAM [11], which generates input heat maps using average gradient weights and feature maps from the final convolutional layer, our TMM-Nets create a lesion area reminder based on physician-annotated lesion areas to ensure that the heat maps generated by the LR-SLE DN are comparable to those produced by clinicians. Gradients from the final convolutional layer are combined with the final feature maps to create attention maps. The attention maps demonstrate how the network detects LR-SLE and which portions of the input images are

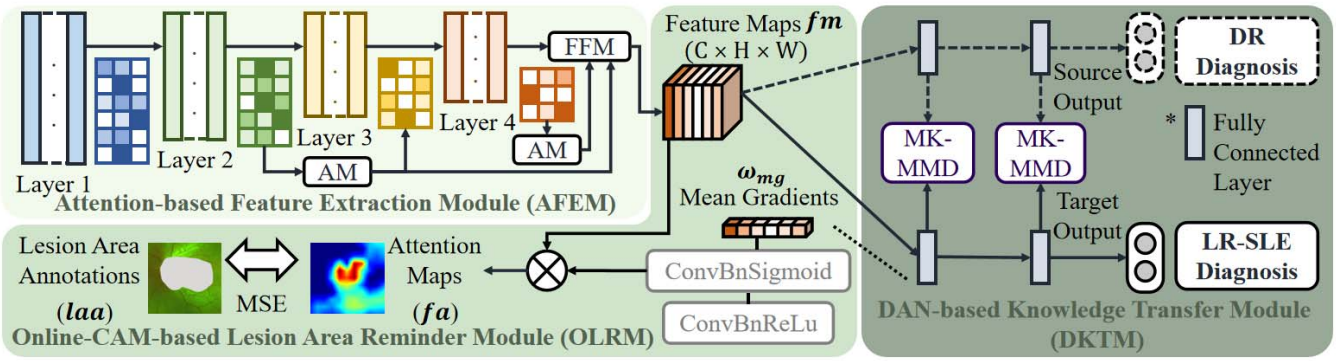

Fig. 3. Network structure of the LR-SLE diagnosis network (DN) which contains three parts: an attention-based feature extraction module (AFEM), a deep adaptation network (DAN)-based knowledge transfer module (DKTM), and an online-CAM based lesion area reminder module (OLRM). The AFEM uses the attention mechanism to capture and enhance high-level and low-level features. The DKTM learns diagnosis knowledge in DR to transfer into LR-SLE diagnosis. The OLRM uses simple lesion area annotation to make the network learn in a way similar to physician diagnosis.

significant. To ensure consistency of judgment, we calculated the mean square error (MSE) loss of each attention map and lesion annotation map from physicians. This procedure was used exclusively for patients with LR-SLE, and we did not limit the attention maps of SLE patients with normal fundus. The structure is depicted in Fig. 3 (OLRM).

Although we increased the data scale by converting multi-modal data into mono-modal data, further research is also necessary to address the lack of data for LR-SLE disease lesions. Due to some similarity between DR and LR-SLE lesions, we used a UWF-FP image dataset collected from diabetic patients with or without DR using the same type of device. We designed a knowledge transfer module using a DAN to transfer information from DR diagnosis to LR-SLE diagnosis. The structure is depicted in Fig. 3 (DKTM). The DAN-based DKTM uses MK-MMD to measure the distance between source and target distributions. In transfer learning tasks, MMD is more suitable for minimizing the difference between source and target domains due to the different but correlated sample data. MK-MMD is developed based on the original MMD, which proposes to use multiple kernels to construct this total kernel and thus enhance the effect.

C. Diagnosis-Guide Networks Refining

In the previous steps, we obtained three models: a UWF FFA-to-FP translator (G_{fa-fp}), a fake UWF-FP discriminator (D_{fp}), and an LR-SLE diagnosis network (DN). We introduced a diagnosis-guide networks refining strategy to use the UWF FFA-to-FP translator as the generator, and the fake UWF-FP discriminator and LR-SLE DN as two discriminators. The generation function G_{fa-fp} attempts to capture the underlying data density and confounds the diagnosis function D_{fp}, whereas the optimization process DN aims to achieve the correct disease discrimination for the generated images (Fig 2). The risk function for optimizing this objective is

$$\mathcal{V}(G_{fa-fp}, D_{fp}, DN) = arg \min_{G_{fa-fp}} \max_{D_{fp}, DN} \mathcal{L}_{G_{total}}(G_{fa-fp}, D_{fp}, DN) \quad (3)$$

The loss function $\mathcal{L}_{G_{total}}$ for the GAN-based training strategy can be written as

$$\mathcal{L}_{G_{total}} = \mathcal{L}_{idt} + \omega_1 \cdot \mathcal{L}_{SSIM} + \omega_2 \cdot \mathcal{L}_{G_{D_{fp}}} + \omega_3 \cdot \mathcal{L}_{G_{DN}} + \omega_4 \cdot \mathcal{L}_{OLRM} \quad (4)$$

where ω_1, ω_2, ω_3 and ω_4 are the relative weights of each term. $\mathcal{L}_{idt}$ is the identity loss [12], and $\mathcal{L}_{SSIM}$ is the structual loss [39]. The GAN loss for the fake UWF-FP discriminator is $\mathcal{L}_{G_{D_{fp}}}$, and that for LR-SLE diagnosis is $\mathcal{L}_{G_{DN}}$. The lesion area reminding loss for the OLRM is $\mathcal{L}_{OLRM}$.

The GAN loss in adversarial training of the UWF FFA-to-FP translator contains of two parts: $\mathcal{L}_{G_{D_{fp}}}$ and $\mathcal{L}_{G_{DN}}$. $\mathcal{L}_{G_{D_{fp}}}$ is the GAN loss for the fake and real images discriminator and is given by

$$\mathcal{L}_{G_{D_{fp}}} = \mathbb{E}_{r_{fa} \sim p_{data}(r_{fa})}[(1 - D_{fp}(G_{fa-fp}(r_{fa})))^2] + \mathbb{E}_{r_{fp} \sim p_{data}(r_{fp})}[D_{fp}(r_{fp})^2] \quad (5)$$

In Eq. 5, the generator G_{fa-fp} intends to fool the discriminator, which means that the generated image is discriminated by the discriminator with a true result, and the discriminator D_{fp} intends to distinguish the true from the false, i.e., it can distinguish the generated images from the real images. To constrain the training of the UWF FFA-to-FP translator G_{fa-fp}, the discrimination loss L_D was defined to train the discriminator D_{fp} to ensure that the distribution of the generated images matched that of the real UWF-FP images.

