Biomedical Signal Processing and Control 63 (2021) 102167

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Multi-label ocular disease classification with a dense correlation deep neural network

Junjun He [a], Cheng Li [b], Jin Ye [c,d], Yu Qiao [c,d], Lixu Gu [a,*]

[a] School of Biomedical Engineering/the Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai 200240, China
[b] Paul C. Lauterbur Research Center for Biomedical Imaging, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[c] Shenzhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China
[d] SIAT Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen 518000, China

ARTICLE INFO

Keywords:
Ocular disease classification
Dense correlation network
Patient-level diagnosis
Multi-label annotation

ABSTRACT

Early diagnosis and timely treatment of ocular diseases are vital to prevent irreversible vision loss. Color fundus photography is an effective and economic tool for fundus screening. Since few symptoms are visible in the early disease stages, automatic and robust diagnosing algorithms according to color fundus photographs are in urgent need. Existing studies concentrate on image-level diagnoses treating the eyes independently without utilizing the useful correlation information between the left and right eyes. Besides, they commonly target only one or several ocular disease categories at a time. Considering the importance of both patient-level diagnosis correlating bilateral eyes and multi-label disease classification, we propose a patient-level multi-label ocular disease classification model based on convolutional neural networks. Specifically, a dense correlation network (DCNet) is designed to tackle the problem. DCNet consists of three major modules, a backbone CNN for feature extraction, a spatial correlation module for feature correlation, and a classifier for classification score generation. The backbone CNN extracts two sets of features from the left and right color fundus photographs, respectively. Subsequently, the spatial correlation module captures the pixel-wise correlations between the two feature sets. Then, the processed features are fused to get a patient-level representation. The final disease classification is conducted with the patient-level representation. Adopting a multi-label soft margin loss, the effectiveness of the proposed model is evaluated on a publicly available dataset, and the classification performance is improved with a large margin compared with multiple baseline methods.

1. Introduction

Retinal damage caused by persistent ocular diseases can lead to irreversible vision loss and even blindness [1–3]. Timely diagnosis is vital for effective treatment. To aid in detecting ocular diseases, different imaging techniques have been developed. Among them, optical coherence tomography (OCT) and color fundus photography (CFP) are widely employed [4]. OCT generates cross-section images of the retina, and retinal thickness can be measured to evaluate the eye conditions. CFP records the interior surfaces of the eyes to monitor possible disorders. Both tools have been proved to be effective for early-stage ocular disease diagnosis. Nevertheless, CFP is a more economical and efficient approach, and periodic fundus examination with CFP is recommended for asymptomatic adults, especially for the elder populations [5]. Unfortunately, common ocular diseases, such as diabetic retinopathy, cataract, age-related macular degeneration, etc. (Fig. 1a), progress with few initial visible symptoms, which makes it difficult to achieve accurate diagnoses in the early stages. Moreover, manual inspection of the generated large quantities of CPFs is laborious and time-consuming. In backward areas, there are far from enough radiologists available to perform the manual analysis. Automatic models are in urgent need not only to alleviate the pressure on ophthalmologists but also to improve the accuracy of imaging-based diagnosis.

Recently, deep neural networks (DNN), particularly convolutional neural networks (CNN), have made significant contributions to the medical imaging field [6–8]. In respect to ocular disease diagnosing, CNNs have shown promising performance in various aspects ranging from disease classification to object detection. A pixel-wise classification approach was used by Liefers et al. to detect the fovea centers in OCT images [9]. A two-stage CNN model was designed by Meng et al. to detect the optic disks in CFPs [10]. CNNs were adopted by Lee et al. to segment intraretinal fluid in OCT images [11]. Similarly,

* Corresponding author.
E-mail address: gulixu@sjtu.edu.cn (L. Gu).

https://doi.org/10.1016/j.bspc.2020.102167
Received 2 June 2020; Received in revised form 2 July 2020; Accepted 15 August 2020
Available online 25 August 2020

