Convolutional Network With Twofold Feature Augmentation for Diabetic Retinopathy Recognition From Multi-Modal Images

Cam-Hao Hua, Kiyoung Kim, Thien Huynh-The, *Member, IEEE*, Jong In You, Seung-Young Yu, Thuong Le-Tien, *Member, IEEE*, Sung-Ho Bae, *Member, IEEE*, and Sungyoung Lee, *Member, IEEE*

***Abstract*—*Objective:* With the scenario of limited labeled dataset, this paper introduces a deep learning-based approach that leverages Diabetic Retinopathy (DR) severity recognition performance using fundus images combined with wide-field swept-source optical coherence tomography angiography (SS-OCTA). *Methods:* The proposed architecture comprises a backbone convolutional network associated with a Twofold Feature Augmentation mechanism, namely TFA-Net. The former includes multiple convolution blocks extracting representational features at various scales. The latter is constructed in a two-stage manner, i.e., the utilization of weight-sharing convolution kernels and the deployment of a Reverse Cross-Attention (RCA) stream. *Results:* The proposed model achieves a Quadratic Weighted Kappa rate of 90.2% on the small-sized internal KHUMC dataset. The robustness of the RCA stream is also evaluated by the single-modal Messidor dataset, of which the obtained mean Accuracy (94.8%) and Area Under Receiver Operating Characteristic (99.4%) outperform those of the state-of-the-arts significantly. *Conclusion:* Utilizing a network strongly regularized at feature space to learn the amalgamation of different modalities is of proven effectiveness. Thanks to the widespread availability of multi-modal retinal imaging for each diabetes patient nowadays, such approach can reduce the heavy reliance on large quantity of labeled visual data. *Significance:* Our TFA-Net is able to coordinate hybrid information of fundus photos and wide-field SS-OCTA for exhaustively exploiting DR-oriented biomarkers. Moreover, the embedded feature-wise augmentation scheme can enrich generalization ability efficiently despite learning from small-scale labeled data.**

***Index Terms*—Convolutional network, diabetic retinopathy recognition, fundus photograph, multi-modal images, SS-OCT angiography, twofold feature augmentation.**

I. Introduction

DIABETIC Retinopathy (DR), the complication resulted from being afflicted with diabetes mellitus over a long period of time, is among the most common causes of visual impairment and blindness [1]. Traditionally, the DR grade is determined based on the combined evaluation of different structural features presented in the color fundus images, for instance, existence of microaneurysms, exudates, hemorrhages, and neovascularization [2]–[4]. Accordingly, five severity scales including no apparent retinopathy (no DR), mild nonproliferative DR (NPDR), moderate NPDR, severe NPDR, and proliferative DR (PDR) have been proposed as international clinical classification system based on the Early Treatment Diabetic Retinopathy Study [5]. However, such grading process remains time-consuming and challenging due to the heavy dependence on examiner as well as errors from low quality of the screened photos or missing subtle details. Moreover, it can be observed that effective diagnosis of DR severity level allows the ophthalmologists to deploy proper treatment procedure for the prevention of vision deterioration. These lead to the fact that the research topic of automatic DR detection from retinal-based images shows great interest in both ophthalmology and modern computer vision domains nowadays.

To this end, with the qualitative and quantitative advancements of computational resources and images, respectively, deep learning (DL) technique has been intensively exploited in computer vision. Particularly, Convolutional Neural Network (CNN), a powerful DL architecture, has been applied in diverse vision-oriented areas ranging from natural image classification [6]–[9], human action recognition [10], [11] to

Manuscript received June 13, 2020; revised September 22, 2020 and November 3, 2020; accepted November 28, 2020. Date of publication December 2, 2020; date of current version July 20, 2021. This work was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2017-0-01 629) supervised by the IITP (Institute for Information & communications Technology Promotion); by IITP grant funded by the Korea government (MSIT) (No. 2017-0-00 655); by the MSIT, Korea, under the Grand ITRC support program (IITP-2020-0-01 489) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation); by NRF-2016K1A3A7A03951968; and by NRF-2019R1A2C2090504. *(Cam-Hao Hua and Kiyoung Kim are co-first authors.) (Corresponding authors: Sung-Ho Bae; Sungyoung Lee.)*

Cam-Hao Hua, Sung-Ho Bae, and Sungyoung Lee are with the Department of Computer Science and Engineering, Kyung Hee University, Yongin, Gyeonggi-do 17 104, Republic of Korea (e-mail: hao.hua@oslab.khu.ac.kr; shbae@khu.ac.kr; sylee@oslab.khu.ac.kr).

Kiyoung Kim, Jong In You, and Seung-Young Yu are with the Department of Ophthalmology, Kyung Hee University Medical Center, Kyung Hee University, Dongdaemun-gu, Seoul 02 447, Republic of Korea (e-mail: pourma@khu.ac.kr; jiyou@khu.ac.kr; syyu@khu.ac.kr).

Thien Huynh-The is with the ICT Convergence Research Center, Kumoh National Institute of Technology, Gumi, Gyeongsangbuk-do 39 177, Republic of Korea (e-mail: thienht@kumoh.ac.kr).

Thuong Le-Tien is with the Department of Electrical and Electronics Engineering, Ho Chi Minh City University of Technology, Ho Chi Minh City 700 000, Vietnam (e-mail: thuongle@hcmut.edu.vn).

Digital Object Identifier 10.1109/JBHI.2020.3041848

biomedicine [12], [13], medical image segmentation [14], and DR risk prognosis [1], to name a few. As a result, utilizing CNN to recognize DR severity scales from fundus images has also attracted numerous researches [15]–[25] thanks to the representational power of the above-mentioned biomarkers for automatic diagnosis. Nonetheless, it is worth noting that large-scale dataset [26] was utilized in those studies for attaining an excellent performance of detecting the DR grades. Additionally, the model training cost becomes a significant concern due to dealing with a large number of high-resolution images.

On the other hand, optical coherence tomography angiography (OCTA), a novel type of noninvasive retinal imaging, can provide the visualization of retinal microvasculature changes, which is incapable in fundus photo. As such, several OCTA-based researches [27]–[31] investigated that the decrease of capillary density and increase of non-perfusion area signify the exacerbation of DR within a limited field of view (FOV). More recently, swept-source OCTA (SS-OCTA) has been shown to offer sufficient capillary details of mid-peripheral retina in a wide-field single scan (12×12 mm^2 or 15×9 mm^2) owing to its longer wavelength and faster acquisition time [32]. Accordingly, novel studies of wide-field SS-OCTA in DR [27]–[29] remarkably reported that higher percentage of non-perfusion region on peripheral sector is a more robustly diagnostic marker of DR severity compared to that of central sectors. In a nutshell, compared to fundus photograph, the wide-field SS-OCTA offers better visualization of retinal vein, artery, and especially capillary, from which the measures of avascular (a.k.a. non-perfusion) areas and/or density can be considered as important biomarkers related to DR severity.

From those observations, we hypothesize that amalgamating SS-OCTA 12×12 mm^2 with fundus image in a well-regularized DL model is able to detect DR severity effectively in the case of limited dataset size. There are four remarkable reasons supporting such problem statement as follows. Firstly, the availability of multi-modal imaging to each diabetes mellitus patient is widespread nowadays for facilitating the ophthalmologists to diagnose abnormalities more efficiently. Secondly, fundus images and SS-OCTA 12×12 mm^2 are noninvasive, which allows to conveniently capture both of them from the patients in practice. Thirdly, they share similar FOV, wherein different clinical manifestations of these two modalities can supplement each other. Fourthly, the additional involvement of wide-field SS-OCTA enables the DL model to acquire better generalized DR-related features at each learning step. As a result, the reliance on single-modal (e.g., conventional fundus images [26]) dataset, which should be large enough to cover all-round DR properties, is alleviated for the training process.

Accordingly, we introduce a Convolutional Neural Network with Twofold Feature Augmentation strategy, namely TFA-Net, to leverage the capability of feature-level generalization from the bimodal inputs given a small amount of observed data. In particular, the proposed architecture includes weight-sharing convolutional layers followed by a pretrained backbone CNN having Reverse Cross-Attention (RCA) stream, wherein the augmentation at feature space is twofold. With respect to the first one, the early convolutional layers are utilized for concurrently learning both types of inputs, i.e., fundus photograph and wide-field SS-OCTA, from which the same weights (parameters) are updated interchangeably as an augmentation procedure for boosting the generalization. Regarding the second stage, a RCA stream is deployed to stair-wisely refine typical fine-grained details of low-level feature maps using depth-wisely semantic context of high-level counterparts [14], [33] in the proposed network. As a consequence, not only spatially informative representations of DR-oriented clues are extensively involved but also the learned feature responses are continuously manipulated in a feedback-like manner for further enriching the generalization ability. In other words, our TFA-Net is able to exhaustively exploit the hybrid information in the bimodal inputs at feature levels, which subsequently benefits the process of learning from limited data for DR grade prediction. Notably, to the best of our knowledge, this is the first work that incorporates fundus image and wide-field SS-OCTA for DR detection using DL especially on the small-scale dataset.

