nnMobileNet: Rethinking CNN for Retinopathy Research

Wenhui Zhu[1]* Peijie Qiu[2]* Xiwen Chen[3]* Xin Li[1] Natasha Lepore[4] Oana M. Dumitrascu[5]
Yalin Wang[1]†
[1] School of Computing and Augmented Intelligence, Arizona State University
[2] McKeley School of Engineering, Washington University in St. Louis
[3] School of Computing, Clemson University
[4] CIBORG Lab, Department of Radiology Children's Hospital Los Angeles
[5] Department of Neurology, Mayo Clinic
wz52@asu.edu, ylwang@asu.edu

Abstract

Over the past few decades, convolutional neural networks (CNNs) have been at the forefront of the detection and tracking of various retinal diseases (RD). Despite their success, the emergence of vision transformers (ViT) in the 2020s has shifted the trajectory of RD model development. The leading-edge performance of ViT-based models in RD can be largely credited to their scalability—their ability to improve as more parameters are added. As a result, ViT-based models tend to outshine traditional CNNs in RD applications, albeit at the cost of increased data and computational demands. ViTs also differ from CNNs in their approach to processing images, working with patches rather than local regions, which can complicate the precise localization of small, variably presented lesions in RD. In our study, we revisited and updated the architecture of a CNN model, specifically MobileNet, to enhance its utility in RD diagnostics. We found that an optimized MobileNet, through selective modifications, can surpass ViT-based models in various RD benchmarks, including diabetic retinopathy grading, detection of multiple fundus diseases, and classification of diabetic macular edema. The code is available at https://github.com/Retinal-Research/NN-MOBILENET

1. Introduction

Retinal diseases (RD), such as diabetic retinopathy (DR), age-related macular degeneration, inherited retinal conditions, myopic maculopathy, and retinopathy of prematurity, are major contributors to blindness worldwide [37]. Deep neural networks, particularly convolutional neural networks (CNNs), have been extensively used in retinal image analysis over the past decades, achieving cutting-edge results in various RD-related tasks [4, 15, 17, 27, 36, 39, 40, 42, 43]. The effectiveness of CNNs in these applications is largely due to their built-in architectural inductive biases, such as spatial hierarchies, locality, and translation invariance. These characteristics enable CNNs to transform local visual elements like edges and textures into complex, high-level abstracted features. Building on this approach, numerous CNN-based RD models [4, 15, 36, 39] have incorporated disease-specific biases into their designs. However, the specialized nature of these CNN-based models for RD limits their versatility across a range of RD tasks.

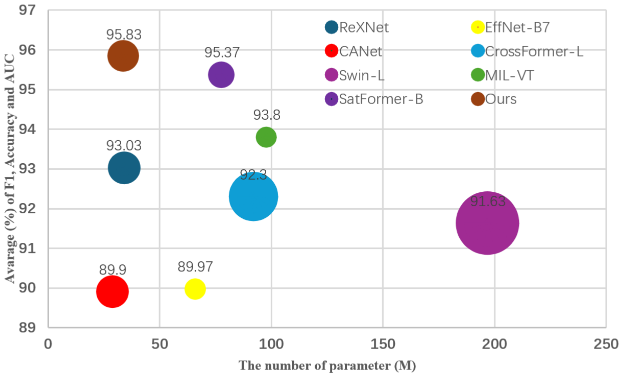


Figure 1. Model size vs average performance (F1, Accuracy and AUC) on retinal multi-disease abnormal detection using RFMid dataset. Our method demonstrates superiority over other CNN/ViT based methods in terms of performance and efficiency.

Recent advancements in RD modeling [14, 16, 30, 35, 38] have largely revolved around the vision transformer [7] (ViT) since its debut in the 2020s. The prowess of ViT-based models in RD is primarily due to their capacity to

*These authors contributed equally to this paper

†Corresponding Author

scale effectively; their performance improves as the model size grows [21]. Consequently, ViT-based methods typically exhibit superior performance over CNNs but with a cost of computational burden (as evident in Fig. 1). In addition, ViT-based models are advantageous for capturing long-range global dependencies via self-attention mechanism. Nonetheless, the quadratic time and memory complexity of self-attention operation make ViTs computationally intense and data-hungry. Accordingly, ViT-based RD models typically necessitate pretraining on large-scale datasets [14, 16, 38]. Furthermore, unlike CNNs, these models generally lack locality, since they operate at the image patch level [16].

To address these challenges, new iterations of ViT designs bring back convolution-like features to recover local context sensitivity [16, 20]. This adaptation is particularly beneficial for RD research since RD lesions are typically small and have heterogeneous appearances. One representative example can be found in the DR task, where four distinct types of lesions (i.e., microaneurysms, soft exudates, hemorrhages, and hard exudates) exhibit variations in shape, size, structure, and contrast. Among these lesions, the microaneurysms are too tiny to be easily detected. In addition, the DR grading task inherently contains hierarchical information, e.g., a proliferative DR image may consist of all types of lesions. These intrinsic properties in RD tasks naturally align with inductive biases in CNNs, where hierarchical and fine-grained local contexts can be better detected than ViTs. This observation prompts a reconsideration: *could CNNs be inherently more suited to RD tasks than ViTs?*

We also note that recent studies have shown well-tuned CNNs surpassing ViTs in general image classification tasks [21]. Motivated by these findings, our research recalibrated a CNN, specifically MobileNetV2 [28], for RD applications, focusing on training strategies and architectural refinements such as the inverted bottleneck, dropout optimization, and activation functions. To this end, we introduce nnMobileNet ("no-new" MobileNet), a model that implements strategic yet minimal enhancements. Our empirical results confirmed nnMobileNet's superiority over many leading RD models across a spectrum of benchmarks (see Fig. 1). Our work not only substantiates the potential of CNNs in RD research but also emphasizes the critical aspects of CNN optimization. We anticipate that our findings will spark a renewed interest in the adaptability and fine-tuning of CNNs within the field.