Furthermore, the calculation of the loss function $\mathcal{L}_{G_{DN}}$ is as follows:

$$\begin{aligned}\mathcal{L}_{G_{DN}} = {} & \mathbb{E}_{r_{fp} \sim p_{data}(r_{fp})}[(DN(r_{fp}) - L_{fp})^2] \\ & + \mathbb{E}_{r_{fa} \sim p_{data}(r_{fa})}[(DN(G_{fa-fp}(r_{fa}) - L_{fa})^2] \\ & + \mathbb{E}_{r_{fp} \sim p_{data}(r_{fp})}[(DN(G_{fa-fp}(r_{fp})) - L_{fp})^2] \end{aligned} \quad (6)$$

where L_{fp} is the diagnosis label for UWF-FP image r_{fp}, and L_{fa} is the diagnosis label for UWF-FP image r_{fa}.

Finally, the last part of $\mathcal{L}_{G_{total}}$ is $\mathcal{L}_{OLRM}$ which is given by

$$\begin{aligned}\mathcal{L}_{OLRM} = {} & \mathbb{E}_{l_{fp} \sim p_{data}(l_{fp})}[(DN'(l_{fp}) - O_{fp})^2] \\ & + \mathbb{E}_{l_{fa} \sim p_{data}(l_{fa})}[(DN'(G_{fa-fp}(l_{fa})) - O_{fa})^2] \end{aligned} \quad (7)$$

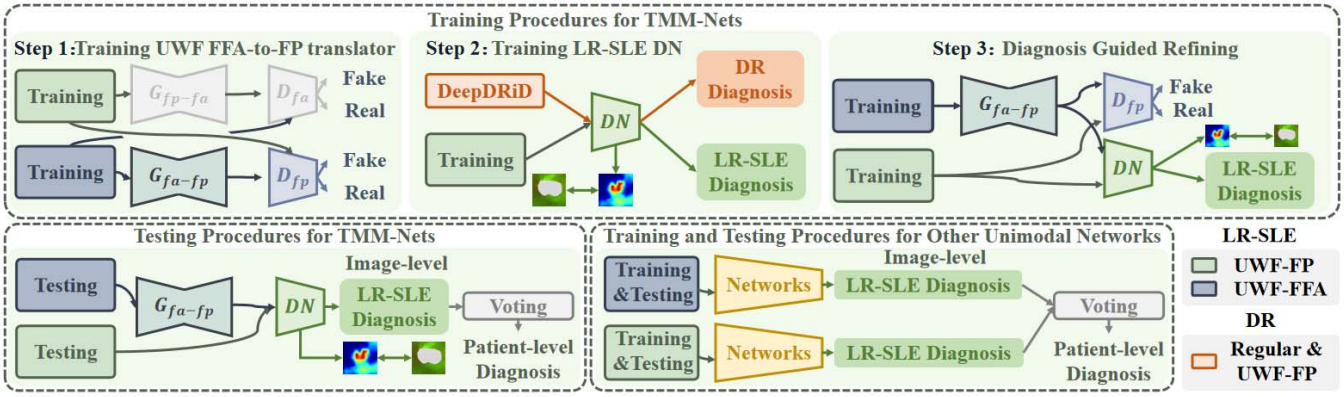

Fig. 4. Training and testing procedures for TMM-Nets and other compared networks. In the TMM-Net training procedures, the structural CycleGAN is firstly trained based on the UWF FFA-to-FP translator, and then the LR-SLE DN is trained. Finally, the UWF FFA-to-FP translator and LR-SLE DN are refined. In TMM-Net testing, the UWF-FFA image is firstly translated into a UWF-FP image using the UWF FFA-to-FP translator, and then both the translated and real UWF-FP images are employed in the LR-SLE DN to obtain the diagnosis results. Finally, a voting strategy is used to obtain the diagnosis based on the patient images. In other networks, UWF-FFA and UWF-FP images are utilized to train two separate models, and then voting is performed to obtain the final diagnosis.

where L_{fp} and L_{fa} are the UWF-FP and UWF-FFA images with LR-SLE lesion, respectively; DN' outputs the class attention map, which is matched with the lesion annotation of the physician; and O_{fp} and O_{fa} are the lesion annotation masks for the UWF-FP and UWF-FFA images, respectively.

IV. Experimental Results

A. Databases

We used two datasets in this study: a publicly available dataset from the ISBI 2020 Open Challenge—DeepDRiD and the private UWF dataset from Shanghai Jiao Tong University Affiliated Renji Hospital. All procedures were performed during routine patient checks. Institutional review board and ethics committee approval was obtained at all locations. The first dataset was acquired from the ISBI 2020 Open Challenge—DeepDRiD organized by our team and contains two parts [40].

DeepDRiD-Regular contains de-identified 2000 regular fundus images from 500 patients. Each Regular fundus image was acquired using a TOPCON digital fundus camera, and four fundus images were produced for each patient.

DeepDRiD-UWF includes 256 ultra-wide-field images from 128 patients. Optomap P200Tx (Optos, Dunfermline, UK) was used to acquire ultra-wide-field retinal images, and two images were obtwined from each patient, centered on the macula and optic disc.

The DeepDRiD dataset was previously published on GitHub: https://github.com/deepdrdoc/DeepDRiD. More information about DeepDRiD can be found in the paper [40].

Furthermore, we created the **Renji UWF dataset**, which contained de-identified UWF-FFA and UWF-FP images of 307 SLE patients, captured between 2014 and 2017 from the ophthalmology outpatient of Renji Hospital. We extracted 553 ultra-wide-field fundus images from 269 patients and 333 ultra-wide-field fluorescence angiography images from 129 of these SLE patients from the thousands of available examination reports to create our dataset. Because each patient underwent an eye examination, clinicians will chose either fundus photography or fluorography, or a combination of the two, depending on the patient's specific condition, the UWF-FFA and UWF-FP images in the Renji UWF dataset were unpaired and unmatched.