As for the evaluation step, we employ a domestic small-scale dataset of 297 patients taken from Kyung Hee University Medical Center (KHUMC), Seoul, Republic of Korea, wherein each individual owns a pair of fundus image and wide-field SS-OCTA. Subsequently, the experimental performance in terms of Quadratic Weighted Kappa (QWK) is an indicator for the effectiveness of the proposed approach in classifying DR scales given a small number of labeled visual data. Additionally, we employ the Messidor dataset [34] to prove the superiority of the proposed RCA stream over the existing approaches regarding popular Accuracy and Area Under Receiver Operating Characteristic (AUROC) metrics.

In overall, this study offers four key contributions as below

- We introduce the study that learns the amalgamation between fundus images and wide-field SS-OCTA using DL on a limited dataset size for DR severity screening.
- We propose a Convolutional Network with Twofold Feature Augmentation to intensively enrich the generalization capability at feature level, which is smoothly compatible with multi-modal small-scale dataset.
- We perform a 5-fold cross validation procedure to prove the effectiveness of the proposed methodology on a domestic small-scale dataset from Kyung Hee University Medical Center, Seoul, Republic of Korea.
- We show the robustness of the RCA stream (i.e., the second stage of the TFA mechanism) compared to state-of-the-art approaches through a single-modal evaluation on the public Messidor dataset [34].

Subsequent sections in this article are organized as follows. Section II discusses about related work, wherein fundus images are handled by DL technique for DR grading task. Section III delivers overview of the whole architecture and then provides an in-depth description of the proposed Twofold Feature Augmentation scheme. Section IV presents the benchmark datasets, evaluation metrics, implementation details, and experimental results with corresponding analyses. Section V mentions existing limitations and future work. Section VI encapsulates the findings of this study.

II. Related Work

We categorize the existing CNN-based methods using fundus images for DR grade detection into two primary groups: (i) the employment of the CNN built by common trainable and non-linear layers for classification [15]–[20] and (ii) deep architecture of multiple network streams learning through ensemble scheme [21]–[25].

Regarding the first branch, the corresponding models are custom 11-layer [15], 17-layer [16], 18-layer [17], and 20-layer [18] CNNs, of which the fundamental constituents are convolutional, ReLU activation, max/average pooling, and FC layers, followed by a Softmax classifier. Notably, in SI2DRNet-v1 [18], a convex post-prediction mechanism attached at the end of the network is argued to be a key determinant gaining an impressive recognition performance. Meanwhile, both Gulshan *et al.* [19] and Sahlsten *et al.* [20] fine-tuned a lightweight deep learning architecture, called Inception-v3 [35], to address the DR severity grading issue. Model in the former work is trained with a very large-sized dataset (more than 128 000 retinal images) and subsequently achieves impressive performance in terms of specificity and sensitivity for referable DR recognition. On the other hand, the latter further exploits the correlation between higher-resolution inputs and smaller-sized sample pool by various experiments with different private datasets.

In the second branch, Vo *et al.* [21] introduced two modified versions of VGG [6] and GoogleNet [7], namely VNXK- and CKML-Net, respectively, to predict the DR grades. Also, the authors adopted L-, green-, and I_1-channel versions of the original fundus images as inputs of those two networks to combine the prediction scores for boosting classification performance. Recently, Ting *et al.* [22] proposed a similar idea, wherein a raw input fundus image and its local contrast-normalized modality are fed into two separated VGG networks [6]. Then, the corresponding output scores are combined for finalizing the DR grading result. Different from the input-based ensemble learning manner in [21], [22], manifold CNN streams comprising Main, Crop, and Attention-based Networks [23] were taken into account to aggregate DR-oriented signs from various viewpoints on a same input image at feature space. On the other hand, a deep multiple instance learning method [24] was introduced to comprehensively extract DR-oriented information. Particularly, an input fundus image is exploited at multiple scales, wherein various patches of each version are fed into different predefined CNNs. As a result, the acquired patch-level feature maps are aggregated through all the scales to produce an averaged DR map for the final classifier. In contrast, to reduce the expensive computation of multi-stream networks, Junjun *et al.* [25] embedded an auxiliary spatial attention mechanism, which infers region scoring maps (RSMs), into the backbone ResNet-18 [8]. Consequently, the features calibrated by the RSMs were proved to be more discriminative and robust for DR grade prediction than those learned in the baseline model.

In brief, for inferring the DR severity levels effectively, the existing approaches either only rely on high-level features extracted from the deep CNNs or extensively adopt ensemble strategies by involving multiple input formats/scales and corresponding late feature maps. It is obvious that the former faces the issue of losing small-sized clinical signs (e.g., microaneurysms, hemorrhages) due to manifold downsampling stages in the CNN, while the latter carries costly computation because of employing the multi-network regime. On the contrary, as aforementioned, the proposed architecture aims to aggregate features of different semantic levels through a lightweight stream for encoding DR-related biomarkers of various scales efficiently. Furthermore, the unique combination between wide-field SS-OCTA and fundus images in our work is able to enrich the pool of representational features for tackling the difficulty of training with small-scale dataset.

III. Methodology

Notably, this study was performed in accordance with the Declaration of Helsinki. Kyung Hee University Medical Center institutional review board (IRB) approved the study protocol (IRB No. KHUH202008028, date of approval: 2020-03-15).

A. Architecture Overview

Regarding the architecture illustrated in Fig. 1, our TFA-Net comprises a backbone network acquiring deep features through multiple blocks of convolutional layers (enclosed by the area having a boundary of dashed gray lines), which are associated with a trainable two-stage feature augmentation mechanism (covered by the beige regions). About the former, we adopt the residual blocks in ResNet [8] as the base feature extractors thanks to its powerful skip-connection strategy for model's parameters optimization. Specifically, there are four sequentially residual-based convolution blocks in the ResNet, for each of which the amount of convolutional and non-linear activation layers may be different from one another. Moreover, the final output features of each block are further involved in the attached RCA stream. As for the latter constituent, as the core contribution of this work, the two stages of feature-level augmentation are taken into account to combat the overfitting issue in the case of training with multi-modal inputs under the limitation of dataset size. Subsequently, details of the TFA strategy are conveyed in following sub-sections.

B. Twofold Feature Augmentation (TFA)

1) Stage 1 - Weight-Sharing Convolution Kernels: Since the two input images possess a highly similar FOV as presented in Fig. 1, we aim to employ convolutional layers which should acquire hybrid finely-patterned features in an effective and generalized way. It is worth noting that a convolutional layer is defined as the pipeline of convolution kernels, batch normalization [36], and rectified linear unit (ReLU) activation function in this work. As mentioned before, fundus photo exposes the emergence of microaneurysms, soft/hard exudates, hemorrhages, etc. Meanwhile, the SS-OCTA 12×12 mm^2 contributes better visualization of non-perfusion areas and retinal vessel density. Therefore, weight-sharing convolution layers followed by depth-wise concatenation are adopted as they can perceive those distinct DR-oriented biomarkers from different modalities,

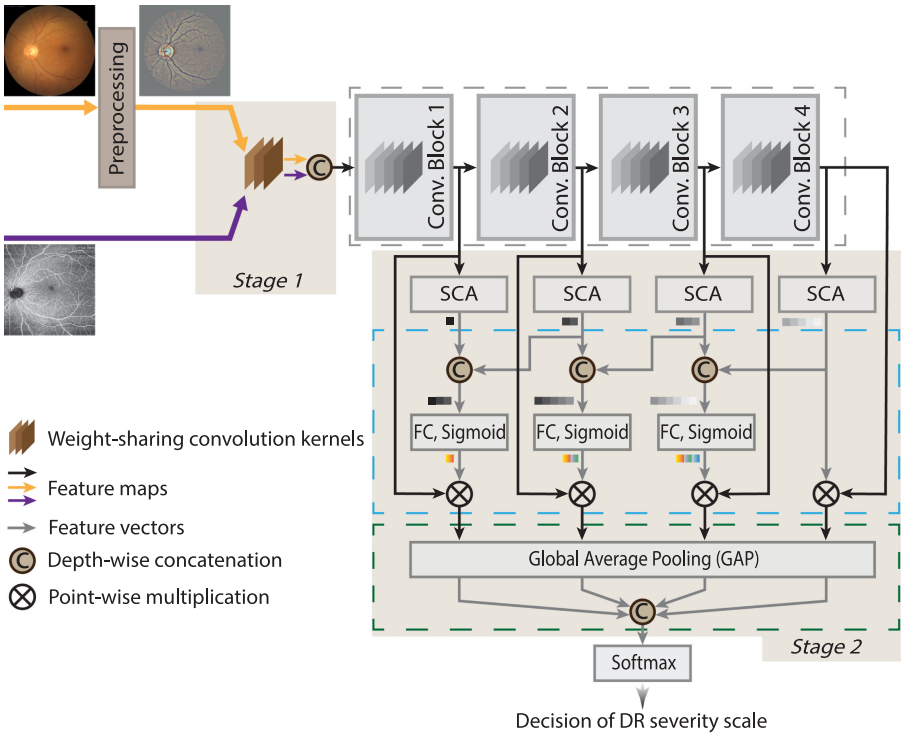


Fig. 1. Architecture of our TFA-Net for DR recognition. Notably, the input wide-field SS-OCTA is mirrored to form into a three-channel image like the preprocessed fundus version so that these two modalities can then be together transformed by the weight-sharing convolution kernels. 'Conv. Block' and 'SCA' stand for the blocks of predefined convolutional layers and Self-Context Aggregation, respectively. The area having a boundary of dashed blue lines represents processes of the Pairwise Reverse Attention component, while that of dashed green lines embodies operations of the Multi-level Fusion component in Stage 2. 'FC, Sigmoid' means fully connected layers followed by the Sigmoid activation function. View in color is recommended.

and then learn to adapt the shared characteristics for superior generalization.