2. Related Works

2.1. Diabetic Retinopathy Assessment

In the domain of deep learning, the evaluation of diabetic retinopathy (DR) encompasses two tasks: DR grading and the classification of referable DR. The process of DR grading adheres to a protocol that categorizes the progression of diabetic retinopathy into distinct stages based on lesion examination, facilitating a multi-class classification task. This grading delineates five levels of severity: no retinopathy, mild non-proliferative DR (NPDR), moderate NPDR, severe NPDR, and proliferative DR (PDR). In contrast, the task of identifying referable DR focuses on detecting that may lead to blindness or significant vision loss resulting from DR, thereby being treated as a binary classification framework. The framework distinguishes between non-referable DR, characterized by the absence or mild presence of NPDR without significant pathological manifestations, and referable DR (rDR), which encompasses conditions of moderate severity or higher. In the past decade, deep learning has achieved state-of-the-art performance in automating the diagnosis of DR. Following the trend, convolutional neural networks (CNN) dominated the early stage of development [4, 15, 17, 27, 36, 38, 39]. Among them, Zoom-in-Net [36] took a biomimetic method (medical experts utilized the magnification to locate the lesion in the diagnosis) that incorporated the multiple scale information into CNN. Zhou et al. proposed a semi-supervised learning framework, which coordinated lesion segmentation and classification tasks by utilizing pixel-level supervision [39]. CANet [15] integrated two attention modules to jointly generate disease-specific and disease-dependent features for grading DR and diabetic macular edema (DME). Che et al. [4] achieved good performance via robust disentangled features of DR/DME. Essentially, These CNN-based DR classification methods rely on the extra auxiliary task and prior knowledge, which inevitably introduce more complex models and specialized multi-task datasets(e.g., DME Classification and lesion segmentation). Vision Transformers (ViT) have recently gained much attention in various visual tasks by leveraging the self-attention mechanism to capture long-term feature dependencies. Along this direction, MIL-VT [38] proposed using multiple-instance pooling to aggregate the features extracted by a ViT. Sun et al. [30] proposed a lesion-aware transformer (LAT) to learn the diabetic lesion-specific features via a cross-attention mechanism. Although those methods achieved state-of-the-art performance, most heavily relied on pretraining on large-scale datasets due to the data-hungry nature of ViT whose complexity quadratically grew concerning the input size. In addition, the DR features are localized in nature, e.g..fine-grained lesions such as microaneurysms typically occupy only a minor fraction of the image area and are discretely distributed in vessels. It was challenging for pure transformer-based feature extractors to focus more on global representations. In this paper, we contend that the capabilities of CNNs for DR tasks remain underutilized. We argue that through fine-tuning techniques, CNNs can achieve significant performance improvements, potentially surpassing ViT. To substantiate our claim, we initiate our investigation with a lightweight framework, MobileNet, and conduct a series of empirical studies.

2.2. Multi Retinopathy abnormal detection

The Multi-Retinopathy delineates a broader subclassification of Retinopathy, introducing more precise representations of lesions. Several fundus images may carry one or multiple labels, such as asteroid hyalosis, anterior ischemic optic neuropathy, age-related macular degeneration, branch retinal vein occlusion, Choroidal folds, etc. Notably, many of these pathological changes are interrelated; for instance, the presence of cotton wool spots on the retina is a characteristic ocular manifestation of various medical conditions, including diabetes mellitus, systemic hypertension, leukemia, and AIDS [3]. CNNs remain dominant as the foundational design approach for multi-retinopathy abnormal detection. A significant portion of the benchmark methods based on CNNs originates from DR grading models. A notable example of such work is the development of CANet [15], which leverages multi-task learning to extract additional semantic information, thereby aiding the classification model. Most subsequent advancements in CNN-based methods have followed this conceptual framework [4, 15]. Contrasting with this trend, some studies argue that establishing long-range dependencies and capturing global semantic information learning is a potentially more effective strategy for advancing model capabilities. The MIL-VT introduces the ViT and incorporates multiple instance learning head to force the token to capture the lesion information [38]. However, this method processes each individual patch without emphasizing the semantics of smaller lesions, resulting in a lack of localized information modeling. Furthermore, they employed extensive external datasets for pre-training due to data-hangry nature of ViT. In contrast, SatFormer enhances the ViT framework by integrating multi-scale CNNs to detect small lesions, such as microaneurysms and exudates. This approach enriches the model's capability to represent features of small lesions and to capture a wide range of pathological semantics [14]. This transition and amalgamation from ViT back to CNN prompt us to ponder whether CNNs are more suited for RD than ViTs or whether the potential of CNNs remains underexploited. This curiosity underpins our motivation for conducting deeper research into the CNNs in various RD tasks.