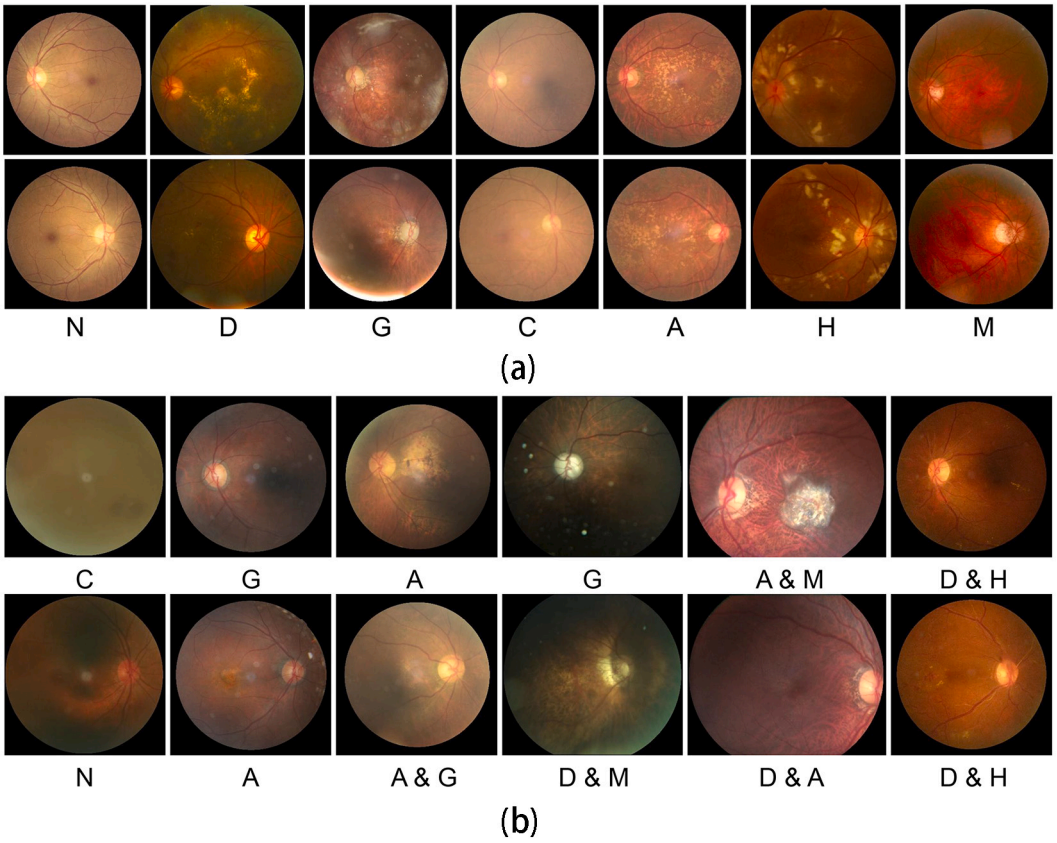

Fig. 1. (a) Example color fundus photographs (CFPs). (b) Complicated cases where the two eyes are infected with different ocular diseases or infected with more than one disease type. The ocular diseases infected by respective eyes are extracted from the provided diagnostic keywords, which cannot be directly transformed into the different classification labels. The first row presents the CFPs from the left eyes and the second row displays the corresponding CFPs from the right eyes. N, D, G, C, A, H and M refer to patients diagnosed to be normal, diabetes, glaucoma, cataract, age-related macular degeneration, hypertension, and myopia.

Roy et al. developed an encoder–decoder network, ReLayNet, to segment the different retinal layers as well as accumulated fluid in OCT images [12]. Zhao et al. combined CNNs and fully connected random fields to segment retinal vessels in CFPs [13]. Palyout et al. utilized image-level annotations to help improve the segmentation accuracy of retinal lesions with a multi-task CNN architecture [14].

Compared to object detection and segmentation, much more attention has been paid to classifying ocular diseases with CNNs. Gulshan et al. classified CFPs according to the grade of diabetic retinopathy [15]. Li et al. conducted glaucomatous optic neuropathy classification [16]. Age-related macular degeneration classification was also frequently investigated [17–19]. Transfer learning with ImageNet pretrained models, such as the Inception network, was found to be very effective in this task of ocular disease classification [20,21]. Ensemble learning with multiple base networks could achieve even better classification performance [22]. These studies validate the feasibility of applying CNNs to achieve high sensitivity and specificity for ocular disease classification.

Despite the encouraging performance achieved, only a few works address the issue of multi-label ocular disease classification with CFPs [23], where one patient can have more than one type of ocular disease. Considering the high possibilities of patients get affected by multiple ocular diseases, optimizing models that can accommodate multi-label ocular disease classification is necessary. In the work of Li et al. it was identified that the coexistence of myopia leads to high false-negative predictions for glaucoma patient classification [16]. Consequently, even though existing studies generated satisfactory results in their specific tasks, they might not be applicable to real situations where complicated cases are inevitable (Fig. 1b).

In addition to the aforementioned issue, there is a lack of studies conducting patient-level ocular disease diagnosing. The majority of existing publications solve the problem in an image-level manner through independently analyzing the CFPs obtained from the left and right eyes. Works have suggested that the bilateral eyes are highly correlated with respect to the ocular disease progression [24], which indicates that patient-level diagnosis considering information from bilateral CFPs should contribute to a more effective approach. Besides, patient-level diagnosis can work as a pre-screening of patients with high risks. It is also more applicable when long-terms continuous monitoring of high-risk patients is in need [25]. Furthermore, most current CNN-based ocular disease classification studies directly utilize the networks developed for natural image analysis without architecture optimization, which limits the classification accuracy.

In this study, we tackle the problem of CFP-based patient-level multi-label ocular disease classification. Specifically, An elaborately designed CNN network, named as dense correlation network (DCNet), is proposed. The inputs to DCNet are pairs of CFPs obtained from the left and right eyes. The outputs are the corresponding possibilities of the patient getting infected by different ocular diseases. DCNet is composed of three modules, a backbone CNN module for the extraction of features from the individual left and right CFPs, a spatial correlation module for the feature refinement and fusion, and a classification module for the generation of classification outputs. Our proposed DCNet achieves inspiring classification performance on a public CFP dataset. A preliminary version of this work has been presented as a conference abstract [26]. In this paper, we comprehensively analyze the results. Particularly, we discuss the performance enhancement with regard to increasing network depth and clearly show the relationship between model complexity and model performance. In summary, our contributions are three-fold: (1) We propose a CNN model for the task of

patient-level multi-label ocular disease classification. Seven types of ocular diseases (six types in Fig. 1 and a remaining type of other diseases) can be processed simultaneously with a single network. (2) A novel module, SCM, is designed to effectively fuse the features extracted from the left and right CFPs. SCM refines the extracted features by taking their correlations into consideration. (3) Significantly enhanced classification performance is achieved on a publicly available CFP dataset by our proposed model with the elaborately designed feature correlation and fusion method compared to multiple baseline methods with direct feature concatenation.

2. Method

2.1. Network architecture

The overall architecture of the proposed DCNet is shown in Fig. 2a. DCNet has three major modules, the backbone CNN, the spatial correlation module (SCM), and the final classifier.

2.1.1. Backbone CNN

The backbone CNN extracts two independent sets of features from the inputted pairs of CFPs. Given the left and right CFPs, I_l and I_r ($I_l, I_r \in \mathbb{R}^{H\times W\times 3}$, H and W refer to the height and width of the input CFPs, 3 is the three color channels), the outputs of backbone CNN are F_l and F_r ($F_l, F_r \in \mathbb{R}^{h\times w\times c}$, h, w, and c are the height, width, and number of extracted features), respectively. Since there is no information exchange or fusion during the feature extraction process, no registration between the paired CFPs is required. We adopt different ResNet architectures truncating the fully-connected layers as our backbone CNNs [27].