In the case of using separated convolutional layers for each input type, the amount of training data should be large enough to make the obtained features more robust. Meanwhile, by the utilization of weight-sharing strategy, the same kernels can selectively collect clinical details of interest from the two modalities right at the beginning of our model. Subsequently, the learned parameters are continuously revised in order to effectively represent essential cross-modal features. Thus, we argue that such progressive manipulation of the shared weights with respect to multi-modal inputs can be considered as a feature-wise augmentation stage, which may ease the significant dependence on large-scale dataset for training.

Let I_f and $I_o \in \mathrm{R}^{\mathrm{H}\times\mathrm{W}\times 3}$ be the preprocessed input fundus image and wide-field SS-OCTA, respectively. Note that we replicate the raw SS-OCTA itself to form into a three-channel image like the fundus version. Then, the first stage of our TFA scheme is executed as follows

$$F_c = \mathcal{C}[ReLU(\mathbf{W}_{shared} * I_f), ReLU(\mathbf{W}_{shared} * I_o)] \quad (1)$$

where $\mathbf{W}_{shared} \in \mathrm{R}^{7\times 7\times 3}$ are trainable weights of 32 shared convolution kernels having stride of 2 and padding of 3; $\mathcal{C}[.]$ indicates the depth-wise concatenation; and $F_c \in \mathrm{R}^{\frac{\mathrm{H}}{2}\times\frac{\mathrm{W}}{2}\times 64}$ are the learned feature maps gained from the first stage of our feature-level augmentation mechanism. For simplicity, the operation of batch normalization is not included in (1). Afterwards, the F_c are fed into the aforementioned backbone CNN for higher-level feature representation.

2) Stage 2 - Reverse Cross-Attention Stream: It is obvious that semantically-richer features are encoded along channel dimension by deeper layers, but subject to lower spatial resolution (with strides of 4, 8, 16, and 32 following the four convolution blocks in Fig. 1) and vice versa. Therefore, as the Stage-2 region in Fig. 1 illustrates, we propose an RCA stream coupled with the backbone ResNet [8] for leveraging the impact of finely-patterned features at earlier layers on the final prediction. Concretely, channel-wise semantic details of the higher-level features are utilized to enhance the informative responses while mitigating the less effective ones in feedback-like manner. As a consequence, such reverse refinement brings two noticeable benefits as follows. Firstly, it allows encoded DR-oriented intimations that rely on spatial representations (e.g., soft/hard exudates and avascular zone) to gain extra emphases on the final output. This activity is achievable since finer-resolution (i.e., low-level) feature maps, of which the semantic information is much enhanced by higher-level context reversely, can be early engaged to the Softmax classifier without significant obscurity. Secondly, it acts as an extensive augmentation procedure at multiple feature levels because our TFA-Net has an additional learning stream of backward and parallel styles besides the main feedforward path.

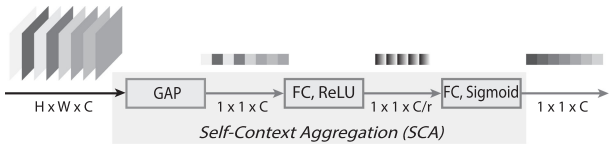


Fig. 2. Functional layers in the Self-Context Aggregation module. 'GAP' signifies the global average pooling layer.

In general, the proposed RCA stream consists of three major components, i.e., (i) Self-Context Aggregation (SCA) inspired from Hu *et al.* [9], (ii) Pairwise Reverse Attention (PRA), and (iii) Multi-level Fusion (MLF). The corresponding details are given as follows.

(a) Self-context aggregation: At first, the four chosen feature maps (i.e., final output of the fundamental blocks of convolutional layers in ResNet) are fed into corresponding SCA modules for individually exploiting semantic context encoded along the depth dimension. Let $\boldsymbol{F}_n$ denote those feature maps of interest, where $n = 1, \ldots, 4$ such that larger n indicates the higher-level features, which have semantically-richer context but smaller spatial size. Subsequently, the process of aggregating self-context shown in Fig. 2 is initially performed by a global average pooling (GAP) layer, which is $\mathcal{G} : \boldsymbol{F}_n \in \mathrm{R}^{\mathrm{H_n x W_n x C_n}} \rightarrow \boldsymbol{g}_n \in \mathrm{R}^{\mathrm{C_n}}$. The corresponding formulation of $\mathcal{G}$ is defined as

$$\boldsymbol{g}_{nc} = \mathcal{G}(\boldsymbol{F}_n) = \frac{1}{H_n \times W_n} \sum_{h=1}^{H_n} \sum_{w=1}^{W_n} \boldsymbol{F}_{n(h,w,c)} \tag{2}$$

where $h = 1, \ldots, H_n$; $w = 1, \ldots, W_n$; and $c = 1, \ldots, C_n$ are height, width, and channel coordinates of pixels in the considered feature maps $\boldsymbol{F}_n$, respectively.

Then, fully connected (FC) layers followed by ReLU activation are applied to exploit underlying cross-channel interactions of the retrieved vectors $\boldsymbol{g}_n$. Formally,

$$\begin{aligned} \boldsymbol{i}_n &= ReLU(\mathbf{W}_{fc1_n}^T \boldsymbol{g}_n + \mathbf{B}_{fc1_n}) \\ \boldsymbol{s}_n &= \sigma(\mathbf{W}_{fc2_n}^T \boldsymbol{i}_n + \mathbf{B}_{fc2_n}) \end{aligned} \tag{3}$$

where $\{\mathbf{W}_{fc1_n} \in \mathrm{R}^{\mathrm{C_n} \times \frac{\mathrm{C_n}}{\mathrm{r}}}$, $\mathbf{B}_{fc1_n} \in \mathrm{R}^{\frac{\mathrm{C_n}}{\mathrm{r}}}\}$ and $\{\mathbf{W}_{fc2_n} \in \mathrm{R}^{\frac{\mathrm{C_n}}{\mathrm{r}} \times \mathrm{C_n}}$, $\mathbf{B}_{fc2_n} \in \mathrm{R}^{\mathrm{C_n}}\}$ are respectively trainable parameters of two *FC* layers in use; $\boldsymbol{i}_n \in \mathrm{R}^{\mathrm{C_n}/\mathrm{r}}$ and $\boldsymbol{s}_n \in \mathrm{R}^{\mathrm{C_n}}$ are intermediate and final outputs of the SCA module, respectively; and $\sigma(.)$ symbolizes the Sigmoid activation function that weights vectors' entries from 0 to 1 based on corresponding utilities. It is noted that value of r, the compression rate for saving computational cost, is set to 16 following Hu *et al.* [9]. Besides that, the lengths C_n of $\boldsymbol{s}_n$, where $n = 1, 2, 3, 4$, are determined based on the ultimate output's channel size of the four fundamental convolution blocks. For instance, using ResNet-18 as the backbone introduces $C_n = \{64, 128, 256, 512\}$ while the 50- and 101-layer counterparts give $C_n = \{256, 512, 1024, 248\}$.

Remarkably, in the original work [9], the output representational vector of this SCA module is subsequently used to re-calibrate its input feature map only at every layer, which can be referred to as intra-feature attention. Meanwhile, the counterpart in our TFA-Net is employed to further incorporate with the corresponding version at lower scale for performing both intra- and inter-feature (in a reversely cross manner as described at next sub-section) attention tasks. Another noteworthy difference is that the SCA blocks in the proposed model are only involved at the end of the four predefined convolution blocks in the backbone network.

(b) Pairwise reverse attention: Clearly, previous step only introduces the utilization of intra-relationships across channels within each individual feature map taken into account. To this end, we additionally exploit semantic inter-dependencies between the considered features by uniquely learning all pairwise concatenation of the self-context vectors, i.e., $\boldsymbol{s}_n$ and $\boldsymbol{s}_{n+1}$, where $n = 1, 2, 3$. This allows deeper feature maps involved from the backbone CNN to enrich semantic representations onto the shallower counterparts reversely, which then suggests two advantages. Firstly, the refined low-level features, which possess high resolution, have stronger contributions since they can be alternatively applied as a shortcut to the final classifier. As a result, characterizations of small-sized factors related to early DR (e.g., microaneurysms, hemorrhages, or capillary abnormalities), which may certainly get loss at later layers due to spatial pooling operations, can be apprehended extensively to improve the recognition performance. Secondly, it is argued that incorporating a stream of manipulating multi-level features in reverse fashion can be considered as another intensive procedure of feature-level augmentation for avoiding overfitting issue.