2.3. Myopic maculopathy grading

Recent trends show a growing interest in leveraging deep learning for the automatic diagnosis and analysis of myopic macular degeneration, the most prevalent complication of myopia and the leading cause of vision loss in individuals with pathological myopia. At the recently concluded MICCAI 2023, the Automated Detection of Myopic Maculopathy in MMAC 2023 challenge featured tasks in myopic maculopathy grading, segmentation, and prediction of spherical equivalent. Insights from the released solutions reveal that the first-place winner utilized a two-stage pre-training method with a CNN backbone, incorporating vision-language pre-training and self-supervised visual representation learning. The second-place team employed a Swin Transformer backbone combined with ArcFace loss. Interestingly, the third-place entry achieved commendable results using a lightweight CNN model without needing external retinal datasets for pre-training or self-supervised learning [13, 23, 41]. This outcome suggests that CNNs can still excel in performance, potentially outpacing ViTs in RD tasks. Moreover, the equitable testing environment of such challenges lends credibility to the results [8]. These findings corroborate our initial hypothesis and further pique our interest in exploring the fine-tuning of CNNs for RD tasks.

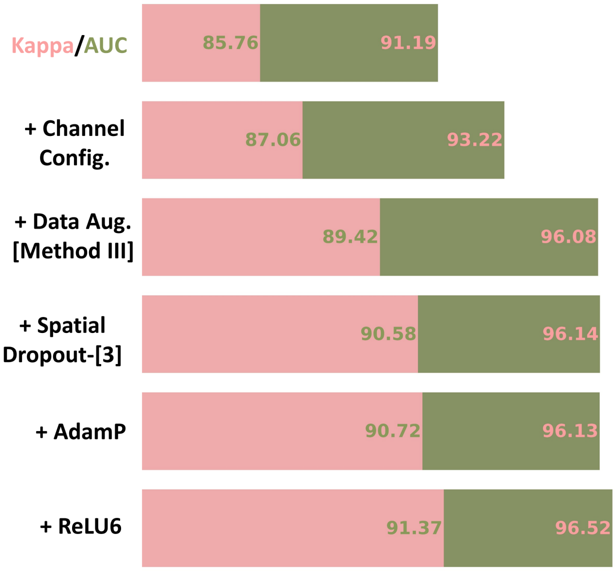


Figure 2. The roadmap of modifying a MobileNetV2 to the proposed no-new MobileNet (nnMobileNet) on the Messidor-2 dataset;

3. Roadmap of a nnMobileNet

Our investigation started with a standard MoblieNetV2 [28] on the Messidor-2 dataset (see the first row in Fig.2) [6]. We chose MobileNetV2 because of its efficiency, achieved by replacing the traditional residual bottleneck [10] with an inverted linear residual bottleneck (ILRB) where channel attention was included by default (see details in Fig.3). We conducted empirical studies on its key components, including channel configuration, data augmentation, dropout, op-

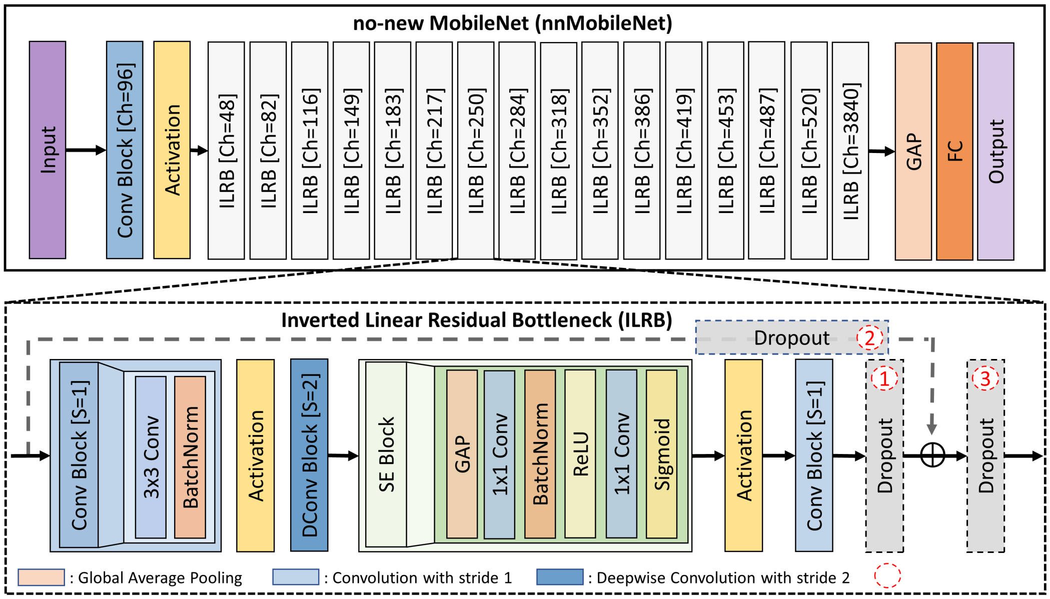
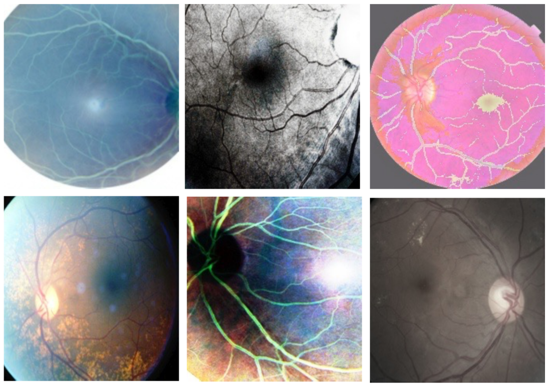