We counted 82 UWF-FP images and 31 UWF-FFA images that were manually annotated by the physicians to train and evaluate the attention lesion finder. Furthermore, we adopted a regional annotation method because of the laboriousness and difficulty of achieving consistent lesion annotation down to the pixel level. In this method, the original fundus image was divided into 8×8 squares, the physician determined the presence or absence of lesions for each square, and the square labeled with lesions was used as the lesion reminder area. For each image, two junior doctors performed separate evaluations, and the disputed areas were labeled by a senior doctor. The training set contained 56 diseased images, of which 44 UWF-FP images and 12 UWF-FFA images had manual lesion annotation. The test set included 26 diseased images, of which 18 UWF-FP images and 8 UWF-FFA images had manual annotation. The validation set consisted of 24 diseased images, of which 14 UWF-FP images and 10 UWF-FFA images had manual annotation. We further collected an external dataset for further validating the algorithm, captured between 2017 and 2018 from the ophthalmology outpatient of Renji Hospital. To create the external dataset, we extracted 60 ultra-wide-field fundus images from 34 SLE patients and nine ultra-wide-field fluorescence angiography images from nine SLE patients. There are seven images with LR-SLE in the external validation set, of which one was a UWF-FP image, and six were UWF-FFA images, all with manual annotation.

In addition, we construct a multiple fundus disease simulation (MFDS) dataset from a public dataset (RFMiD) [42] to further explore the performance of our model. The MFDS dataset has two kinds of fundus diseases: diabetic retinopathy, DR, and central retinal vein occlusion, CRVO, and consists of sub-dataset (the MFDS-DR dataset and the MFDS-CRVO

TABLE I

MODEL PERFORMANCE FOR LR-SLE DIAGNOSIS(I.E., NORMAL VERSUS LR-SLE) AND COMPARISON OF OUR TMM-NETS AND OTHER LEARNING MODELS, INCLUDING, RES-NET [41], CM-NET [14], CAB-NET [8], AND TDR-NET [13]. ALL OF THE INVOLVED VALIDATION METRICS ARE SHOWN ON THE IMAGE-LEVEL

Data used	Method	FLOPs (GMac)	Parameters (M)/	Testing Dataset				External Dataset			
				ACC(%)	SPE(%)	SEN(%)	F1(%)	ACC(%)	SPE(%)	SEN(%)	F1(%)
UWF-FP	Res-Net [41]	21.59	23.51	88.43	91.43	68.75	61.11	60.00	62.96	33.33	48.57
	CM-Net [14]	1477.20	345.27	89.26	94.29	56.25	58.06	66.67	68.52	50.00	64.91
	CAB-Net [8]	100.23	23.57	90.08	92.38	75.00	66.67	66.67	68.52	50.00	64.91
	TDR-Net [13]	695.27	19.25	91.74	94.29	75.00	70.59	68.33	70.37	50.00	64.96
UWF FFA & FP	Res-Net [41]	43.17	47.02	93.75	92.93	100.0	78.79	62.32	66.13	28.57	43.27
	CM-Net [14]	2954.60	690.55	95.54	95.96	92.31	82.76	71.01	72.58	57.14	71.00
	CAB-Net [8]	200.46	47.15	94.64	95.96	84.62	78.57	66.67	69.35	42.86	58.37
	TDR-Net [13]	1390.55	38.50	95.54	94.95	100.0	83.87	71.01	74.19	42.86	58.48
	TMM-Nets	357.20	42.24¥	**99.11**	**98.99**	**100.0**	**96.30**	**89.86**	**91.94**	**71.43**	**82.13**

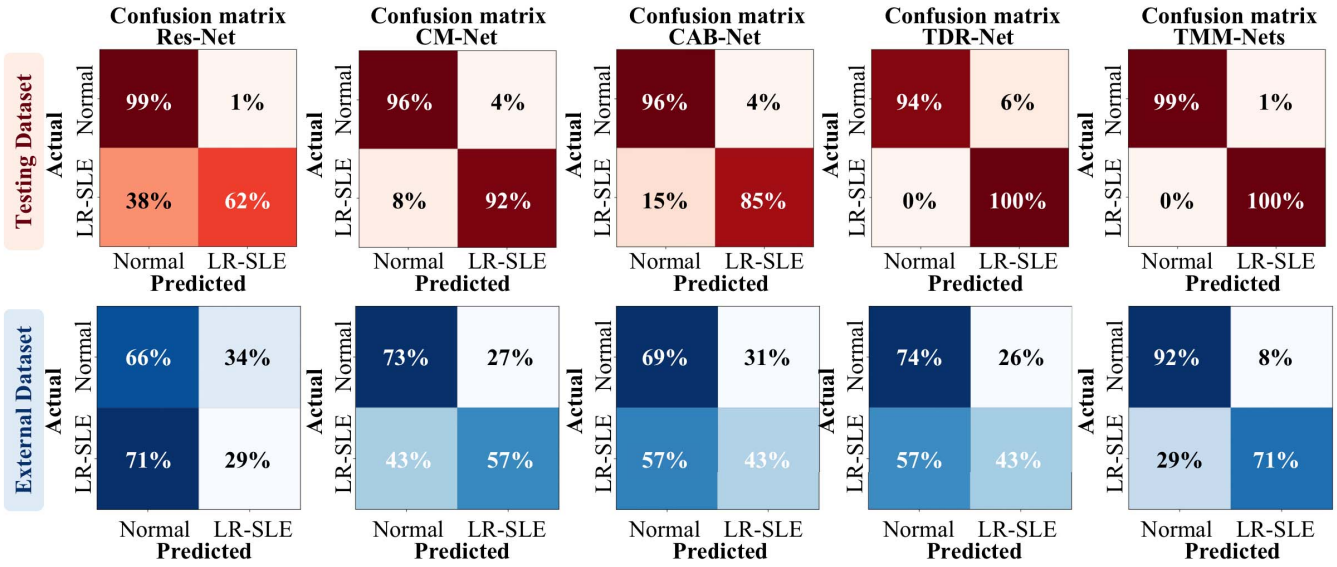

Fig. 5. Diagnostic performance comparison (confusion matrices) of our TMM-Nets and other methods (Res-Net [41], CM-Net [14], CAB-Net [8], and TDR-Net [13]). The percentages assigned to each category on the test set are given in each graph, with darker colors representing higher percentages. Red indicates the testing dataset results, and blue corresponds to the external testing dataset results.

dataset). The MFDS-DR dataset has 1070 regular fundus (401 with DR, 669 are normal), and the MFDS-CRVO dataset has 711 regular fundus images (42 with CROV, and 669 are normal). There are no fundus images containing DR and CROV in these two datasets. In the MFDS-DR dataset, the training set has 641 regular fundus images (240 with DR). The testing set of MFDS-DR has 214 regular fundus images (80 with DR), and there are 215 regular fundus images in the validation set of MFDS-DR (81 with DR). Moreover, the MFDS-CRVO dataset has 427 regular fundus images in the training set (42 with CRVO). The testing and validation sets have 142 regular fundus images (eight with CROV for testing and validation).