2.1.2. Spatial correlation module

SCM receives the two feature sets from the backbone CNN and outputs two corresponding feature sets by taking the correlations between them into consideration. The details of the proposed SCM is illustrated in Fig. 2b. With the two input feature sets F_l and F_r, pixel-wise relationship between them is calculated. Inspired by the design of non-local neural networks [28], each of the two feature sets are firstly transformed into query, key, and value features (F_{lq}, F_{lk}, and F_{lv} for features of the left CFPs, and F_{rq}, F_{rk}, and F_{rv} for features of the right CFPs, where $F_{lq}, F_{lk}, F_{rq}, F_{rk} \in \mathbb{R}^{h\times w\times c'}$ and $F_{lv}, F_{rv} \in \mathbb{R}^{h\times w\times c''}$) by 1×1 convolutions:

$$F_{lk} = \text{Linear}(F_l;\theta_{lk}), \qquad F_{rk} = \text{Linear}(F_r;\theta_{rk}), \tag{1}$$

$$F_{lq} = \text{Linear}(F_l;\theta_{lq}), \qquad F_{rq} = \text{Linear}(F_r;\theta_{rq}), \tag{2}$$

$$F_{lv} = \text{Linear}(F_l;\theta_{lv}), \qquad F_{rv} = \text{Linear}(F_r;\theta_{rv}). \tag{3}$$

where "Linear" represents a 1×1 convolution with θ referring to the relevant parameters. c' and c'' are the dimensions of the transformed query/key features and value features, respectively. We empirically set $c' = c'' = 512$.

Then, with the transformed features, the pixel-wise correlation weights ($R_{l\leftarrow r} \in \mathbb{R}^{(h\times w)\times(h\times w)}$) to aggregate information extracted from the right CFP to that extracted from the left CFP is obtained as the inner product normalized by sigmoid function:

$$R_{l\leftarrow r} \in \mathbb{R}^{(h\times w)\times(h\times w)} = \text{Sigmoid}(F_{lq}F_{rk}^T) \tag{4}$$

The correlation weights ($R_{r\leftarrow l} \in \mathbb{R}^{(h\times w)\times(h\times w)}$) to aggregate information extracted from the left CFP to that extracted from the right CFP can be calculated in a similar way:

$$R_{r\leftarrow l} \in \mathbb{R}^{(h\times w)\times(h\times w)} = \text{Sigmoid}(F_{rq}F_{lk}^T) \tag{5}$$

The correlation weights capture the interactions between every locations in the paired CFPs. Once the weights are acquired, the two feature sets from the backbone CNN are refined to $F_{l\_update} \in \mathbb{R}^{w\times h\times c''}$ and $F_{r\_update} \in \mathbb{R}^{w\times h\times c''}$ by multiplying the respective weight maps:

$$F_{l\_update} = R_{l\leftarrow r} \times F_{rv}, \tag{6}$$

$$F_{r\_update} = R_{r\leftarrow l} \times F_{lv}. \tag{7}$$

The last step in SCM is to fuse the features obtained from the bilateral CFPs. As shown in Fig. 2b, the fusion is realized through concatenation of four feature sets. The input left CFP feature set F_l is concatenated with the corresponding updated left CFP feature set F_{l_update}, and the input right CFP feature set F_r is concatenated with the updated right CFP feature set F_{r_update}:

$$S_l = \text{Linear}([F_l,\ F_{l\_update}]^T;\theta_{sl}), \tag{8}$$

$$S_r = \text{Linear}([F_r,\ F_{r\_update}]^T;\theta_{sr}). \tag{9}$$

where $S_l, S_r \in \mathbb{R}^{h\times w\times c1}$ are the two outputs of SCM.

2.1.3. Classifier

The two output feature sets from SCM are transformed into two feature vectors by global average pooling. They are then concatenated before inputting into the final classifier module. The classifier consists of two fully-connected layers, one with ReLu activation and one without. The dimension of the concatenated features is reduced by the first fully-connected layer. The exact feature dimension depends on the backbone CNN utilized. Take ResNet-50 backbone CNN as an example, the concatenated feature has a dimension of 4096. It is reduced to 512 by the first fully-connected layer. The second fully-connected layer further reduces the features into a dimension of eight, which equals to the classification categories. The eight-dimension features can then be compared with the ground-truth disease classification labels, and the network loss can be calculated accordingly.

2.2. Loss function

The classic cross-entropy loss for image classification is usually employed following a softmax activation and it is more applicable when the categories are exclusive. Thus, according to similar existing studies [29–31], we employ a multi-label soft margin loss instead for our multi-label ocular disease classification task:

$$L = -\frac{1}{C}\sum_{c=1}^{C} y[c]log(\sigma(\hat{y}[c])) + (1-y[c])log(1-\sigma(\hat{y}[c])) \tag{10}$$

where c refers to the categories, $\sigma(\cdot)$ is the sigmoid activation, $y \in \{0,1\}$ is the reference label, and $\hat{y}$ is the output of the network.

2.3. Dataset description

We instantiate our proposed model with a publicly available CFP dataset. The dataset is provided by the 2019 University International Competition on Ocular Disease Intelligent Recognition (ODIR-2019).[1] The dataset contains eight different classification categories, including the normal control group (N) and seven disease groups (diabetes (D), glaucoma (G), cataract (C), age-related macular degeneration (A), hypertension (H), myopia (M), and other diseases/abnormalities (O)). The patient-level labels are generated with reference to both the CFPs and additional information, such as the age of the patient. In total, there are 5000 cases with the original CFPs, and annotations of 3500 cases are made publicly available. The distribution of the 3500 patient cases among the eight categories is shown in Fig. 3. These cases are utilized to investigate the effectiveness of our proposed model. Considering the relatively small data size, the 3500 cases are split into three folds randomly with 1167, 1167, and 1166 cases, and cross-validation is conducted by training on each combination of two folds and testing on the remaining fold. Finally, the average results of the test folds of three cross-validation experiments are reported.

[1] https://odir2019.grand-challenge.org.