Figure 3. The detailed architecture of the no-new MobileNet (Including the Channel configuration) and the inverted linear residual bottleneck used in the no-new MobileNet.



	Details of augmentation
Method I	random horizontal flips, vertical flips and random cropping. rotation, scaling, gaussian blur, color jittering and mirroring.
Method II	Method 1 with Mixup and CutMix.
Method III	resized crop, horizontal flip, auto contrast, equalize, invert, rotate, posterize, Solarize, color, sharpness, shear, brightness, random erasing, Translate + Mixup and CutMix.

Figure 4. Examples of data augmentation (Method III) and details of three sets of data augmentation we used.

timizer, and activation functions.

3.1. Channel Configuration of ILRB

Recent findings in natural images have revealed that changing stage-wise channel configuration (primarily computation distribution between layers) led to a remarkable performance gain [9, 21]. This improvement is primarily due to two factors: first, the expressiveness of a layer is affected by the rank of its output matrix [9]. Second, ViTs generally use a different stage ratio from CNNs to change its computation distribution [21]. Both of these factors indicated the necessity of changing the channel configuration in MobileNetV2.

As consistent with findings in natural images, we empirically found that changing the channel configuration led to a performance gain of 2.03% in Kappa and 1.30% in AUC (see the second row in Fig. 2). It is worth noting that we follow the channel configuration in [9] (see Fig.3 no-new MobileNet architecture for channel configuration details).

3.2. Data Augmentation

In the field of RD, a common belief was that heavy data augmentation (e.g., Mixup and CutMix) should be avoided, as it dramatically distorts structures of images [14, 15, 30]. However, recent research has revealed that heavy data augmentation even boosted the performance in retinal vessel segmentation [33]. We hypothesize that this finding can be transferred to the RD tasks because introducing noise from the heavy data augmentation (e.g., unrealistic images shown in Fig.4) may help the model generalize better. To validate those hypotheses, we conducted experiments on three different sets of data augmentation combinations from light to heavy (as detailed by Methods I, II, III in Fig.4).

Interestingly, we found that the heaviest data augmentation (i.e., Method III) achieved the best performance compared to the other two strategies (see Fig. 5). We con-

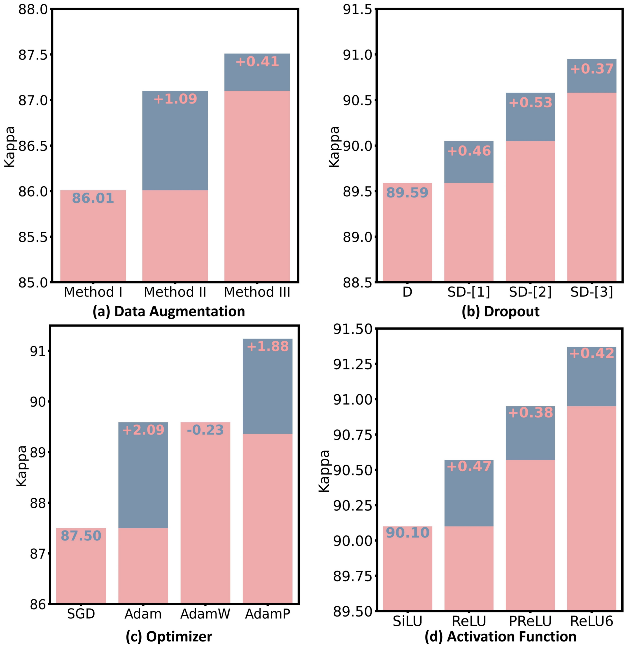


Figure 5. Empirical studies on Messidor-2 dataset where subpanel pictures (a), (b), (c), and (d) represent different experimental groups, each of which is independent of the others. D and SD-[x] in subpanel (b) denote Dropout and SpatialDropout in position [x] as shown in Fig.3(c), respectively.

jectured that different from ViTs which perform classification by comparing patches, the lack of non-local context in CNNs may necessitate heavy data augmentation to find the most discriminative local feature representation for abstracting heterogeneous RD lesions. Integrating this set of data augmentation into our nnMobileNet model design can improve Kappa by 2.36% and AUC by 2.86% (see the third row in Fig. 2).