B. Implementation Details

1) Dataset Splitting: We divided the Renji UWF dataset into two parts: the UWF-FP dataset and UWF-FFA dataset, each of which was subsequently divided into three parts. Specifically, the UWF-FP dataset was divided into a training set (312 UWF-FP images from 60 patients, 17 of whom had LR-SLE), test set (121 UWF-FFA images from 61 patients, of which 7 had LR-SLE), and validation set (120 UWF-FP images from 58 patients, 7 of whom had LR-SLE). Similarly, the UWF-FFA dataset was divided into a training set (130 UWF-FFA images from 60 patients, 9 of whom had LR-SLE), test set (103 UWF-FFA images from 35 patients, 5 of whom had LR-SLE), and validation set (100 UWF-FFA images from 34 patients, of which 5 had LR-SLE). There was no overlap between the data in the UWF-FFA and UWF-FP validation sets and the other datasets on either the patient or image level.

2) Training Strategy: The structural Cycle-GAN was trained first to generate consistent images, and the Adam optimizer with recommended parameters was used to train the UWF FFA-to-FP translator, with the size of the mini-batch set to two. The generative model resized the images to 512 × 512. Next, we pre-trained the DN of LR-SLE using the DeepDRiD dataset to learn from lesions in DR, and we utilized the UWF-FP training set without introducing the generative images to diagnose LR-SLE. Finally, we refined the FFA-to-FP translator and LR-SLE DN using the diagnosis-guide network refining strategy. The network was trained on the Renji UWF training dataset, then fine-tuned on the Renji UWF validation dataset.

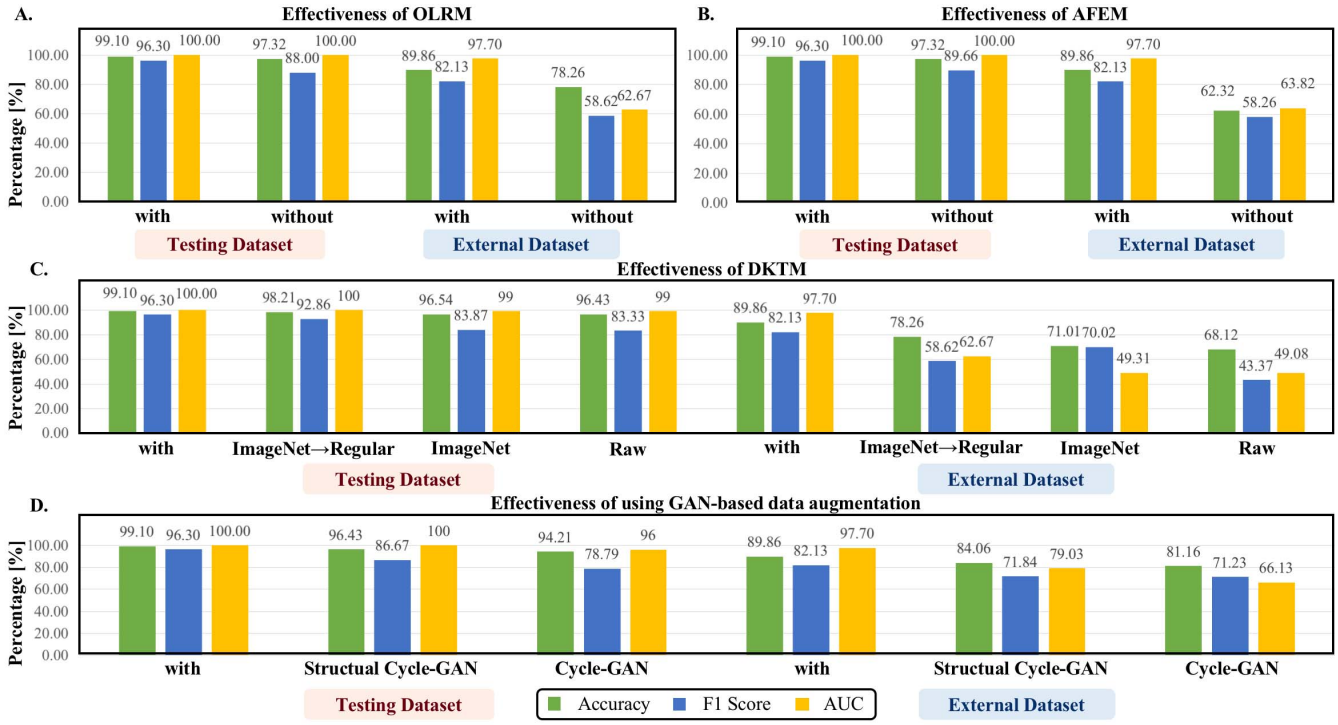

Fig. 6. Bar chart comparing the effects of different modules on TMM-Nets. **A.** Effects of OLRM on TMM-Nets. **B.** Effects of AFEM on TMM-Nets. **C.** Effects of DKTM on TMM-Nets for different cases. **D.** Effects of GAN-based data augmentation on TMM-Nets.

Finally, the model performance was validated on the Renji UWF test set, with the diagnostic model and associated tuning parameters selected based on the validation performance. The proposed network was implemented in Python using the Pytorch package and run on a computer equipped with two GPUs (NVIDIA GTX 2080ti 10 GB).

3) Inference Process: In the TMM-Net, the inference process was as follows. Firstly, the UWF-FFA images in the test set were fed into the UWF FFA-to-FP translator (G_{fa-fp}) to obtain the transferred UWF-FP images. Secondly, the transferred UWF-FP images were input into the LR-SLE diagnostic network (DN) together with the real UWF-FP images in the test set to obtain the diagnostic results and the lesion hints generated by the OLRM module for the positive diagnostic results. For the other lesion diagnostic networks being compared, the inference process was as follows. Firstly, the UWF-FFA images in the test set were input into the diagnostic network trained with UWF-FFA images to obtain the diagnostic results. After that, the UWF-FP images in the test set were input into the diagnostic network trained with UWF-FP images to obtain the diagnostic results. The final diagnostic results were obtained by integrating the two image results.