According to center part of the Stage-2 region in Fig. 1, the workflow of this PRA module is formulated as follows.

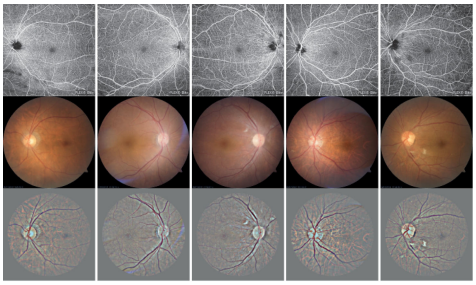
$$\begin{aligned} \boldsymbol{F}_{pra_4} &= \boldsymbol{F}_4 \otimes \boldsymbol{s}_4 \\ \boldsymbol{F}_{pra_n} &= \boldsymbol{F}_n \otimes \sigma(\mathbf{W}_{fc3_n}^T (\mathcal{C}[\boldsymbol{s}_n, \boldsymbol{s}_{n+1}]) + \mathbf{B}_{fc3_n}) \end{aligned} \tag{4}$$

where $n = 1, 2, 3$ in this step; $\otimes$ refers to as the point-wise multiplication at each channel; and $\{\mathbf{W}_{fc3_n} \in \mathrm{R}^{(\mathrm{C_n}+\mathrm{C_{n+1}}) \times \mathrm{C_n}}$, $\mathbf{B}_{fc3_n} \in \mathrm{R}^{\mathrm{C_n}}\}$ denote the parameters of the *FC* layers followed by another Sigmoid activation function. These learning layers manage the integration of features having semantically-richer information into those with finer representation of spatial-based details. Notably, such reverse combinations only take place in pairwise fashion to ensure reasonable increment of computational burden and refrain low-level self-context vectors from overwhelming acquisition of heterogeneous higher-level information. Then, the cross-context output vectors are utilized for re-calibrating the corresponding feature maps $\boldsymbol{F}_n$ via point-wise multiplication along channel dimension. Afterwards, the retrieved results, denoted as $\boldsymbol{F}_{pra_n}$, are the finalized representatives of typical semantic and spatial scales adopted for multi-level learning by Softmax classifier in the proposed architecture.

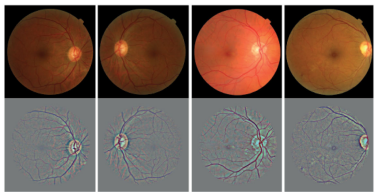
(c) Multi-level fusion: To this end, we feed each $\boldsymbol{F}_{pra_n}$ into the GAP layer followed by channel-wise concatenation for gaining the mixture of multi-level context, which smoothly carries finely-patterned and semantically-rich features of DR-related factors. Such procedure is given as follows

$$\boldsymbol{F}_{dr} = \mathcal{C}\left[\mathcal{G}(\boldsymbol{F}_{pra_1}), \mathcal{G}(\boldsymbol{F}_{pra_2}), \mathcal{G}(\boldsymbol{F}_{pra_3}), \mathcal{G}(\boldsymbol{F}_{pra_4})\right] \tag{5}$$

where $\boldsymbol{F}_{dr}$ stands for the final features handled by the subsequent Softmax classifier. Eventually, the severity grading performance



(a) Top to bottom: raw wide-field SS-OCTA, raw and preprocessed fundus images. Left to right: no DR, mild NPDR, moderate NPDR, severe NPDR, and PDR grades.



(b) Top to bottom: raw and preprocessed fundus images. Left to right: two referable DR and two non-referable DR grades.

Fig. 3. Example samples of: (a) internal KHUMC dataset and (b) Messidor dataset [34].

can be improved since DR-oriented clinical signs in various spatial scales are involved exhaustively and unambiguously thanks to the RCA stream.

IV. Experiments

A. Benchmark Datasets and Metrics

1) Internal KHUMC Dataset: This dataset is privately acquired from Kyung Hee University Medical Center, Seoul, Republic of Korea. There are totally 297 pairs of color fundus images (with resolution of 3608 × 3608) and wide-field SS-OCTA (1024 × 1024) involved in this research. Notably, the ground-truth classes of DR severity scales are manually graded and finalized by four experienced ophthalmologists in the Department of Ophthalmology. As a consequence, there are 65, 48, 64, 84, and 36 pairs of fundus and SS-OCTA 12 × 12 mm^2 images corresponding to the grades of no DR, mild NPDR, moderate NPDR, severe NPDR, and PDR, respectively. Fig. 3(a) presents several examples of wide-field SS-OCTA, fundus photographs of inconsistent appearances and corresponding preprocessed versions from this internal dataset. Obviously, training a DL model for remarkable DR screening performance with such kind of small-scale dataset is a challenging matter. Therefore, this difficulty motivates the proposal of our TFA-Net, which learns from multi-modal images, as a promising solution in this paper.

Following the existing DR-oriented studies [26], [37], we adopt QWK as the main metric assessing the performance of the proposed model with regard to different settings on this dataset. According to Scikit-learn framework [38], QWK is the measure representing the agreement degree of classification results between two raters, i.e., the group of grading experts and the prediction of our deep network in this scenario. The corresponding formulation is defined as follows

$$QWK = 1 - \left(\sum_{x=1}^{L}\sum_{y=1}^{L}\mathcal{W}_{x,y}\mathcal{O}_{x,y}\right) \Big/ \left(\sum_{x=1}^{L}\sum_{y=1}^{L}\mathcal{W}_{x,y}\mathcal{E}_{x,y}\right) \tag{6}$$

where L denotes the number of DR severity labels; $\mathcal{W}$ represents the weighting matrix that shows penalty of difference between prediction and corresponding ground-truth labels; $\mathcal{O}$ refers to as the observed confusion matrix computed from classifier's results; and $\mathcal{E}$ stands for the expected matrix inferred by the outer product between the L-length ground-truth and prediction vectors, which carry the occurrences of actual and predicted labels counted from all the test data, respectively. The final result of higher value indicates stronger agreement while the zero or negative rate demonstrates that the agreement is random.

2) Messidor Dataset [34]: This public dataset comprises 1200 fundus images of various resolutions graded in four scales by three domain experts. The quantitative distribution corresponding to these categories is 546, 153, 247, and 254, respectively. However, following the existing work [18], [21], [23], the data of those severity levels are further grouped into two classes only, i.e., the referable (representing those of the first two lower grades) and the non-referable DR (amounting to those of the latter two). Note that some sample images of this dataset with their preprocessed counterparts are exhibited in Fig. 3(b). Since this becomes a traditional binary classification problem, Accuracy, AUROC, and QWK metrics are involved for assessing the effectiveness of a learning model. As mentioned before, we design an ablation study on such relatively small-scale dataset to benchmark the robustness of the RCA stream regarding DR severity recognition in Section IV-C3.

B. Implementation Details

This article applies Pytorch [39] to train and quantify the proposed TFA-Net on one NVIDIA 1080TI GPU. Regarding the model's input resolution, all the images in use are re-scaled to different sizes such as 224 × 224, 448 × 448, or 600 × 600 for relevant ablation studies described at next subsections. In addition, since the raw wide-field SS-OCTA is grayscale while the fundus photograph being of RGB format, unless otherwise specified, the former is replicated itself by three times to form into a three-channel image like the latter as a default configuration that we mentioned before. About the preprocessing step applied into the fundus images, the local mean intensity subtraction method [37] is utilized for combating the large variety of brightness and contrast, which is caused by the inconsistency of nonmydriatic cameras. Besides that, we set the mini-batch size at 60 pairs of fundus images and wide-field SS-OCTA for the internal KHUMC dataset and 120 fundus images for the Messidor dataset.

Remarkably, in addition to the feature-wise augmentation in the TFA scheme of our model, conventional data-level augmentation and learnable regularization scheme are of great importance to elevate the generalization capability. Hence, we involve various manipulations onto the raw inputs such as random

Gaussian noise insertion, vertical and horizontal flipping, arbitrary rotation, and channel-wise normalization. Concurrently, the weight decay term with coefficient of 0.0001 is attached to the cross-entropy objective function.

As for the initialization of our TFA-Net's parameters, we apply He's method [40] to those of the learnable layers in the TFA scheme (for all experiments) as well as in the backbone CNN (for several ablation studies only). Otherwise, as a default setting, the parameters belonging to the backbone network are those from the corresponding ResNet [8] pretrained with ImageNet [41]. Then, the gradient descent algorithm having learning rate initialized at 0.001 with momentum fixed at 0.9 is applied to optimize the model's parameters. Note that a learning rate decay schedule in 'poly' style [42] is also adopted. Finally, a 5-fold cross validation strategy is deployed to perform the evaluation procedure. Note that training period in each validation fold is 400 epochs and the final reported QWK includes mean and standard deviation values across five folds. Meanwhile, we take a 10-fold cross validation protocol as in [18], [21], [23], where the training process lasts 50 epochs at each fold, for the evaluation on Messidor dataset.