3.3. Dropout

Dropout is commonly employed to alleviate overfitting and enhance the representational capacity of individual layers. However, we argue that the standard Dropout [12], which randomly zeros out neurons in a feature map, is unsuitable for general RD tasks. This is mainly due to the fact that specific lesion information is primarily encoded in certain image color channels. An example can be seen in the case of DR, where the information of microaneurysm predominantly resides in the red channel, while the exudate information is primarily encoded in the green channel. Based on this observation, we conjecture that spatial Dropout, which randomly removes channels from a feature map, is a better choice for RD tasks. Additionally, spatial Dropout naturally preserves spatial and local structures by randomly dropping out strongly activated local patterns [32], which is highly needed in RD. However, *where to place spatial Dropout remains an open problem.* Here, we conducted experiments to investigate the strategic placement of spatial Dropout in ILRB (see Fig.3 inverted linear residual bottleneck).

We observed that i) spatial Dropout is indeed more effective than standard Dropout for RD tasks and ii) the performance varies when placing the spatial Dropout at different positions (see Fig.5(b)). Integrating spatial Dropout into the proposed model led to an improvement of 0.06% in Kappa and 1.16% in AUC (see the forth row in Fig.2).

3.4. Optimizer

The training of ViTs is typically performed by an AdamW [22] optimizer, which makes us wonder if the performance gain in ViT-based RD models comes from the more advanced optimizer [21]. Alternatively, *would a more advanced optimizer boost the performance of a CNN-based RD model ?*

Our empirical studies revealed that training the network with AdamP [11] optimizer, which better accommodates the step size adaptively, showed better performance compared to other optimizers (see Figure 5(c)). Applying the AdamP to train nnMobileNet contributed to a performance gain of 2.2% in Kappa (see the fifth row in Fig. 2).

3.5. Activation Function

ReLU is extensively used in traditional CNNs due to its simplicity and computational efficiency. However, increasing research indicates that smoother variants of ReLU (e.g., SiLU), commonly used in ViTs, can lead to performance improvements [9, 21]. Based on that, we investigate the most suitable activation functions within inverted linear residual blocks (see Fig.3 inverted linear residual bottleneck) for retinal disease applications. Here, we consider four variants of ReLU, including SiLU, ReLU, PReLU, and ReLU6. As shown in Fig.5(d), the ReLU6 activation was the best among all options. After we replaced the ReLU with ReLU6 in each ILRB, it led to an improvement of 0.65% in kappa and 0.39% in AUC (see the sixth row in Fig. 2).

4. Experiments and Results

An optimal set of network structures and training strategies is summarized in Section 3. We used cross-entropy loss for training all the models in this work. All models were trained for 1000 epochs with a batch size of 32. The initial learning rate was set to 0.001 decayed according to a cosine decay learning rate scheduler with 20 epochs of linear warm-up. A weight decay rate of 0.05 was applied to prevent overfitting. All experiments were implemented in PyTorch and were performed on a Nvidia RTX 3090 GPU with a memory of 24G.

4.1. Datasets and Evaluation Metrics

Messidor-1 dataset [6] contains 1200 fundus images with four DR grades. We conducted referral and normal DR classification in this dataset. In the referral and non-referral DR classification, Grades 0 and 1 are considered non-referable, while Grades 2 and 3 are considered referable DR (rDR). For normal and abnormal classification, only Grade 0 will be labeled as normal, and the other grades will be recognized as abnormal. We followed the experimental settings in [15] by using 10-fold cross-validation on the entire dataset. The area under the curve (AUC) was used as the evaluation metric.

Messidor-2 dataset [6] contains 1748 fundus images with five DR grades. As no official split of the training and testing dataset was provided, we used this dataset to conduct ablation studies to demonstrate the effectiveness of each component of our proposed method on DR grading evaluated by the AUC and quadratic Cohen's kappa (Kappa).

RFMiD dataset [24] contains 1920 training, 640 validation, and 640 testing images with 45 different types of pathologies (central serous retinopathy, central retinal vein occlusion, asteroid hyalinosis, etc.). Following the protocol in [14, 38], we performed normal and abnormal binary classification on this dataset whose performance is measured by accuracy (ACC), AUC, and F1.

APOTS dataset [1] contains 3662 fundus images for DR grading with the severity on a grade of 0 to 4 (no DR, mild, moderate, severe, proliferative DR). Following the experimental setting of 5-fold cross-validation in [38], we evaluated the performance of DR grading in terms of ACC, AUC, weighted F1, and kappa.

IDRiD dataset [25] contains 413 training and 103 testing images for both DR grading and DME severity grading tasks. we used the training and testing data provided by the official split. Different from method [15] that re-labeled DR grading into two categories, we trained the multi-class DR grading task and reported the evaluation metrics of ACC, AUC, and F1. Both grading experiments followed the protocol in [4].

MICCAI 2023 MMAC (Myopic Maculopathy Analysis Challenge) contains 1143 fundus images with four myopic maculopathy grades. There are 404 images for grade 0, 412 images for grade 1, 224 images for grade 2,60 images for grade 3, and 43 images for grade 4. We used 5-fold stratified cross-validation on the training set. The Quadratic-weighted Kappa (kappa), F1 score, and Specificity were used as the evaluation metric. For this experiment, we felt the raw data was good quality and did not need to apply any preprocessing.