C. Evaluation Metrics

The classification performance was evaluated by utilizing four metrics, including accuracy (ACC= $\frac{TP+TN}{TP+TN+FP+FN}$), sensitivity (SEN= $\frac{TP}{TP+FN}$), and specificity (SPE= $\frac{TN}{TN+FP}$), where TP, TN, FP, and FN denote, respectively, the true positive, true negative, false positive, and false negative values.

TABLE II
CAM FIGURES COMPARED WITH PHYSICIAN ANNOTATIONS IN TERMS OF PRECISION, RECALL, AND DICE COEFFICIENT (MEAN ± SD)

Datasets	Precision	Recall	Dice
Testing Dataset	0.983±0.022	0.580±0.107	0.723±0.085
External Dataset	0.960±0.056	0.632±0.223	0.736±0.142

The F1 score was also given for evaluation and is defined as F1= $2 \cdot \frac{PRE \cdot SEN}{PRE+SEN}$, where Precision (PRE)= $\frac{TP}{TP+FP}$. The area under three receiver operating characteristic curve (AUC) was calculated based on all possible pairs of SEN and 1-SPE obtained by changing the thresholds performed on the classification scores yielded by the trained networks.

D. Diagnosis Performance

We have compared the performance of the LR-SLE classifications among our TMM-Nets and other state-of-art methods (i.e., Res-Net [41], CM-Net [14], CAB-Net [8], and TDR-Net [13]), are shown in Table I. Although these previous learning-based methods are among the most advanced currently proposed for diagnosing fundus lesions, our TMM-Nets obtained the outstanding results at the patient level. The F1 score of 96.30% was higher than that of the TDR-Net (83.87%) in the testing dataset that obtained the best performance among these state-of-art methods. In the external dataset, the TMM-Nets also achieved the best performance (F1: 82.13 %), whereas the best compared method obtain an F1 score of 71.00 %. Compared with the these previous methods that are not designed for unpaired multi-modal UWF-FP and

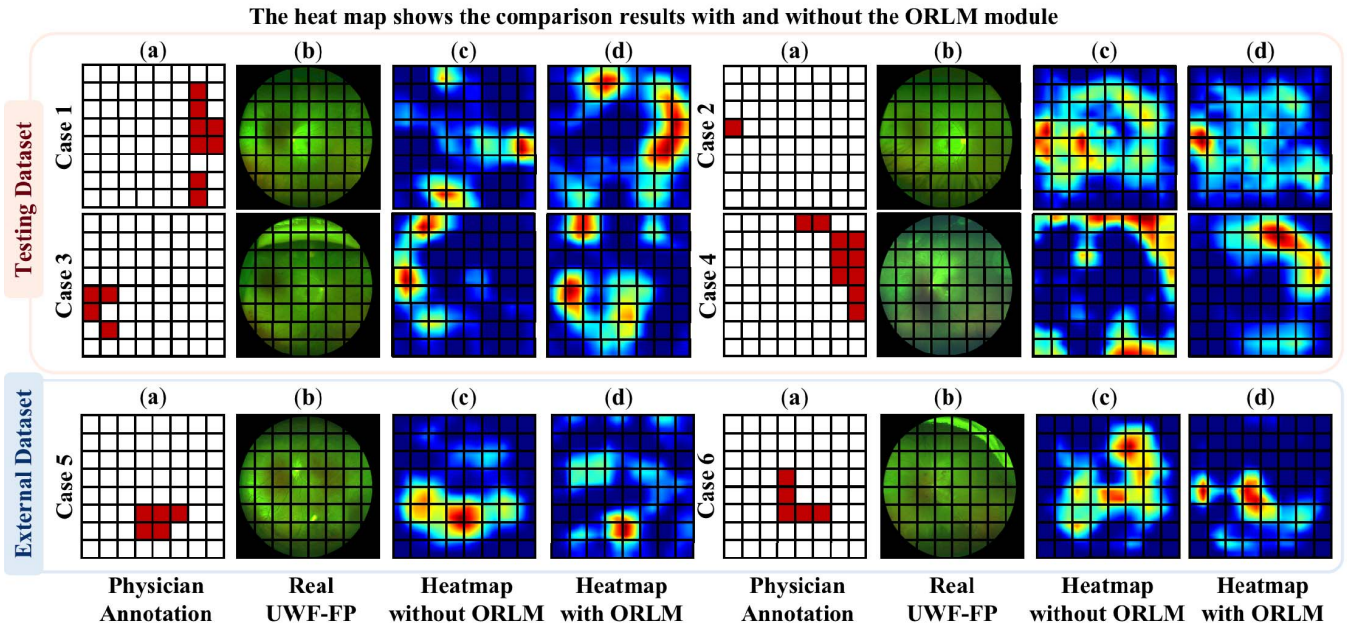

Fig. 7. The heatmaps using Grad-CAM method show the results with and without lesion area reminding module. In the annotation figures, the red squares are areas with lesion, and the black blocks represent 8×8 uniformly distributed cells of the UWF-FP images. (a) the lesion area annotations by senior physicians; (b) the raw UWF-FP images with 8×8 area segmentation blocks; (c) and (d) show the heatmaps under Grad-CAM without and with lesion area reminding module, respectively.

UWF-FFA data, the reason that our TMM-Nets achieved the highest performance is threefold. Firstly, the integration of the attention modules and lesion region enables the TMM-Nets to focus on the lesion-related region and to emphasize the characteristics of this region. Secondly, our DR lesion-based knowledge transfer and structural Cycle-GAN-based data augmentation combines and generalizes two distinct types of inputs into a mono-modal data distribution. Lastly, diagnosis-guided refinement for both generation and diagnosis can accurately capture the lesion-aware phenotype information and thereby increase the diagnosis precision. Fig. 5 compares the performance of these previous methods and our TMM-Nets. Due to the low incidence and clinical severity of LR-SLE, LR-SLE diagnosis should be highly sensitive and accurate. Our TMM-Nets has satisfied this clinical need by achieving a sensitivity of 100.0% and a precision rate of 92.86% in the testing dataset, which are significantly higher than the sensitivities and precisions of the other methods. In the external dataset, the TMM-Nets also achieved the highest sensitivity, specificity, precision and F1 Score.

Furthermore, TMM-Nets exhibit equal or better classification performance than the current state-of-the-art network on the MFDS dataset. On both the test and validation sets of MFDS, TMM-Nets obtained an F1 score of 93.33% and a sensitivity of 87.50%. Among the four networks (ResNet, CM-Net, CAB-Net, and TDR-Net) compared, only the CAB-Net was able to obtain a performance equal to that of TMM-Nets.