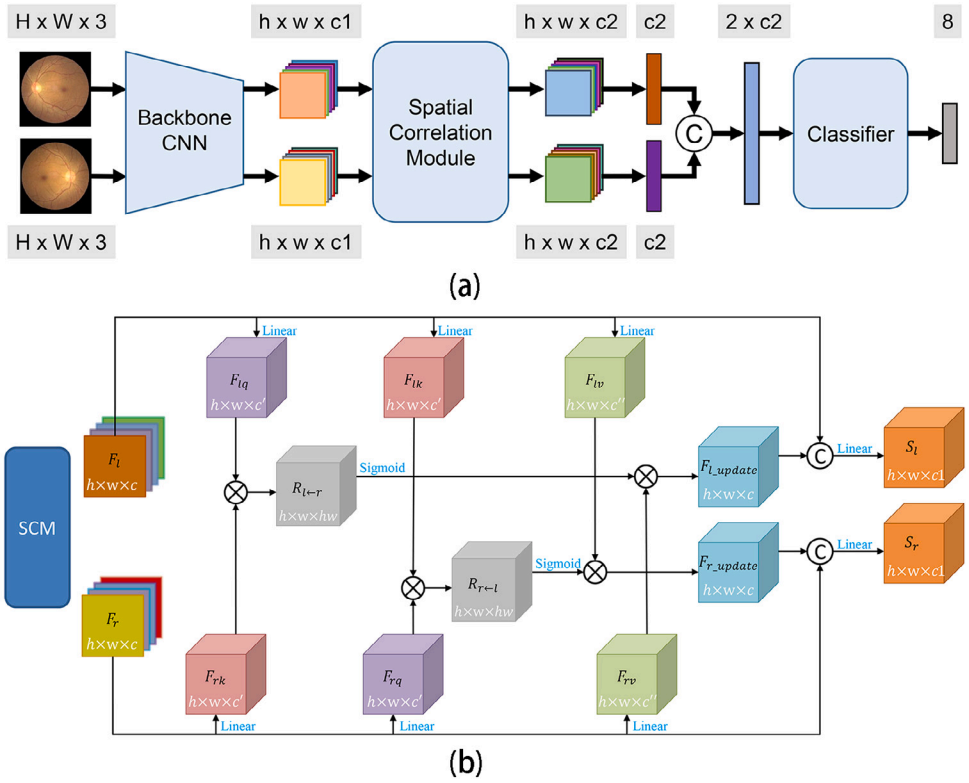

Fig. 2. (a) Network architecture of the proposed DCNet. The backbone CNN extracts features from both left and right CFPs from the same patients. The spatial correlation module (SCM) simulate the relationship between the two sets of extracted features. The classifier generates the classification scores of the eight categories (the seven categories in Fig. 1 and one additional category for other diseases/abnormalities. "C" is feature concatenation. (b) Detailed architecture of SCM. "Linear" is a convolution layer with a kernel size of 1 × 1. "×" stands for matrix multiplication and "C" is feature concatenation.

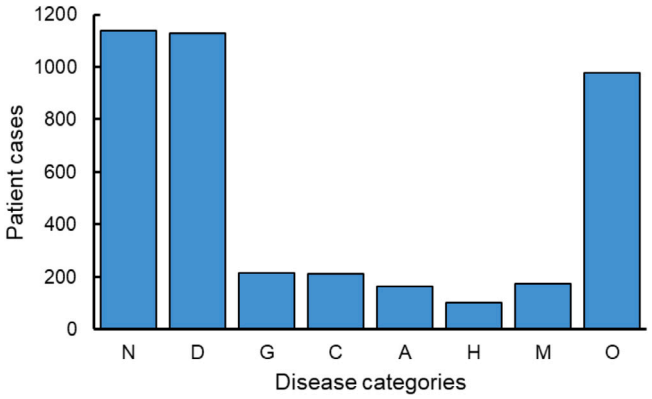

Fig. 3. Patient distribution over the eight classification categories. Characters N, D, G, C, A, H, and M have the same meaning as Fig. 1. O refers to other diseases/abnormalities.

2.4. Implementation details

ResNets of different depths are utilized as our backbone CNNs to investigate their influence on the feature extraction process. ImageNet-pretrained models are employed as the initialization of the backbones. Since the ODIR-2019 dataset was collected from different centers or hospitals with different cameras, the CFPs come with different resolutions. All CFPs are firstly resized to the same image resolution of 512 × 512. Then, random cropping of 448 × 448 image patches is conducted as an image augmentation method for the network training. During testing, center cropping is used instead.

All our deep neural networks are implemented with PyTorch [32]. Experiments are run on NVIDIA GeForce 1080Ti GPUs. Stochastic gradient decent (SGD) optimizer is adopted to train the networks with the multi-label classification loss function in Eq. (10). The learning rate is initially set to 0.007, which is decayed according to the poly learning rate decay policy $lr = initial_{lr} \times (1 - \frac{iter}{total_{iter}})^{power}$ [33]. The power is set as 0.9. All experiments are run for 50 epochs and the results of the last epoch are recorded.

2.5. Evaluation metrics

Four evaluation metrics, kappa score (k in Eq. (11)), F_1 score (F_1 in Eq. (12)), area under the receiver operating curve (AUC in Eq. (13)), and average of the three (AVG in Eq. (14)), are calculated to evaluate the classification performance of different models as suggested by the official ODIR-2019 challenge. The statistics are calculated over the whole dataset and averaged over the patients.

$$k = \frac{p_o - p_e}{1 - p_e}$$
$$p_o = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C}(TP_c + FN_c)} \tag{11}$$
$$p_e = \frac{\sum_{c=1}^{C} TP_c * (TP_c + FN_c)}{N^2},$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} = \frac{2TP}{2TP + FN + FP} \tag{12}$$

$$AUC = \int_{x=0}^{1} TPR(FPR^{-1}(x))dx \tag{13}$$
$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{FP + TN}$$

$$AVG = \frac{1}{k + F_1 + AUC} \tag{14}$$

Table 1
Ocular disease classification results with different backbone CNNs (without SCM).