4.2. Comparison to State-of-the-art Methods

DR task performance. We compared the proposed method to a variety of existing state-of-the-art (SOTA) methods on three DR datasets (i.e., Messidor1, APOTS, and IDRID). The proposed method achieved the best performance on the IDRID dataset (AUC=91.6, ACC=73.1). Remarkably, the proposed method outperformed the best model (DETACH-DAW [4]) by 8% and 23.5% in AUC and ACC (Table 4), respectively, on the IDRID dataset. We also found that the

Table 1. Comparison of rDR and normal classification on the Messidor-1 dataset [6]. Annotations denote whether pixel-level or patch-level supervision was applied. (†: methods implemented by us; while the other benchmarks are taken from [15, 30, 34].)

Method	Annotations	Referral AUC	Normal AUC
VNXK [34]	-	88.7	87.0
CKML [34]	-	89.1	86.2
Comp. CAD [27]	-	91.0	87.6
Expert A [27]	-	94.0	92.2
Expert B [27]	-	92.0	86.5
Zoom-in-Net [36]	-	95.7	92.1
AFN [17]	patch	96.8	-
Semi + Adv [39]	pixel	97.6	94.3
†CANet [15]	-	96.3	-
LAT [30]	-	98.7	96.3
Ours	-	**98.7**	**97.5**

Table 2. Performance comparison of multi-disease abnormal detection on the RFMiD dataset [24]. Param are the parameter numbers, indicating model complexity of models. (Due to some methods codes not being made publicly available, †: methods reproduced by us; while the other benchmarks are taken from [14].)

	Normal/Abnormal		
Method	ACC	AUC	F1
† CANet [15]	88.3	91.0	90.4
† EffNet-B7 [31]	88.2	91.0	90.7
† ReXNet [9]	91.3	94.5	93.3
† CrossFormer-L[35]	90.6	94.3	92.0
† Swin-L [16]	89.5	93.8	91.6
† MIL-VT [38]	91.1	95.9	94.4
SatFormer-B [14]	93.8	96.5	**95.8**
Ours	**94.4**	**98.7**	94.4

Table 3. Performance comparison of DR grading on the APOTS dataset [1]. Due to some methods codes not being made available in public, † denotes methods reproduce performances at the same level as reported while the other benchmarks are taken from [38].

	DR Grading			
Method	AUC	ACC	F1	Kappa
DLI [26]	-	82.5	80.3	89.0
† CANet [15]	-	83.2	81.3	90.0
GREEN-ResNet50 [19]	-	84.4	83.6	90.8
GREEN-SE-ResNext50 [19]	-	85.7	85.2	91.2
† MIL-VT [38]	**97.9**	85.5	85.3	92.0
Ours	97.8	**89.1**	**88.9**	**93.4**