C. Experimental Results and Discussions

1) Coordination Between Fundus Image and Wide-Field SS-OCTA for DR Severity Labeling Given Small-Scale Supervised Dataset: The experiments in this part are conducted using the internal KHUMC dataset, wherein all of the considered input images are resized to 448 × 448. Besides that, the chosen backbone CNN is ResNet-18 [8] pretrained with ImageNet dataset [41]. By such baseline settings, we then train the proposed network with different configurations using three separated groups of inputs, i.e., fundus images only, wide-field SS-OCTA only, and the amalgamation between those two modalities, respectively. At each evaluation phase, manifold settings of enabled/disabled Stage 1 (Weight-sharing convolution kernels followed by channel-wise concatenation) and Stage 2 (RCA stream) are in-turn performed as shown in Table I. Notably, because the Stage 1 of TFA mechanism is not applicable to single-modal learning cases, we follow the conventional layer setting wherein 64 initial convolution kernels applied to the input images have size of 7 × 7 × 3, stride of 2, and padding of 3 for producing half-spatial-sized feature maps to be fed into the first residual block in the backbone network. Regarding the scenario only using wide-field SS-OCTA for learning the DR grade classification model, the ground-truth labels are determined through screening the fundus counterparts. Although this procedure is obviously meaningless in terms of conventional DR-oriented study, we still involve it to further express that despite the restriction of dataset size, DL technique remains efficient if its architecture is well-regularized at feature space and trained with multi-modal images.

Compared to the baseline model of single-modal learning, the extra attachment of the proposed RCA stream improves the QWK rates by 2.1% (for the case of only using fundus inputs) and 2.4% (wide-field SS-OCTA). Notably, the corresponding trade-off is the increase by 2.7% of trainable parameters' amount

TABLE I

QWK (%) ON THE INTERNAL KHUMC DATASET WITH DIFFERENT INPUT TYPES (WITH SAME SIZE OF 448 × 448) AND STRATEGIES OF TWOFOLD FEATURE AUGMENTATION GIVEN THE RESNET-18 [8] AS BACKBONE CNN. $\mathbf{W}_{f64} \in \mathrm{R}^{7\times7\times3}$ (APPLIED TO INPUT FUNDUS IMAGES ONLY) AND $\mathbf{W}_{o64} \in \mathrm{R}^{7\times7\times1}$ (APPLIED TO WIDE-FIELD SS-OCTA) DENOTE TRAINABLE WEIGHTS OF 64 CONVOLUTION KERNELS HAVING STRIDE OF 2 AND PADDING OF 3. $\mathbf{W}_{f32} \in \mathrm{R}^{7\times7\times3}$, $\mathbf{W}_{o32} \in \mathrm{R}^{7\times7\times1}$, AND $\mathbf{W}_{shared} \in \mathrm{R}^{7\times7\times3}$ (APPLIED TO BOTH INPUT MODALITIES) STAND FOR 32 KERNELS WITH OTHER CONFIGURATIONS SAME AS ABOVE. $\mathcal{C}[.]$ INDICATES THE DEPTH-WISE CONCATENATION USED IN MULTI-MODAL LEARNING CASES. '✓' SIGNIFIES THE UTILIZATION OF STAGE 2, OTHERWISE THE SOFTMAX CLASSIFIER DIRECTLY PROCESSES THE GLOBAL-POOLED FEATURES OF THE 4^{th} CONVOLUTION BLOCK'S OUTPUTS AS IN THE ORIGINAL RESNET-18 [8]

Inputs	Strategy		No. params	QWK (%)
	Stage 1	Stage 2		
Preprocessed Fundus	$\mathbf{W}_{f64}$		11.18M	82.0±1.7
	$\mathbf{W}_{f64}$	✓	11.48M	84.1±1.5
Wide-field SS-OCTA	$\mathbf{W}_{o64}$		11.18M	81.1±1.7
	$\mathbf{W}_{o64}$	✓	11.48M	83.5±2.0
Preprocessed Fundus + Wide-field SS-OCTA	$\mathbf{W}_{f32}$+$\mathbf{W}_{o32}$+$\mathcal{C}[.]$		11.18M	80.4±2.1
	$\mathbf{W}_{shared}$+$\mathcal{C}[.]$		11.18M	86.7±2.1
	$\mathbf{W}_{f32}$+$\mathbf{W}_{o32}$+$\mathcal{C}[.]$	✓	11.48M	88.7±2.8
	$\mathbf{W}_{shared}$+$\mathcal{C}[.]$	✓	11.48M	90.2±2.4

for backbone ResNet-18, which is insignificant. About multi-modal utilization, naively combining the fundus images and wide-field SS-OCTA as inputs of the baseline network, i.e., each modality is learned by separated convolutional layers prior concatenation, even deteriorates the QWK-based performance (with only 80.4%, the lowest rate in Table I). The primary cause for such bottleneck is basically the severe lack of data, due to which the conventional DL network struggles to efficiently coordinate the contextual mixture of those two input types. Therefore, the involvement of the proposed TFA mechanism is vastly important for leveraging the generalization capability of the whole model given the constraint of very small-scale dataset. Particularly, the results in Table I show that individually using the weight-sharing convolution kernels (Stage 1) and the RCA stream (Stage 2) can boost the QWK values by 6.3% and 8.3%, respectively, which are a great deal of performance advancement. Furthermore, simultaneously applying the two stages of feature-level regularization strategy nails the highest QWK rate at 90.2%.

Those quantitative results extensively deliver the following two major remarks. Firstly, the weight-sharing kernels enable the network to earlier orchestrate informative hybrid features of the two different modalities, which should facilitate the learning process at later layers more optimally. Meanwhile, the exploitation of mutual context when using separated filters only takes place after depth-wise concatenation layer, which may easily lead to sub-optimal state. Secondly, the QWK's improvement rate of the RCA stream for the multi-modal inputs strongly dominates the single-modal cases (8.3% vs. 2.1% and 2.4%). This outcome indicates that the cross-modal representations of the two input types perform a critical role in strengthening

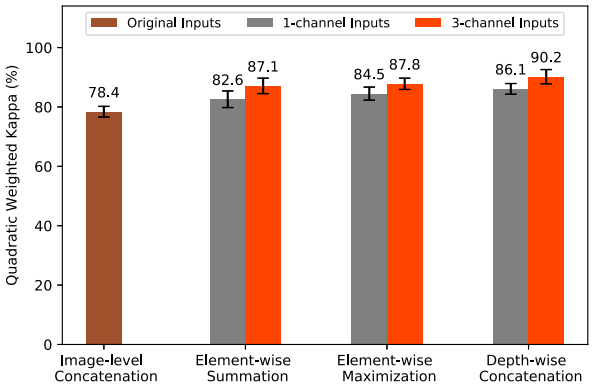


Fig. 4. QWK (%) on the internal KHUMC dataset with different input combination styles in Stage 1 of the TFA-Net: channel-wise concatenation at image level; 1-channel (grayscale) or 3-channel (RGB) inputs followed by various feature-level fusion manners (element-wise summation, element-wise maximization, and depth-wise concatenation).

the augmentation ability of the RCA stream, especially under the scenario of learning DR-oriented biomarkers from such a small-scale dataset.

2) Correlation Between Modalities' Format and Feature-Level Combination Styles for Cross-Modal Learning in Stage 1 of the TFA Mechanism: Since the original channel dimension between the two input types is different, wherein the fundus modality is RGB (3-channel) image while the wide-field SS-OCTA is grayscale (1-channel), we deploy three separated approaches as follows. Firstly, the two input modalities are directly concatenated along the channel dimension. Secondly, 64 weight-sharing convolution kernels $\mathbf{W}_{shared} \in \mathrm{R}^{7\times7\times1}$ are applied to the grayscale version of the preprocessed fundus input and the original wide-field SS-OCTA. Thirdly, those of $\mathbf{W}_{shared} \in \mathrm{R}^{7\times7\times3}$ are adopted to extract features from the preprocessed fundus image and the 3-channel form of the raw wide-field SS-OCTA (generated from channel-wise replication as described previously). Then, besides the feature-level combination scheme of depth-wise concatenation in Stage 1 of the proposed TFA-Net, we further experiment with element-wise summation [8] and maximization [43] to examine the impact of these mixture styles on the subsequent cross-modal learning process. It is noteworthy that the input resolution is fixed at 448 $\times$ 448 and the rest of the architecture is a backbone of pretrained ResNet-18 [8] connected with the RCA stream (Stage 2).

As illustrated in Fig. 4, the image-level concatenation approach yields the lowest QWK value since each input modality has its own DR-oriented biomarkers, from which the direct fusion may introduce significant ambiguities to subsequent learning stages. Meanwhile, QWK rates of utilizing 3-channel inputs outperform those of using 1-channel counterparts by 3.3-4.5% given the above-mentioned combination techniques at feature space. These results arguably indicate that the DR-related biomarkers encoded in the original 24-bit format of preprocessed fundus modality are more profitable to the model learning process than those attenuated in the 8-bit image representation. Moreover, it can be realized that the extra operation of mirroring the raw wide-field SS-OCTA along the depth dimension does not cause any adverse effects. Particularly, since the feature extraction basis of the convolution kernels is linear combination within the receptive field, applying those filters to the original (1-channel) image and the 3-channel replicated version shall produce corresponding output features having linear scaling relationships. For instance, let us suppose that all weights of the convolution kernels in use are same, output features' intensities inferred from the 3-channel input are equal to those acquired from the 1-channel counterpart multiplied by a scale factor of 3.