Backbone CNN	Fusion strategy	Kappa	F_1	AUC	AVG
ResNet-18	Summation	0.496 ± 0.009	0.891 ± 0.003	0.909 ± 0.003	0.765 ± 0.005
	Multiplication	0.306 ± 0.123	0.869 ± 0.016	0.866 ± 0.026	0.680 ± 0.054
	Concatenation	0.527 ± 0.003	0.894 ± 0.001	0.914 ± 0.002	0.778 ± 0.001
ResNet-34	Summation	0.547 ± 0.041	0.899 ± 0.007	0.920 ± 0.009	0.789 ± 0.019
	Multiplication	0.426 ± 0.148	0.884 ± 0.016	0.897 ± 0.028	0.736 ± 0.064
	Concatenation	0.554 ± 0.008	0.898 ± 0.001	0.917 ± 0.001	0.790 ± 0.003
ResNet-50	Summation	0.577 ± 0.024	0.904 ± 0.004	0.925 ± 0.004	0.802 ± 0.011
	Multiplication	0.588 ± 0.028	0.904 ± 0.005	0.922 ± 0.007	0.804 ± 0.013
	Concatenation	0.598 ± 0.014	0.905 ± 0.003	0.927 ± 0.003	0.810 ± 0.007
ResNet-101	Summation	0.593 ± 0.007	0.906 ± 0.003	0.927 ± 0.001	0.809 ± 0.003
	Multiplication	0.604 ± 0.002	0.907 ± 0.001	0.928 ± 0.004	0.813 ± 0.001
	Concatenation	0.604 ± 0.016	0.907 ± 0.003	0.927 ± 0.006	0.812 ± 0.009

Table 2
Ocular disease classification results with different backbone CNNs with SCM.

Backbone CNN	Kappa	F_1	AUC	AVG
ResNet-18	0.545 ± 0.009	0.895 ± 0.001	0.914 ± 0.004	0.785 ± 0.005
ResNet-34	0.596 ± 0.014	0.904 ± 0.003	0.924 ± 0.001	0.808 ± 0.006
ResNet-50	0.628 ± 0.005	0.910 ± 0.002	0.928 ± 0.002	0.822 ± 0.002
ResNet-101	0.637 ± 0.007	0.913 ± 0.002	0.930 ± 0.004	0.827 ± 0.003

where $C = 8$ refers to the eight categories. N is the number of samples in the test set. TP, FP, TN, and FN refer to true positive predictions, false positive predictions, true negative predictions, and false negative predictions. TPR and FPR are true positive rate and false positive rate.

3. Experimental results and analysis

Two sets of experiments are conducted. The first set is to investigate the influence of backbone CNNs on the classification performance. The second set to validate the effectiveness of the proposed spatial correlation module.

3.1. Classification performance with different backbone CNNs

The backbone CNN is responsible for the independent extraction of features from bilateral CFPs. Backbone CNNs of different depths extract features of different abstraction levels. The proposed spatial correlation module (SCM in Fig. 2) is not included in this set of experiments. The generated features by the backbone CNN are fused through summation, pixel-wise multiplication, or concatenation directly without a special fusion process.

Table 1 lists the detailed results. Three phenomena can be observed. Firstly, among the different fusion strategies, feature concatenation works the best. Pixel-wise multiplication is more suitable when deep backbone CNN is utilized. Secondly, better performance is achieved with deeper backbone CNNs. Comparing the results of models with ResNet-101 backbone and feature concatenation to those with ResNet-18, the kappa score, F_1, AUC, and the final AVG are increased by 7.7%, 1.3%, 1.3%, and 3.4%, respectively. It indicates a better ocular disease distinction ability of higher abstraction features. Finally, it is also observed that the performance plateaus at models with ResNet-50 backbone, and the enhancement is very limited when replacing it with ResNet-101 backbone (Fig. 4). There are studies showing similar patterns that network performance cannot improve linearly with network depth [34,35]. Three possible causes can be summarized. The first is relevant to the gradient vanishing issue. The difficulty of network optimization increases with the network depth [34]. The diminishing of feature reuse is the second cause, which leads to insufficient usage of the generated large number of features for deep networks [35]. The last is due to limited available training samples, the network might not be properly trained.

Table 3
Computational complexities of different network configurations with/without SCM.

Backbone CNN	SCM	FLOPs (G)	Params (M)
ResNet-18	Without	14.6	11.7
ResNet-18	With	14.9	13.6
ResNet-34	Without	29.4	21.6
ResNet-34	With	29.8	23.7
ResNet-50	Without	33.0	25.8
ResNet-50	With	38.8	55.2
ResNet-101	Without	62.9	44.8
ResNet-101	With	68.7	74.2

3.2. Performance enhancement by the proposed spatial correlation module

Table 2 presents the results when SCM is enabled. The overall performance trend with increasing of network depth is the same as that without SCM. Introducing dense feature correlations through SCM into the classification models consistently improves the classification accuracy regardless of the backbone CNNs (Fig. 5). Take the model with ResNet-50 backbone CNN as an example, utilizing SCM increases the four metrics by 3.0%, 0.5%, 0.1%, and 1.2%, respectively. Paired t-tests between the results obtained without and with SCM confirm that SCM can significantly improve the classification performance characterized by the kappa score, F_1 score, and the final average score (with a p value smaller than 0.05).

A similar performance plateau is observed when different backbone CNNs are used. Characterizing by the final average score, networks with ResNet-34 backbone increases AVG by 2.3% compared to ResNet-18 backbone. Networks with ResNet-50 backbone increase AVG by 1.4% over ResNet-34 backbone. On the other hand, networks with ResNet-101 backbone increases AVG by only 0.5% over ResNet-50 backbone. Considering the increased computational complexity with the increased network depth, it might not be necessary to utilize the deepest backbone CNN. We will discuss this in the following section.

3.3. Computational complexities of different models

To compare the classification performance in a fair manner, we present the network complexities in Table 3 and plot the classification metrics with regard to the floating point of operations (FLOPs, Fig. 6a) and the network parameters (Fig. 6b).