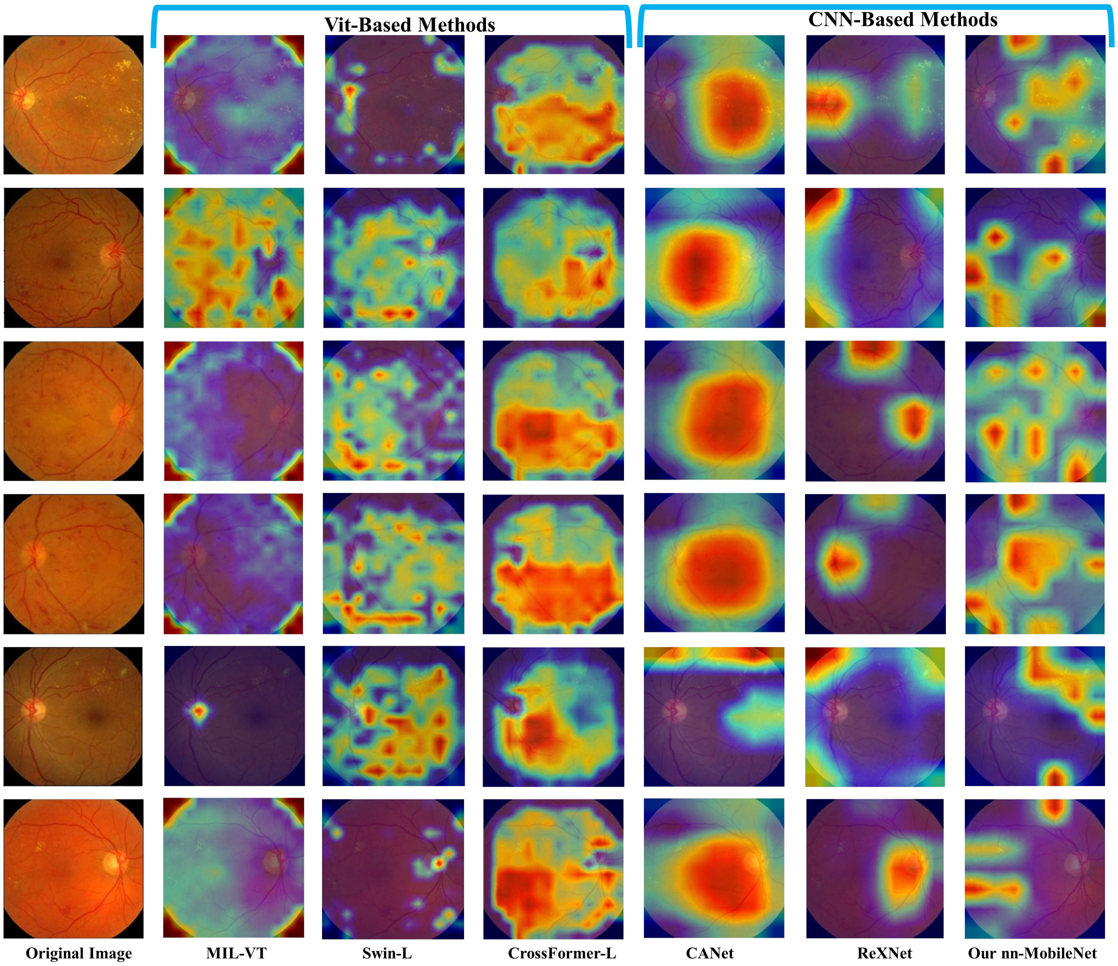


Figure 6. The comparative visualization on the Messidor-1 dataset was performed utilizing CAM [29]. We chose representative CNN/ViT-based methods with publicly available code, including MIL-VT [38], Swin-L [16], CrossFormer-L [35], CANet [15] and ReXNet [9].

Table 4. Performance comparison of DR and DME grading on the IDRID dataset [25]. † denote methods implemented by us while the other benchmarks are taken from [4].

	DME			DR		
Method	AUC	F1	ACC	AUC	F1	ACC
† CANet [15]	87.9	66.1	78.6	78.9	42.3	57.3
Multi-task net [5]	86.1	60.3	74.8	78.0	43.9	59.2
† MTMR-net [18]	84.2	61.1	79.6	79.7	45.3	60.2
† DETACH + DAW [4]	89.5	72.3	82.5	84.8	49.4	59.2
Ours	**95.3**	**84.8**	**86.5**	**91.6**	**72.6**	**73.1**

Table 5. The performance comparison in 2023 MMAC Challenge. The results are tested by the officially challenge platform [2].

Method	Kappa	F1	SPE	Average	CPU time (s)
Rank 1st [13]	90.1	78.1	94.5	87.5	2.1283
Rank 2nd [23]	88.9	76.8	94.1	86.6	0.8047
Ours (can reach 3nd)	90.0	75.1	94.1	86.4	0.2750

proposed method had the highest ACC and Kappa on the APOTS dataset with a similar AUC to the best-performed model (Table 3). The proposed method achieved an equal performance to the best model (LAT [30]) on the referral DR classification task and the best performance on the normal DR classification task (Table 1). It is worth noting that most works (i.e., MIL-VT [38], LAT [30], Zoom-in-Net [36], Semi + Adv [39], and CKML [34]) were pre-trained on large-scale external datasets. Whereas, the proposed method was trained from scratch using the same benchmark datasets.

Multi-disease abnormal detection performance. We also conducted experiments and comparisons to current SOTA methods on the multi-disease detection task. The proposed method achieved the best performance in terms of ACC and AUC, while the SatFormer-B [14] achieved the best performance in F1 (Table 2). However, our model (Param=34M) has fewer than half number of parameters of the SatFormer-

B [14] (Param=78M). Even though the proposed model had a similar stem architecture to the RexNet, our heavy data augmentations and spatial dropout improved the ACC by 3.4% and AUC by 4.4%.

DME classification performance. The DME classification task was evaluated on the IDRID dataset following the protocol in [4]. Table 4 demonstrated that the proposed method surpassed the model with the best performance (DETACH+DAW [4]) by 17.3% on F1, 6.5% on AUC, and 4.8% on ACC. Compared to other SOTA methods (i.e., CANet [15], Multi-task net [5], MTMR-net [18], and DETACH-DAW [4]) that were jointly trained on multiple tasks, our proposed model was trained from scratch on the DME task only.

MICCAI MMAC 2023 Challenge. Our well-calibrated nnMobileNet secured the third position in MICCAI MMAC 2023 Challenge [2] and was remarkably close to the top-ranking models (Table 5). Whereas models that won the first and second places were ViT-based models pre-trined on large-scale external datasets using self-supervised learning. Consequently, their models were at least three times slower than ours' regarding the inference time on CPU (see Table 5).

5. Visual Interpretability

We visualize the most discriminative regions of several representative methods using the gradient-weighted class activation map (Grad-CAM) [29] in the Messidor-1 dataset for the DR task. As shown in Fig. 6, the proposed method showed the most accurate localization of diabetic lesions compared to the other baseline methods, e.g.. hard exudates, and hemorrhages. This observation aligns with our initial hypothesis that ViTs are typically employed to model the similarities between different patches. When dealing with small lesion blocks, localized lesions within many patches tend to be averaged out and overlooked, with ViTs favoring semantic comparisons between patches. Consequently, this leads to methods like Swin-L and CrossFormer producing CAM regions that are overly broad, hindering the precise localization of smaller lesions. It is noteworthy that MIL-VT compels each patch token to pass through a MIL (Multiple Instance Learning) head, essentially engaging in a pseudo-label learning process. We observed that this MIL attention mechanism tends to assign a uniform level of importance to all patch tokens, which disrupts the ability of ViT to learn the relationships between different patches. Compared with CNN methods, the multi-task network of CANet presents fitting challenges, indicating that despite the relatedness of the tasks, DME does not significantly enhance lesion localization in DR, possibly due to divergent interest patterns between the two tasks. Interestingly, The nn-mobilenet and ReXNet share the same model configuration, but the latter still struggles to accurately learn lesion representation. This situation underscores the importance of fine-tuning CNNs for improved performance. Finally, We observed that the ViT-based methods show inferior localization performance compared to CNN-based methods. However, other CNN-based baseline methods (i.e., ReXNet and CANet) only demonstrate coarse localization of the lesions. Whereas, the proposed method can accurately localize diabetic lesions. These findings suggest the importance of CNN in capturing small localized features for retinal disease diagnosis.