Regarding the feature-level combination methods, the depth-wise concatenation strategy delivers superior performance over the remaining schemes by 1.6–3.5% (for 1-channel inputs) and 2.4–3.1% (for 3-channel inputs). These experimental results can be explained by the following two aspects. Firstly, although the two input image types share a similar FOV, their structural details are not per-pixel aligned. Hence, element-wise merging approaches such as summation or maximization may bring about unexpected biases of information propagation in next layers of the network. Secondly, the channel-wise concatenation, which regulates corresponding regions of interest between the two modalities placed along the depth dimension, enables the subsequent convolutional layers to flexibly manipulate the correlations of cross-modal features based on the predefined spatial extent. In other words, this procedure allows the unrestricted exploitation of hybrid features at region level, rather than a fixed combination at pixel level beforehand like the two compared schemes.

3) Robustness of RCA Stream Compared to the State-of-The-Arts: Subsection IV-C1 demonstrated that the engagement of the RCA stream with a base CNN introduces more robust feature representation for higher diagnosis performance. Furthermore, for the purpose of proving this mechanism's novelty in the literature, we use the Messidor dataset [34], a popular DR screening-related fundus dataset having relatively small size, to benchmark and compare the TFA-Net (without Stage 1) with the state-of-the-arts [18], [19], [21]–[23].

Accordingly, the comparison in terms of Accuracy and AUROC along with the total number of trainable parameters are exhibited in Table II. Note that we additionally include the standard deviation and QWK values (which are linearly proportionate to the Accuracy rates in such binary classification problem) in our results. In general, with the lowest-capacity version (i.e., ResNet18-RCA), the proposed model outperforms the compared ones by a large margin for both Accuracy (1.6–6.7%) and AUROC (1.3–11.3%). Moreover, as the total number of layers increases up to 101, the performance gain of those two benchmark metrics is approximately 2% and 1.6% higher, respectively, which are significant in such recognition-related area. Additionally, compared to the baseline counterparts, those attached with the RCA stream achieve remarkable performance gaps of higher Accuracy (5.9–7.5%) and AUROC (10.4–11.2%). These attainments suggest that the proposed RCA stream, by utilizing higher-level semantic details in coarser patterns to reversely refine lower-level ones in finer scales based on depth-wise context, is not only flexible to different backbone CNNs but also robust against overfitting issue on such a relatively small-sized dataset.

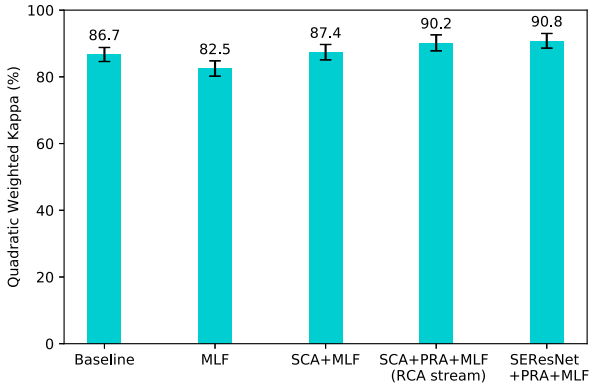
TABLE II
PERFORMANCE COMPARISON ON MESSIDOR DATASET [34]. NOTE THAT THE RESULTS OF OUR MODELS ARE ACHIEVED WITH PREPROCESSED INPUTS. THERE ARE FIXED 1080 TRAINING AND 120 TESTING IMAGES AT EACH ROUND OF THE 10-FOLD CROSS VALIDATION

Approach	Input size	No. params	Accuracy (%)	AUROC (%)	QWK (%)
Inception-v3-DR [19]†	299×299	21.8M	86.1±3.2	86.5±1.3	73.2±3.6
VNXK-Net [21]	449×449	507.4M	89.3	88.7	
CKML-Net [21]	451×451	71.5M	89.7	89.1	
Adaptated-VGGs [22]†	512×512	268.6M	90.7±2.4	95.4±0.9	81.4±5.1
Zoom-in-Net [23]	492×492	55.8M	91.1	95.7	
SI2DRNet-v1 [18]	672×672	10.6M	91.2	96.5	
ResNet-18 [8]‡	448×448	11.18M	85.3±2.3	86.6±1.3	69.6±5.5
ResNet-50 [8]‡		23.51M	87.8±3.5	88.3±1.4	76.6±3.0
ResNet-101 [8]‡		42.51M	87.6±3.8	88.3±0.9	75.2±2.3
ResNet18-RCA	448×448	11.48M	92.8±3.1	97.8±1.8	84.6±3.9
ResNet50-RCA		28.34M	93.7±2.9	98.7±1.2	88.4±4.2
ResNet101-RCA		47.34M	94.8±3.6	99.4±1.1	89.6±2.7

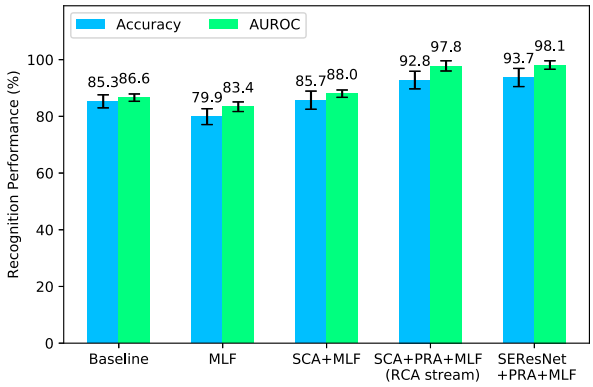
†: Experimental results are achieved by our own implementation following default configurations in the original work.
‡: Experimental results are achieved by finetuning the pretrained baseline networks.

Regarding the computational complexity, all the variants of the proposed network have much fewer parameters compared to ensemble-oriented approaches, which simultaneously involve either different styles of fundus photos [21] or multiple feature learning streams [22], [23]. About approaches adopting conventional structure of CNN like 20-layer SI2DRNet-v1 [18] and Inception-v3 [7] finetuned in [19], although the corresponding models are more lightweight, their input's resolution should be large enough so that spatially-rich representations of DR-related biomarkers are encoded more effectively throughout multiple downsampling layers. Clearly, SI2DRNet-v1 [18] (with large input size of 672 × 672) gives superior Accuracy and AUROC over the Inception-v3-DR [19] (299 × 299) as reported in Table II. On the contrary, our network still grants superior performance given much lower-resolution input images despite slightly trading-off 8% more parameters (in the case of using backbone ResNet-18).

4) Interactions Between SCA, PRA, and MLF in the RCA Stream: The effectiveness of Stage 2 (i.e., RCA stream) has been shown in the earlier ablation studies on both internal KHUMC and Messidor [34] datasets given bimodal and single-modal input images, respectively. To this end, we further evaluate how the constituent components within the RCA stream, i.e., SCA, PRA, and MLF, interact with each other to leverage the performance of DR grade recognition. Note that we continue to apply the baseline setting with input size of 448 × 448 and backbone network of pretrained ResNet-18 [8]. Fundamentally, as manifested in Fig. 5, there are five deployed strategies: (i) baseline ResNet-18 finetuned from pretrained parameters; (ii) ResNet-18 with chosen feature maps of four different scales directly fed into the MLF component; (iii) ResNet-18 with those feature maps re-calibrated using SCA followed by point-wise multiplication before the MLF component; (iv) the proposed TFA-Net (ResNet-18 coupled with the RCA stream comprising SCA, PRA, and MLF); and (v) SEResNet-18 [9] (wherein SCA blocks are embedded along the feedforward pass by default) connected with the PRA and MLF modules. Remarkably, the MLF component is essential when employing SCA and/or PRA so that all the multi-level features of interest are involved in the



(a) Internal KHUMC dataset (inputs: preprocessed fundus images & wide-field SS-OCTA).



(b) Messidor dataset [34] (inputs: preprocessed fundus images).

Fig. 5. Recognition performance on the internal KHUMC and Messidor [34] datasets with following strategies: ResNet-18 [8] with baseline structure (having 11.18 M parameters), with only MLF (11.18 M), with SCA + MLF (11.23 M), with SCA + PRA + MLF (i.e., RCA stream having 11.48 M parameters), and SEResNet-18 [9] with PRA + MLF (11.57 M).

Softmax classifier of the proposed network. As a consequence, the two diagrams of experimental results on both benchmark datasets in Fig. 5 present a similar performance paradigm as follows.

In comparison with the baseline, the involvement of only MLF component to the backbone ResNet-18 generally decreases the recognition rates in terms of QWK (by 4.2%) for the internal KHUMC dataset as well as Accuracy (5.4%) and AUROC (3.2%) for the Messidor [34]. It is apparent that adopting directly the low-level features, which carry spatially rich details but semantically poor context, to the Softmax classifier induces even more ambiguities in the final decision.