Taking FLOPs or network parameters into consideration, the proposed method can still outperform the baseline with a large margin. Networks with ResNet-50 backbone and SCM can perform better than baselines with ResNet-101 backbone with much fewer FLOPs. It again validates that the highly accurate classification results of the proposed method are not solely caused by the increased network complexities. The correlation between left and right CFPs is effective for the patient-level ocular disease classification. As patient-level diagnosis is important [25], it is necessary to take care of the information from

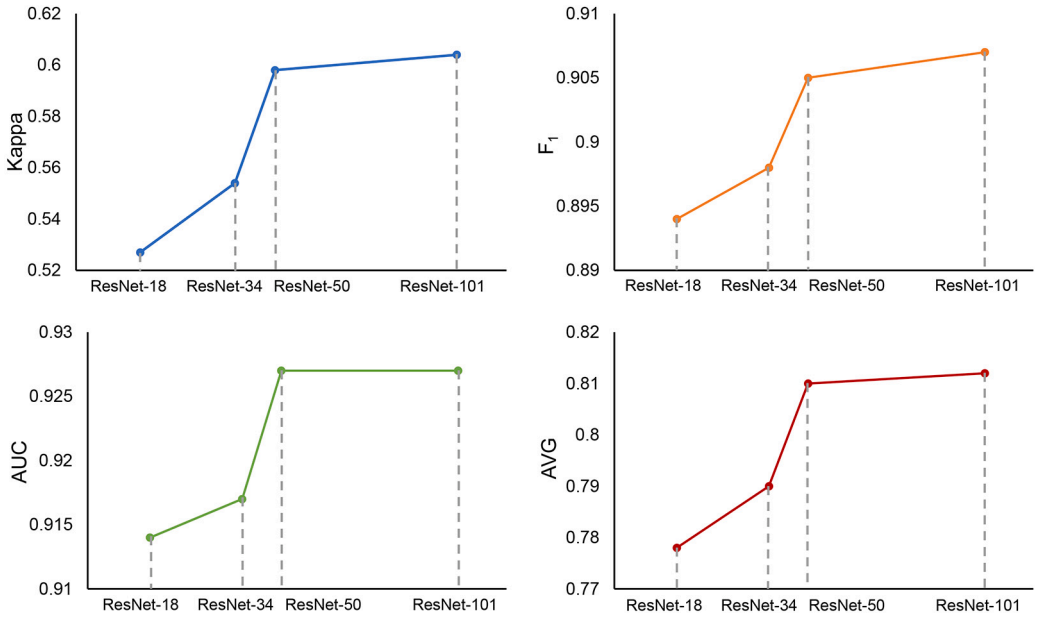

Fig. 4. Classification performance with different backbone CNNs.

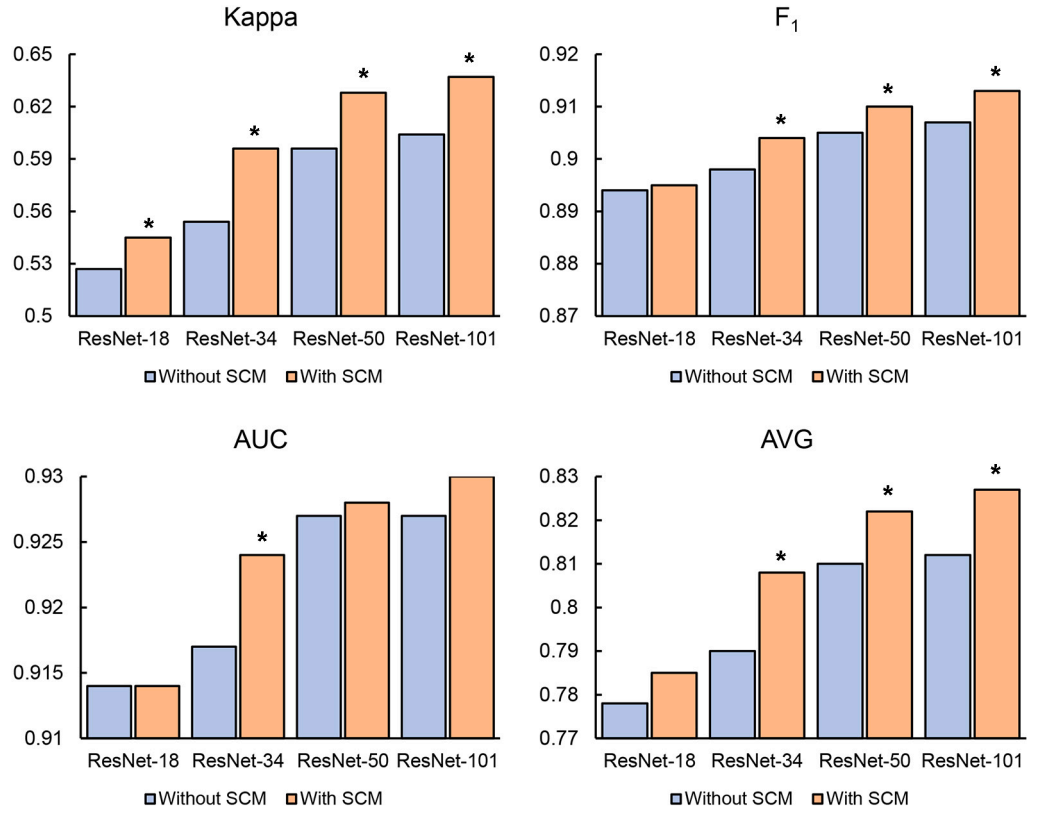

Fig. 5. Classification performance with different network configurations. * indicates significant difference between the two experiments with $p < 0.5$ by t-test.

both CFPs to boost the classification performance. It is to be noted that although the FLOPs increase caused by the introduction of SCM is relatively small, the network parameters increase a lot, especially for baselines with ResNet-50 and ResNet-101 backbones.

At the same time, from Fig. 6, it is clear that with SCM enabled, increasing network depth from utilizing ResNet-18 backbone to utilizing ResNet-50 backbone, the network performance increases almost linearly with FLOPs. But continuously increasing the network depth by adopting the ResNet-101 backbone brings only slight performance enhancement. Therefore, if available computational power is limited, it is not necessary to use very deep backbone CNNs, such as the ResNet-101 backbone.