E. Ablation Study

We conducted experiments to illustrate the effectiveness of each module in the TMM-Nets based on metrics such as ACC, F1, and AUC.

1) Effectiveness of OLRM: We divided the UWF-FP images into small 8×8 cells for region annotation to assist doctors in labeling, and we used region reminder labels to facilitate LR-SLE diagnosis, which is challenging, and to reduce physician workload. The results with and without OLRM are shown in Fig. 6 **A**. In the testing dataset, the OLRM can realize improvements of 1.79% in ACC and 8.3% in F1-Score. In the external dataset, it provides an 11.60% improvement in ACC, a 23.51% improvement in F1, and a 35.03% improvement in AUC.

Fig. 7 also presents the heat maps used in our methods. Fig. 7 demonstrates that the heat maps generated by OLRM are largely consistent with the physician annotations. In addition, the comparison between the cases with and without lesion area reminders demonstrates the superior focusing performance of our method. To analyze the performance of the OLRM-based generated heat map quantitatively, we compared it with the physician-annotated 8×8 lesion attention annotations. We partitioned the heat map into 8×8 squares as well and calculated the Precision, Recall, and Dice Coefficient with the original physician annotations. The detailed results can be found in Table II.

2) Effectiveness of AFEM: We validated the effectiveness of the AFEM in this set of experiments. Firstly, we trained another network without the attention modules and compared its performance to those of the TMM-Nets. The corresponding results are shown in Fig. 6 **B**, indicating that training the network with the attention module can further enhance its diagnostic performance. In addition, to determine the optimal model for combining attention and feature fusion modules, we evaluated various methods for combining these modules. As illustrated in Table III, the results indicate that the accuracy

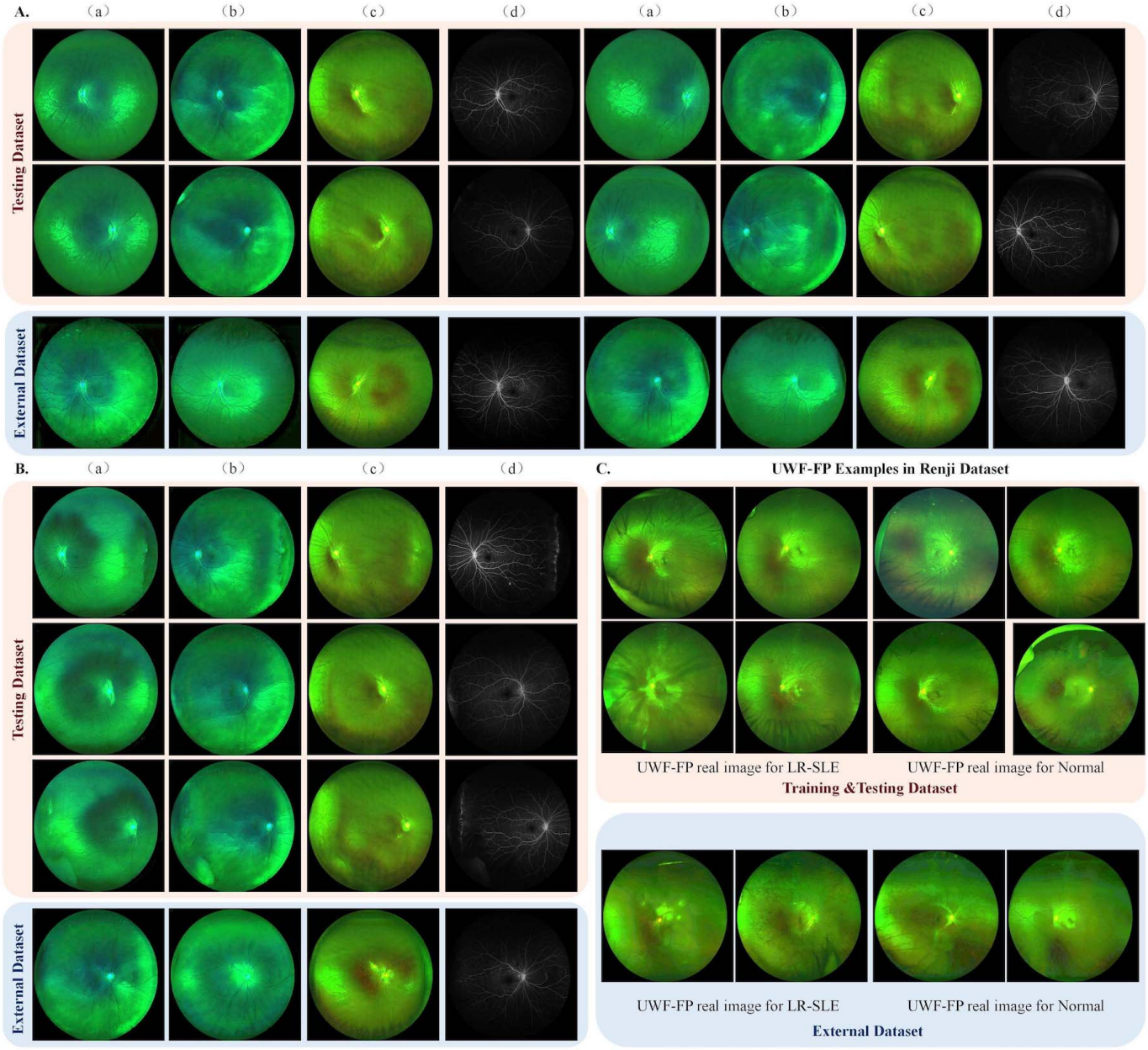

Fig. 8. Generation results. **A.** Results of generating UWF-FP images of patients with LR-SLE. **B.** Results of generating UWF-FP images of patients without LR-SLE. The results obtained using the structural Cycle-GAN and original Cycle-GAN are shown in (a) and (b), respectively; those obtained using the network based on the TMM-Nets are presented in (c); and the original UWF-FFA images are provided in (d). **C.** Original UWF-FP images, including images of normal and LR-SLE patients. Images in red rectangles are the generation results obtained using the testing dataset, and those in blue rectangles are those acquired using the external testing dataset.