Meanwhile, when utilizing the SCA modules, of which the output self-context vectors are used to channel-wisely refine their input feature maps via point-wise multiplication, preceding the MLF scheme, the classification results are slightly improved over those of the baseline in both benchmark cases. Particularly, QWK value rises by 0.7% for the bimodal inputs, while Accuracy and AUROC gain increases of 0.4% and 1.4%, respectively, for the single-modal ones. We argue that the re-calibration procedure on each feature map of interest can facilitate itself to express

TABLE III
QWK (%) ON THE INTERNAL KHUMC DATASET WITH DIFFERENT SETTINGS OF BACKBONE NETWORK CAPACITIES, INPUT RESOLUTIONS, IMAGENET-PRETRAINED PARAMETERS INITIALIZATION, AND FUNDUS-RELATED PREPROCESSING

Backbone capacity			Input resolution			ImageNet pretrained	Fundus preprocessing	QWK (%)
ResNet-18	ResNet-50	ResNet-101	224×224	448×448	600×600			
✓			✓			✓	✓	83.2±1.6
	✓		✓			✓	✓	90.0±2.1
		✓	✓			✓	✓	88.5±3.0
✓				✓		✓	✓	90.2±2.4
✓					✓	✓	✓	90.3±1.7
✓			✓				✓	74.1±2.2
✓			✓			✓		78.1±4.1
✓			✓					69.5±3.1

informative details and vice versa, which then reduces unexpected uncertainties caused by the original low-level features. However, the benefits of only utilizing such intra-contextual dependencies remain trivial.

Accordingly, by further engaging the feedback-like strategy of inter-context exploitation introduced by the PRA component, such models outperform the baseline by a significant margin. In specific, while QWK rate achieves an improvement of 3.5% on the internal KHUMC dataset, the Accuracy and AUROC surge by 7.5% and 11.2% on the Messidor [34], respectively. These results demonstrate the advantage of combining semantically-rich context with fine-grained details, which leverages the usefulness of the concerned multi-scale features for the final prediction as argued previously. Furthermore, the strategy using SEResNet-18 [9] combined with the PRA and MLF modules can improve the evaluation metrics marginally (by 0.3–0.9%) on both benchmark datasets thanks to a more powerful feature extraction process (but trading-off more parameters for manifold embedded SCA blocks).

As for model complexity aspect, the increment of trainable parameters due to the proposed RCA stream is insignificant. Particularly, provided that the MLF is not a learnable component, in comparison with the baseline ResNet-18 having 11.18 M parameters, involving SCA only or SCA plus PRA further raises that quantity by 0.45% (11.23 M) or 2.68% (11.48 M), respectively. These statistics specify the notable efficiency of depth-wisely transforming self-context information in conjunction with the reverse mixture tactic.

5) Sensitivity of TFA-Net to Predefined Hyperparameters: Earlier experiments unveiled the advantages of feeding multi-modal inputs into our well-generalized TFA-Net subject to the challenge of learning from small-scale dataset. The comparison results in Table II also show that backbone network's capacity has an obvious impact on the final performance for the case of single-modal (i.e., fundus photos only) learning. Therefore, we additionally carry out another ablation study about the sensitivity of DR-oriented multi-modal learning with respect to various backbone CNN's capacities, input resolutions, model's parameters initialization manners, and the utilization of fundus image preprocessing technique on the domestic KHUMC dataset. Accordingly, all the related measurements of QWK are presented in Table III.

It can be realized that the behaviors of TFA-Net regarding the number of layers in backbone CNN are slightly different from those discussed in Section IV-C3. Firstly, the QWK values inferred by the deeper backbone networks introduce a big gap of more than 5.3% compared to those of the 18-layer version. Secondly, using ResNet-50 as the base feature extractor surprisingly offers higher QWK rate (approximately 1.5%) than employing the 101-layer counterpart on our internal dataset, as opposed to the experiment with the Messidor [34]. Consequently, the former observation points out that the proposed architecture with feature-level augmentation scheme enables a more proficient utilization of larger-capacity backbone CNNs. Meanwhile, the latter suggests that there exists a peak of the correlation between network capacity's increment and training data volume, over which the performance degradation arises despite the strength of the applied regularization scheme. This accordingly explains for the event that adopting ResNet-50 as backbone extractor yields better classification performance than ResNet-101 for the smaller-scale dataset like ours, and vice versa for the Messidor benchmark [34] with larger image quantity.

As for the dependence of grading performance on input resolution, there is also an analogous orientation to the above capacity factor. In specific, the QWK rates resulted from both spatial sizes of 448×448 and 600×600 are similar while higher than that of 224×224 by around 7%. These outcomes imply that richer spatial details allow the proposed model to represent DR-related manifestations more comprehensively. However, there is also a saturation scale, over which the larger values of input size do not considerably improve the QWK rate while inducing more computations in the model.

Finally, the remaining measures report that the manners of parameters initialization and fundus image preprocessing affect the QWK-based performance as well. Among the corresponding four test cases, those with backbone CNN originated by parameters pretrained on ImageNet [41] hold the first and second ranks of QWK values (with an average higher rate of 8.85% compared to the random initialization settings). Arguably, despite the difference of data domain, ImageNet-pretrained parameters are still able to reasonably characterize general features of interest (e.g., the important edge-based appearance of retinal vessels) from the beginning of the training process. This activity should be more helpful than considering the random initialization plan.

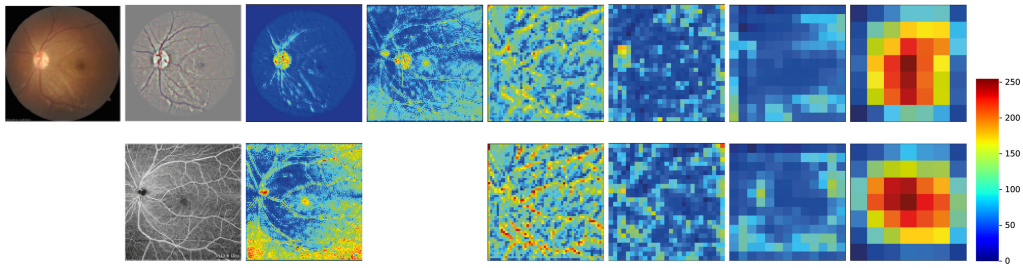


Fig. 6. Illustrations of typical feature maps throughout the proposed architecture. First row (left to right): raw fundus input, preprocessed fundus version, fundus-based output of weight-sharing convolution kernels $\mathbf{W}_{shared}$, F_c (cross-modal output of Stage 1 of the TFA mechanism), F_1 (final output of the first convolution block in the backbone CNN), F_2, F_3, and F_4. Second row (left to right): raw wide-field SS-OCTA input, OCTA-based output of $\mathbf{W}_{shared}$, F_{pra_1} (output of the PRA component, which corresponds to the input F_1 of Stage 2), F_{pra_2}, F_{pra_3}, and F_{pra_4}. View in color is highly recommended.

Moreover, involving the preprocessed fundus inputs instead of the raw formats can leverage the final QWK metric by 4.85% averagely. As a result, the alignment of these two settings profits the final performance in terms of QWK rate greatly.

6) Illustrations of Typical Feature Maps Throughout the TFA-Net: To this end, as illustrated in Fig. 6, we deliver appearances of several pivotal features along the proposed architecture comprising two stages of our TFA mechanism with a backbone CNN in the middle. Let a pair of fundus image and wide-field SS-OCTA be fed into the trained model, for each of the inspected feature maps, the pixel values are averaged over the depth dimension, of which the obtained results are then calibrated to the intensities ranging from 0 to 255 [33]. Note that the regions highly corresponding to strong features are roughly expressed by higher intensity values.

Initially, low-level features of both input modalities, depicted in the third column of Fig. 6, respectively, are acquired by the weight-sharing convolution kernels $\mathbf{W}_{shared}$. Next, they are concatenated in channel-wise manner, of which the cross-modal feature map F_c possessing the combined expressiveness (as shown in the fourth column of Fig. 6) is the final output of Stage 1 in the TFA scheme. Afterwards, a backbone CNN (i.e., ResNet-18 [8]) is engaged to transform that concatenated feature at manifold semantic levels. Notably, spatial resolutions of the extracted feature maps get coarser while the corresponding semantic contexts are enriched more intensively along the feedforward pass of the network. Then, ultimate outputs with different scales of the four constituent convolution blocks in ResNet-18 [8] (as demonstrated by the last four heat maps in the first row of Fig. 6), i.e., F_n where $n = 1, 2, 3, 4$, are further involved in the TFA's second stage (i.e., RCA stream). Subsequently, the corresponding output features F_{pra_n}, which are managed by the regime of SCA followed by PRA components, are manifested in the last four columns of the second row in Fig. 6. Since fine-grained details of lower-level features are refined by the reverse integration of semantically-richer information from higher-level versions, the visual disparities between those original and corresponding manipulated representations are obvious. Finally, the recalibrated multi-level features continue to pass through the MLF component followed by the Softmax classifier for DR severity grade recognition.