4. Conclusion

In this study, we developed a patient-level multi-label ocular disease classification model, dense correlation network (DCNet). DCNet has three major modules, a backbone CNN, a spatial correlation module

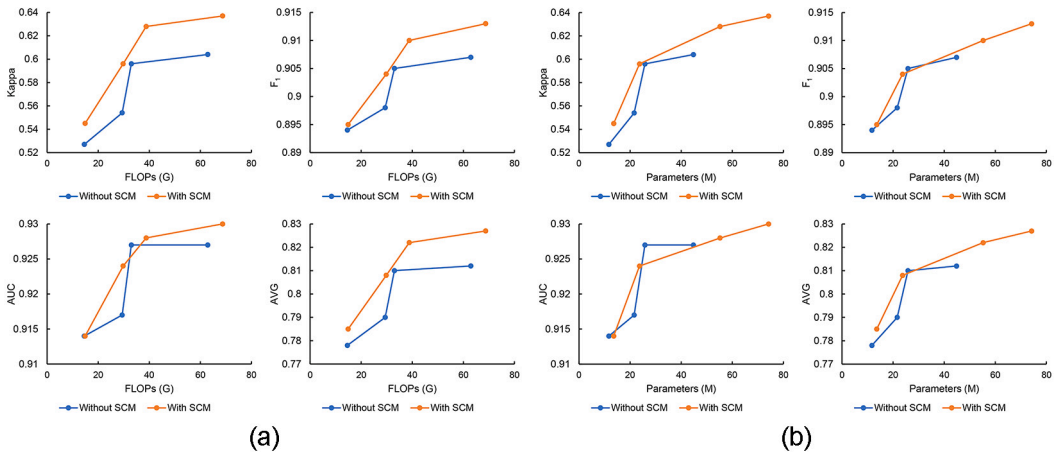

Fig. 6. (a) Classification performance with different network configurations with respect to FLOPs. (b) Classification performance with different network configurations with respect to network parameters.

(SCM), and a classifier. Our major novelty regarding the network design lies in SCM, where pixel-wise feature correlation proceeds with the two sets of features extracted from left and right CFPs by the backbone CNN. The processed features are then concatenated to extract a patient-level feature representation for the final patient-level ocular disease classification. Extensive experiments on a public dataset, ODIR-2019, were conducted to validate the effectiveness of the proposed method, and the proposed method could always generate better results than the corresponding baselines while possessing lower computational complexities.

There are several limitations that exist in the current method that should be addressed in our future studies. Firstly, because only patient-level ocular disease category labels were provided, we were not able to compare the results of the proposed method to those conducting image-level classifications. Secondly, to control the computational complexities, the backbone CNNs to extract features from left and right CFPs were shared. Finally, although the distribution of patient cases over the different categories was heavily unbalanced, we did not address this issue explicitly. Meanwhile, although our current approach did not require the registration between the inputted two CFPs, we believe it is helpful when the two CFPs are roughly registered, which can be done under the guidance of segmented key biomarkers (such as the optic disk and macula).

The proposed method can be easily extended beyond ocular disease classification. It can be modified and optimized to similar tasks, such as the breast cancer diagnosis that needs to take bilateral breasts into consideration. Besides, the method can also be employed to conduct multi-modal image analysis as long as the correlations between the different modal images are important for the end tasks.

CRediT authorship contribution statement

Junjun He: Conceptualization, Methodology, Writing - original draft. **Cheng Li:** Conceptualization, Methodology, Writing - original draft. **Jin Ye:** Software, Visualization, Investigation, Validation. **Yu Qiao:** Supervision, Writing - review & editing. **Lixu Gu:** Supervision, Writing - review & editing.

Acknowledgements

This research is partially supported by the National Key Research and Development Program (No. 2016YFC0106200), Beijing Natural Science Foundation-Haidian Original Innovation Collaborative Fund (No. L192006), and the funding from Institute of Medical Robotics of Shanghai Jiao Tong University as well as the 863 National Research Fund (No. 2015AA043203). This research was also partially supported by the National Key Research and Development Program of China (2016YFC1400704), National Natural Science Foundation of China (U1613211 and U1713208), Shenzhen Basic Research Program (JCYJ20170818164704758 and CXB201104220032A), the Joint Lab of CAS-HK, and Shenzhen Institute of Artificial Intelligence and Robotics for Society.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Sommer, J.M. Tielsch, J. Katz, H.A. Quigley, J.D. Gottsch, J.C. Javitt, J.F. Martone, R.M. Royall, K.A. Witt, S. Ezrine, Racial differences in the cause-specific prevalence of blindness in east baltimore, New Engl. J. Med. 325 (1991) 1412–1417.

[2] N. Congdon, B. O'Colmain, C.C. Klaver, R. Klein, B. Munoz, D.S. Friedman, J. Kempen, H.R. Taylor, P. Mitchell, for the Eye Diseases Prevalence Research Group, Causes and prevalence of visual impairment among adults in the united states, Arch. Ophthalmol. 122 (2004) 477–485.

[3] R.R. Bourne, G.A. Stevens, R.A. White, J.L. Smith, S.R. Flaxman, H. Price, J.B. Jonas, J. Keeffe, J. Leasher, K. Naidoo, S. Resnikoff, H.R. Taylor, on the behalf of the Vision Loss Expert Group, Causes of vision loss worldwide, 1990–2010: a systematic analysis, Lancet Glob. Health 1 (2013) e339–349.

[4] R. Bernardes, P. Serranho, C. Lobo, Digital ocular fundus imaging: a review, Ophthalmologica 226 (2011) 161–181.

[5] S. Rowe, C.H. MacLean, P.G. Shekelle, Preventing visual loss from chronic eye diseases in primary care: scientific review, JAMA 291 (2004) 1487–1495.

[6] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.

[7] W. Zhao, J. Yang, Y. Sun, C. Li, W. Wu, L. Jin, Z. Yang, B. Ni, P. Gao, P. Wang, et al., 3d deep learning from ct scans predicts tumor invasiveness of subcentimeter pulmonary adenocarcinomas, Cancer Research 78 (24) (2018) 6881–6889.