TABLE III

MODEL PERFORMANCE IN TESTING DATASET FOR ATTENTION MODULE AT DIFFERENT LOCATIONS. *i* INDICATES THAT THE ATTENTION MODULE IS ADDED BEHIND THE i^{th} LAYER. + REPRESENTS THE CONNECTION SCHEME USING FEATURE FUSION MODULE

Metrics	Different Locations														
	1	2	3	4	1+2	1+3	1+4	2+3	2+4	3+4	1+2+3	1+2+4	1+3+4	2+3+4	1+2+3+4
ACC	94.64	96.43	95.54	97.32	96.43	97.32	98.21	96.43	99.11	97.32	95.54	97.32	95.54	96.43	94.64
F1	81.25	86.67	80.00	89.66	83.33	88.00	92.86	84.62	96.30	88.00	76.19	89.66	80.00	83.33	78.57

varies depending on the location of the attention module. The best accuracy is achieved by utilizing attention modules behind the second and fourth layers.

3) Effectiveness of DKTM: We verified the effectiveness of the DKTM for network training by performing four experiments utilizing four different methods to find the best model: 1) directly training a network for LR-SLE diagnosis (Raw); 2) using a pre-trained Image-Net (ImageNet); 3) using a DR regular fundus pre-trained based on 2), and then training the model to diagnose LR-SLE (ImageNet → Regular);

4) using an DR UWF fundus pre-trained based on 3), and then training the model to diagnose LR-SLE (with). The results are shown in Fig. 6 **C**. It can be concluded that constant knowledge transfer to initialize the network can further improve the diagnostic performance. The F1-Scores of these four methods indicate that transfer-learning domains that are more relevant to the problem can provide higher diagnostic performance.

We also explore the impact of using different inter-domain distance losses on the overall model performance in the framework of transfer learning. We compare the impact of MK-MMD loss, KL loss, and MMD loss on the performance of the final model. Compared to using MK-MMD, using both KL loss and MMD loss brings a decrease in model performance by 23.96% and 10.77%, respectively.

4) Effectiveness of Using GAN-Based Data Augmentation: We considered three experimental approaches to demonstrate the effectiveness of the GAN-based strategy: 1) using a Cycle-GAN; 2) using a structural Cycle-GAN; 3) using a GAN-based refining strategy for TMM-Nets. The results are shown in Fig. 6 **D**. In addition, structured Cycle-GAN data augmentation is used to assist in training the diagnostic network first; then, the structured Cycle-GAN-based trained generator and diagnostic network trained using the data augmentation results are combined to form a GAN-based framework for further analysis. Training improves the diagnostic performance and generates images that are more representative of the distribution of UWF-FP images.

F. Generation Ability

We conducted three experiments to evaluate the effectiveness of various approaches for generating UWF-FP images from UWF-FFA images during the training and testing phases of the generative model.

Experiment I: The classical unpaired generation training model, Cycle-GAN, was used to train the generative model for mapping UWF-FFA images to UWF-FP images, and the generation results are presented in Figs. 8 **A** (a) and 8 **B** (a). The generated results retain the information about the vascular distribution and optic disc morphology in the UWF-FFA images, but the generation of details in the secondary and tertiary vessels requires improvement. The overall image morphology and color show a slight greenish tint; thus, some differences from the real UWF-FFA image in Fig. 8 **C** remain.

Experiment II: We used the structural Cycle-GAN for training and testing, and the generation results are shown in Figs. 8 **A** (b) and 8 **B** (b). The generated results demonstrate that this scheme produces smoother details in the secondary and tertiary vessels, thus overcoming the challenge of obtaining detailed information. However, the overall image morphology and color rendering remain slightly greenish, and although the additional structure loss improves realism, the over-emphasis on structural similarity obscures some details.

Experiment III: A GAN-based training model consisting of the generative model and two discriminative models was employed, and some of the results are provided in Figs. 8 **A** (c) and 8 **B** (c). In comparison to the methods utilized in Experiments I and II, this method generates more details in the secondary and tertiary vessels. In addition, the overall image morphology and color rendering with a slight greenish tint are significantly improved, resulting in a high degree of subjective similarity to the real UWF-FFA image.

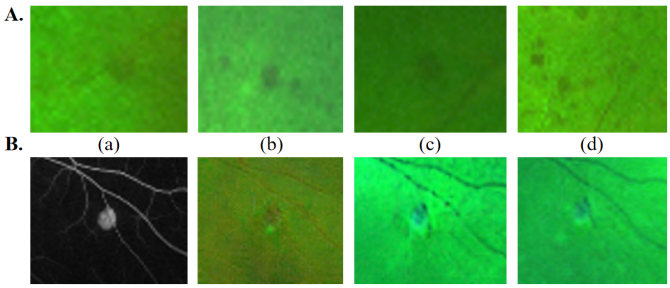

Fig. 9. Generation details in lesion parts. **A.** Lesion part details of real UWF-FP images. **B.** Real UWF-FFA lesion area image (a), image generated by structural Cycle-GAN (b), generation details of original Cycle-GAN model (c), and detailed image generated with diagnosis GAN-based retraining (d).

Furthermore, we include detailed images of the lesion in Fig. 9. Section **A.** in Fig. 9 depicts the various lesion components in real UWF-FP images, whereas Section **B.** presents the generation results obtained using various models based on the UWF-FFA images. The lesion portions visible in the real UWF-FFA images have distinct edges, whereas the lesion edges are fuzzy in real UWF-FP images. The results generated by the diagnosis GAN-based model exhibit the highest degree of similarity to the real UWF-FP image in terms of color tone and structure. The subjective evaluation indicates that the generated image is similiar to the real UWF-FP image, with the vascular and lesion features of the fundus preserved to a large extent, which clearly validates our method.

V. Conclusion

In this study, TMM-Nets and a corresponding training strategy were proposed to use unpaired multi-modal data for LR-SLE diagnosis, and outstanding UWF-FP image generation ability and disease diagnosis performance was achieved. The TMM-Nets combine transfer learning and a structural Cycle-GAN for augmentation. Online-CAM was developed to make the TMM-Nets focus on lesions for clinical interpretation, and GAN-based refinement was designed for model distillation. On the datasets with UWF-FP images of 269 subjects and UWF-FFA images of 129 subjects, the LR-SLE diagnosis effectiveness of our TMM-Nets was extensively evaluated and compared with that of the state-of-the-art methods. Our proposed method demonstrated excellent classification performance and can naturally integrate the unpaired and unstructured image data into the diagnosis workflow, showing great promise for the clinical deployment of LR-SLE. Broadly, TMM-Nets enable high sensitivity and precision in LR-SLE diagnosis and may have general applicability across other rare disease types.