V. Limitations and Future Work

Despite achieving promising performance on DR severity recognition using two different input modalities, there still exists noticeable limitations as follows. Firstly, both the benchmark datasets contain samples of a narrow population and the manually grading process may suffer from subjective biases. Secondly, although several feature visualization tools are available, there are still challenges in transparently interpreting the DR-oriented signs during the feedforward process inside a DL architecture. This subsequently raises concerns of the ophthalmologists in clinical practice about following up the exact risk factors exacerbating in the considered images.

In the future, we target to efficiently involve more image types related to DR domain such as Fluorescein Angiography and/or narrow-field SS-OCTA (3×3 mm^2), which have better representations of foveal avascular zone and vessel density. Moreover, not only the depth-wise attention scheme is reversely exploited as in this work, but also spatially-attentional mechanism incorporated forwardly shall be studied for improving the DR grade classification performance.

VI. Conclusion

In this study, a Convolutional Network with Twofold Feature Augmentation, called TFA-Net, is introduced for DR severity grading given a very small-scale dataset of fundus and wide-field SS-OCTA modalities. Specifically, the proposed model comprises a backbone CNN extracting deep features of various representational scales and an attached feature-wise augmentation scheme operating in two stages. The first one is the employment of weight-sharing convolution kernels for coordinating cross-modal characteristics in a generalized manner. The second is the utilization of the RCA stream, which facilitates the integration of semantically-rich details (higher-level features) to finely-patterned appearances (lower-level counterparts) in feedback-like procedure. Such nonlinear transformations allow features refined at multiple scales to engage with the Softmax classifier more comprehensively. Consequently, the incorporation between data- and feature-level augmentation mechanism in the proposed network attains impressive recognition performance on the small-sized internal KHUMC and Messidor

datasets. Although both of the benchmarks do not exhibit diverse population, we show that applying multi-modal inputs of the same instance to a CNN architecture, which is strongly regularized at feature levels, can be a considerable alternative mitigating the massive dependence on big labeled data.

REFERENCES

[1] C.-H. Hua *et al.*, "Bimodal learning via trilogy of skip-connection deep networks for diabetic retinopathy risk progression identification," *Int. J. Med. Inform.*, vol. 132, Dec. 2019, Art. no. 103926.

[2] K. Kim, E. S. Kim, and S.-Y. Yu, "Longitudinal relationship between retinal diabetic neurodegeneration and progression of diabetic retinopathy in patients with type 2 diabetes," *Amer. J. Ophthalmol.*, vol. 196, pp. 165–172, 2018.

[3] S. Vujosevic *et al.*, "Early microvascular and neural changes in patients with type 1 and type 2 diabetes mellitus without clinical signs of diabetic retinopathy," *RETINA*, vol. 39, no. 3, pp. 435–445, Mar. 2019.

[4] Y. He *et al.*, "Segmenting diabetic retinopathy lesions in multispectral images using low-dimensional spatial-spectral matrix representation," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 2, pp. 493–502, Feb. 2020.

[5] "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales," *Ophthalmology*, vol. 110, no. 9, pp. 1677–1682, 2003.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.

[7] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[9] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[10] T. Huynh-The, C.-H. Hua, and D. Kim, "Encoding pose features to images with data augmentation for 3-d action recognition," *IEEE Trans. Ind. Inform.*, vol. 16, no. 5, pp. 3100–3111, May 2020.

[11] T. Huynh-The, C.-H. Hua, T.-T. Ngo, and D.-S. Kim, "Image representation of pose-transition feature for 3d skeleton-based action recognition," *Inf. Sci.*, vol. 513, pp. 112–126, 2020.

[12] C. Cao *et al.*, "Deep learning and its applications in biomedicine," *Genomic., Proteomic. Bioinf.*, vol. 16, no. 1, pp. 17–32, 2018.

[13] M. Wainberg, D. Merico, A. Delong, and B. J. Frey, "Deep learning in biomedicine," *Nature Biotechnol.*, vol. 36, pp. 829 EP –, 09 2018.

[14] C.-H. Hua, T. Huynh-The, and S. Lee, "Retinal vessel segmentation using round-wise features aggregation on bracket-shaped convolutional neural networks," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Jul. 2019, pp. 36–39.

[15] M. C. A. Trivino, J. Despraz, J. A. L. Sotelo, and C. A. Peña, "Deep learning on retina images as screening tool for diagnostic decision support," *CoRR*, vol. abs/1807.09232, 2018.

[16] J. de la Torre, A. Valls, and D. Puig, "A deep learning interpretable classifier for diabetic retinopathy disease grading," *Neurocomputing*, vol. 396, pp. 465–476, 2020.

[17] S. M. S. Islam, M. M. Hasan, and S. Abdullah, "Deep learning based early detection and grading of diabetic retinopathy using retinal fundus images," *CoRR*, 2018. [Online]. Available: http://arxiv.org/abs/1812.10595

[18] Y. Chen, T. Wu, W. Wong, and C. Lee, "Diabetic retinopathy detection based on deep convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2018, pp. 1030–1034.

[19] V. Gulshan *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 12 2016.

[20] J. Sahlsten *et al.*, "Deep learning fundus image analysis for diabetic retinopathy and macular edema grading," *Sci. Rep.*, vol. 9, no. 1, p. 10750, 2019.

[21] H. H. Vo and A. Verma, "New deep neural nets for fine-grained diabetic retinopathy recognition on hybrid color space," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2016, pp. 209–215.

[22] D. S. 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," *JAMA*, vol. 318, no. 22, pp. 2211–2223, Dec. 2017.

[23] Z. Wang, Y. Yin, J. Shi, W. Fang, H. Li, and X. Wang, "Zoom-in-net: Deep mining lesions for diabetic retinopathy detection," in *Med. Image Comput. Comput. Assist. Interv.*, 2017, pp. 267–275.

[24] L. Zhou, Y. Zhao, J. Yang, Q. Yu, and X. Xu, "Deep multiple instance learning for automatic detection of diabetic retinopathy in retinal images," *IET Image Process.*, vol. 12, no. 4, pp. 563–571, 2018.

[25] P. Junjun, Y. Zhifan, S. Dong, and Q. Hong, "Diabetic retinopathy detection based on deep convolutional neural networks for localization of discriminative regions," in *Proc. Int. Conf. Virtual Reality Visual.*, Oct. 2018, pp. 46–52.

[26] "Kaggle: Diabetic retinopathy detection," https://www.kaggle.com/c/diabetic-retinopathy-detection

[27] T. S. Hwang *et al.*, "Automated quantification of capillary nonperfusion using optical coherence tomography angiography in diabetic retinopathy," *JAMA Ophthalmol.*, vol. 134, no. 4, pp. 367–373, 2016.

[28] A. Alibhai *et al.*, "Quantification of retinal capillary nonperfusion in diabetics using wide-field optical coherence tomography angiography," *Retina*, p. 1, 2018.

[29] F. Wang, S. S. Saraf, Q. Zhang, R. K. Wang, and K. A. Rezaei, "Ultra-widefield protocol enhances automated classification of diabetic retinopathy severity with oct angiography," *Ophthalmol. Retina*, vol. 4, no. 4, pp. 415–424, Apr. 2020.

[30] K. Kim, E. Kim, D. Kim, and S.-Y. Yu, "Progressive retinal neurodegeneration and microvascular change in diabetic retinopathy: Longitudinal study using oct angiography," *Acta Diabetologica*, vol. 56, no. 12, pp. 1275–1282, 2019.

[31] I. P. Okuwobi, Z. Ji, W. Fan, S. Yuan, L. Bekalo, and Q. Chen, "Automated quantification of hyperreflective foci in SD-OCT with diabetic retinopathy," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 4, pp. 1125–1136, Apr. 2020.

[32] Y. Huang *et al.*, "Swept-source OCT angiography of the retinal vasculature using intensity differentiation-based optical microangiography algorithms," *Ophthalmic Surg., Lasers Imag. Retina*, vol. 45, no. 5, pp. 382–389, 2014.

[33] C.-H. Hua, T. Huynh-The, S.-H. Bae, and S. Lee, "Cross-attentional bracket-shaped convolutional network for semantic image segmentation," *Inf. Sci.*, vol. 539, pp. 277–294, 2020.

[34] E. Decencire *et al.*, "Feedback on a publicly distributed database: The messidor database," *Image Anal. Stereol.*, vol. 33, no. 3, pp. 231–234, Aug. 2014.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[36] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, vol. 37, pp. 448–456.

[37] B. Graham, "Kaggle diabetic retinopathy detection competition report," 2015.

[38] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[39] A. Paszke *et al.*, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

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

[41] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014.

[42] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Comput. Vis.–ECCV 2018*. Cham: Springer International Publishing, 2018, pp. 833–851.

[43] Z. Liao and G. Carneiro, "A deep convolutional neural network module that promotes competition of multiple-size filters," *Pattern Recognit.*, vol. 71, pp. 94–105, 2017.