[8] J. Yang, H. Deng, X. Huang, B. Ni, Y. Xu, Relational learning between multiple pulmonary nodules via deep set attention transformers, in: ISBI, 2020, pp. 1875–1878.

[9] B. Liefers, F.G. Venhuizen, V. Schreur, B.V. Ginneken, C. Hoyng, S. Fauser, T. Theelen, C.I. Sanchez, Automatic detection of the foveal center in optical coherence tomography, Biomed. Opt. Express 8 (2017) 5160–5295.

[10] X. Meng, X. Xi, L. Yang, G. Zhang, Y. Yin, X. Chen, Fast and effective optic disk localization based on convolutional neural network, Neurocomputing 312 (2018) 285–295.

[11] C.S. Lee, A.J. Tyring, N.P. Deruyter, Y. Wu, A. Rokem, A.Y. Lee, Deep-learning based, automated segmentation of macular edema in optical coherence tomography, Biomed. Opt. Express 8 (7) (2017) 3440–3448.

[12] A.G. Roy, S. Conjeti, S.P.K. Karri, D. Sheet, A. Katouzian, C. Wachinger, N. Navab, Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks, Biomed. Opt. Express 8 (8) (2017) 3627–3642.

[13] H. Zhao, H. Li, S. Maurer-Stroh, Y. Guo, Q. Deng, L. Cheng, Retinal vessel segmentation of color fundus images using multiscale convolutional neural network with an improved cross-entropy loss function, Neurocomputing 309 (2018) 179–191.

[14] C. Playout, R. Duval, F. Cheriet, A novel weakly supervised multitask architecture for retinal lesions segmentation on fundus images, IEEE Trans. Med. Imaging 38 (10) (2019) 2434–2444.

[15] V. Gulshan, L. Peng, M. Coram, M.C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Guadros, R. Kim, R. Raman, P.C. Nelson, J.L. Mega, D.R. Webster, Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, JAMA 316 (2016) 2402–2410.

[16] Z. Li, Y. He, S. Keel, W. Meng, R.T. Chang, m. He, Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs, Ophthalmology 125 (8) (2018) 1199–1206.

[17] P.M. Burlina, N. Joshi, M. Pekala, K.D. Pacheco, D.E. Freund, N.M. Bressler, Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks, JAMA Ophthalmol. 135 (11) (2017) 1170–1176.

[18] F. Grassmann, J. Mengelkamp, C. Brandl, S. Harsch, M.E. Zimmermann, B. Linkohr, A. Peters, I.M. Heid, C. Palm, B.H.F. Weber, A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography, Ophthalmology 125 (9) (2018) 1410–1420.

[19] S. Matsuba, H. Tabuchi, H. Ohsugi, H. Enno, N. Ishitobi, H. Masumoto, Y. Kiuchi, Accuracy of ultra-wide-field fundus ophthalmoscopy-assised deep learning, a machine-learning technology, for detecting age-related macular degeneration, Int. Ophthalmol. 39 (6) (2019) 1269–1275.

[20] S.P.K. Karri, D. Chakraborty, J. Chatterjee, Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration, Biomed. Opt. Express 8 (2017) 579–592.

[21] D.S. Kermany, M. Goldbaum, W. Cai, C.C.S. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Xu, F. Yan, J. Dong, M.K. Prasadha, J. Pei, M. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V.A.N. Huu, C. Wen, E.D. Zhang, C.L. Zhang, O. Li, X. Wang, M.A. Singer, X. Sun, J. Xu, A. Tafreshi, M.A. Lewis, H. Xia, K. Zhang, Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (2018) 1122–1131.

[22] F. Li, H. Chen, Z. Liu, X.-D. Zhang, M.-S. Jiang, Z.-Z. Wu, K.-Q. Zhou, Deep learning-based automated detection of retinal diseases using optical coherence tomography images, Biomed. Opt. Express 10 (2019) 6204–6226.

[23] D.S.W. Ting, C.Y.-L. Cheung, G. Lim, G.S.W. Tan, N.D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I.Y.S. Yeo, S.Y. Lee, E.Y.M. Wong, C. Sabanayagam, M. Baskaran, F. Ibrahim, N.C. Tan, E.A. Finkelstein, E.L. Lamoureux, I.Y. Wong, N.M. Bressler, S. Sivaprasad, R. Varma, J.B. Jonas, M.G. He, C.-Y. Cheng, G.C.M. Cheung, T. Aung, W. Hsu, M.L. Lee, T.Y. Wong, Development and validation of a deep learning algorithm for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes, JAMA 318 (2017) 2211–2223.

[24] A. Group, A simplified severity scale for age-related macular degeneration: Areds report no. 18, Arch. Ophthalmol. 123 (2005) 1570–1574.

[25] Y. Peng, S. Dharssi, Q. Chen, T.D. Keenan, E. Agron, W.T. Wong, E.Y. Chew, Z. Lu, Deepseenet: A deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs, Ophthalmology 126 (2019) 565–575.

[26] C. Li, J. Ye, J. He, S. Wang, Q. Yu, L. Gu, Dense correlation network for automated multi-label ocular disease detection with paired color fundus photographs, in: International Symposium on Biomedical Imaging, 2020, pp. 1250–1253.

[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 770–778.

[28] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.

[29] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, W. Xu, Cnn-rnn: A unified framework for multi-label image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2285–2294.

[30] M. Lapin, M. Hein, B. Schiele, Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2018) 1533–1554.

[31] Z. Chen, X. Wei, P. Wang, Y. Guo, Multi-label image recognition with graph convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5177–5186.

[32] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch, in: Conference on Neural Information Processing Systems, 2017, pp. 1–4.

[33] W. Liu, A. Rabinovich, A.C. Berg, Parsenet: Looking wider to see better, in: International Conference on Learning Representations, 2016.

[34] R.K. Srivastava, K. Greff, J. Schmidhuber, Training very deep networks, in: Advances in Neural Information Processing Systems, 2015.

[35] S. Zagoruyko, N. Komodakis, Wide residual networks, in: BMVC, 2016.