References

[1] A. Kaul et al., "Systemic lupus erythematosus," *Nature Rev. Disease Primers*, vol. 2, no. 1, p. 16039, Jun. 2016.

[2] A. M. Canamary, J. M. D. Sousa, G. C. D. Andrade, and H. M. D. Nascimento, "Choroidopathy in systemic lupus erythematosus," *Revista Brasileira de Oftalmologia*, vol. 76, no. 2, pp. 230–234, 2017.

[3] T. Mathis et al., "Performance de la rétinophotographie en ultra-grand champ dans le dépistage de la rétinopathie diabétique," *J. Français d'Ophtalmologie*, vol. 42, no. 6, pp. 572–578, 2019.

[4] L. Dai et al., "A deep learning system for detecting diabetic retinopathy across the disease spectrum," *Nature Commun.*, vol. 12, p. 3242, May 2021.

[5] V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Amer. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, 2016.

[6] D. S. W. Ting et al., "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *J. Amer. Med. Assoc.*, vol. 318, no. 22, pp. 2211–2223, 2017.

[7] X. Ouyang et al., "Dual-sampling attention network for diagnosis of COVID-19 from community acquired pneumonia," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2595–2605, May 2020.

[8] A. He, T. Li, N. Li, K. Wang, and H. Fu, "CABNet: Category attention block for imbalanced diabetic retinopathy grading," *IEEE Trans. Med. Imag.*, vol. 40, no. 1, pp. 143–153, Jan. 2021.

[9] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[10] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE CVPR*, Jun. 2019, pp. 3146–3154.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, Oct. 2017, pp. 618–626.

[12] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, Oct. 2017, pp. 2242–2251.

[13] Y. Zhou, B. Wang, L. Huang, S. Cui, and L. Shao, "A benchmark for studying diabetic retinopathy: Segmentation, grading, and transferability," *IEEE Trans. Med. Imag.*, vol. 40, no. 3, pp. 818–828, Mar. 2021.

[14] H. Xie et al., "Cross-attention multi-branch network for fundus diseases classification using SLO images," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102031.

[15] F. Yu et al., "Annotation-free cardiac vessel segmentation via knowledge transfer from retinal images," in *Medical Image Computing and Computer Assisted Intervention*, vol. 11765. Berlin, Germany: Springer, 2019, pp. 714–722.

[16] G. M. Lee et al., "Unsupervised learning model for registration of multi-phase ultra-widefield fluorescein angiography," in *Medical Image Computing and Computer Assisted Intervention*, vol. 12263. Berlin, Germany: Springer, 2020, pp. 201–210.

[17] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. NIPS*, 2014, pp. 3320–3328.

[18] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*.

[19] M. Long, Y. Cao, Z. Cao, J. Wang, and M. I. Jordan, "Transferable representation learning with deep adaptation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 3071–3085, Dec. 2019.

[20] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. ICML*, vol. 70, 2017, pp. 2208–2217.

[21] Y. Tamaazousti, H. Le Borgne, C. Hudelot, M.-E.-A. Seddik, and M. Tamaazousti, "Learning more universal representations for transfer-learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2212–2224, Sep. 2020.

[22] S. Li et al., "Deep residual correction network for partial domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2329–2344, Jan. 2021.

[23] X. Li, X. Hu, L. Yu, L. Zhu, C.-W. Fu, and P.-A. Heng, "CANet: Cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1483–1493, Nov. 2020.

[24] Y. Xiao et al., "Dealing with long-tail issue in diabetic retinopathy and diabetic macular edema grading," in *Proc. 6th Int. Conf. Biomed. Eng. Appl. (ICBEA)*, 2022, pp. 40–46.

[25] I. J. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[26] R. He, J. Cao, L. Song, Z. Sun, and T. Tan, "Adversarial cross-spectral face completion for NIR-VIS face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1025–1037, May 2020.

[27] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101552.

[28] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE CVPR*, Jul. 2017, pp. 5967–5976.

[29] E. Jung, M. Luna, and S. H. Park, "Conditional GAN with an attention-based generator and a 3D discriminator for 3D medical image generation," in *Medical Image Computing and Computer Assisted Intervention* (Lecture Notes in Computer Science), vol. 12906. Berlin, Germany: Springer, 2021, pp. 318–328.

[30] Z. Xu et al., "Adversarial uni- and multi-modal stream networks for multimodal image registration," in *Medical Image Computing and Computer Assisted Intervention*, vol. 12263. Berlin, Germany: Springer, 2020, pp. 222–232.

[31] H. Uzunova, J. Ehrhardt, F. Jacob, A. Frydrychowicz, and H. Handels, "Multi-scale GANs for memory-efficient generation of high resolution medical images," in *Medical Image Computing and Computer Assisted Intervention*, vol. 11769. Berlin, Germany: Springer, 2019, pp. 112–120.

[32] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze- and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Jun. 2020.

[33] R. Liu et al., "NHBS-Net: A feature fusion attention network for ultrasound neonatal hip bone segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3446–3458, Jun. 2021.

[34] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10705–10714.

[35] K. Li, Z. Wu, K. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *Proc. IEEE CVPR*, Jun. 2018, pp. 9215–9223.

[36] Z. Xu et al., "Adversarial uni- and multi-modal stream networks for multimodal image registration," in *Medical Image Computing and Computer Assisted Intervention*, vol. 12263. Berlin, Germany: Springer, 2020, pp. 222–232.

[37] T. de Bel, J.-M. Bokhorst, J. van der Laak, and G. Litjens, "Residual cyclegan for robust domain transformation of histopathological tissue slides," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 102004.

[38] L. Zhou, J. D. Schaefferkoetter, I. W. K. Tham, G. Huang, and J. Yan, "Supervised learning with cyclegan for low-dose FDG PET image denoising," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101770.

[39] X. Jin, Y. Qi, and S. Wu, "CycleGAN face-off," 2017, *arXiv:1712.03451*.

[40] R. Liu et al., "DeepDRiD: Diabetic retinopathy—Grading and image quality estimation challenge," *Patterns*, vol. 3, no. 6, May 2022, Art. no. 100512.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, Jun. 2016, pp. 770–778.

[42] S. Pachade et al., "Retinal fundus multi-disease image dataset (RFMiD): A dataset for multi-disease detection research," *Data*, vol. 6, no. 2, p. 14, 2021.