6. Discussion and Conclusion

In this article, we center our investigation on the question - *Could CNN inherently be more suited to retinal disease (RD) tasks than ViTs ?* To address this, we embarked on a series of empirical studies, starting with fine-tuning a lightweight MobileNetV2. Through this process, we proposed a series of modifications to MobileNetV2, culminating in developing a tailored and lightweight model we denote as nnMobileNet. The proposed method surpasses ViT-based and multitask-driven models across various RD benchmarks. Remarkably, nnMobileNet achieves this superior performance without applying self-supervised pretraining on external datasets, highlighting the potential of CNNs in the domain of RD tasks.

In revisiting CNNs for RD tasks, we do not entirely negate the value of ViTs. It's evident from our findings that ViTs excel at capturing long-range dependencies better than CNNs. However, ViTs relying on extensive data for pretraining poses significant challenges for medical datasets subject to privacy concerns. Meanwhile, patterns of interest in natural images typically occupy a large portion of the image, and lesions in medical images often constitute a small fraction, making patch-based ViT relational understanding insufficient. Therefore, we offer the following recommendations for future model development in RD tasks: (i) CNNs are preferable in scenarios with limited retinal image data. (ii) CNNs have superior capabilities in capturing fine-grained local features, particularly for RD tasks focused on small lesions. (iii) Integrating CNNs with ViTs could be a viable solution. (iv) Emphasize data characteristics and model fine-tuning. (v) Large-kernel convolutions could address limitations in capturing long-range dependencies. In the end, we believe that results will challenge several widely held views and prompt people to rethink the importance of convolution in RD.

ACKNOWLEDGMENTS The work was partially supported by NIH (R01EY032125, R01DE030286) and the state of Arizona via the Arizona Alzheimer's Consortium, United States.

References

[1] Aptos database. 6

[2] https://codalab.lisn.upsaclay.fr/competitions/12441. 7, 8

[3] GARY C Brown, MELISSA M Brown, TYRIE Hiller, DAVID Fischer, WILLIAM E Benson, and LARRY E Maggargal. Cotton-wool spots. *Retina*, 5(4):206–214, 1985. 3

[4] Haoxuan Che, Haibo Jin, and Hao Chen. Learning robust representation for joint grading of ophthalmic diseases via adaptive curriculum and feature disentanglement. In *MICCAI*, pages 523–533, 2022. 1, 2, 3, 6, 7, 8

[5] Q. Chen and et al. A multi-task deep learning model for the classification of Age-related Macular Degeneration. *AMIA Jt Summits Transl Sci Proc*, 2019. 7, 8

[6] Etienne Decencière and et al. Feedback on a publicly distributed image database: The messidor database. *Image Analysis & Stereology*, 2014. 3, 6

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[8] Matthias Eisenmann, Annika Reinke, Vivienn Weru, Minu D Tizabi, Fabian Isensee, Tim J Adler, Sharib Ali, Vincent Andrearczyk, Marc Aubreville, Ujjwal Baid, et al. Why is the winner the best? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19955–19966, 2023. 3

[9] Dongyoon Han, Sangdoo Yun, Byeongho Heo, and YoungJoon Yoo. Rethinking channel dimensions for efficient model design. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 732–741, 2021. 4, 5, 6, 7

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[11] Byeongho Heo, Sanghyuk Chun, Seong Joon Oh, Dongyoon Han, Sangdoo Yun, Gyuwan Kim, Youngjung Uh, and Jung-Woo Ha. Adamp: Slowing down the slowdown for momentum optimizers on scale-invariant weights. *arXiv preprint arXiv:2006.08217*, 2020. 5

[12] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012. 5

[13] Junlin Hou, Jilan Xu, Fan Xiao, Bo Zhang, Yiqian Xu, Yuejie Zhang, Haidong Zou, and Rui Feng. Towards label-efficient deep learning for myopic maculopathy classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 31–45. Springer, 2023. 3, 7

[14] Yankai Jiang and et al. Satformer: Saliency-guided abnormality-aware transformer for retinal disease classification in fundus image. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 987–994, 2022. 1, 2, 3, 4, 6, 7, 8

[15] X. Li, X. Hu, L. Yu, L. Zhu, C. W. Fu, and P. A. Heng. CANet: Cross-Disease Attention Network for Joint Diabetic Retinopathy and Diabetic Macular Edema Grading. *IEEE Trans Med Imaging*, pages 1483–1493, 2020. 1, 2, 3, 4, 6, 7, 8

[16] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*, 2021. 1, 2, 6, 7

[17] Zhiwen Lin, Ruoqian Guo, Yanjie Wang, Bian Wu, Tingting Chen, Wenzhe Wang, Danny Z. Chen, and Jian Wu. A framework for identifying diabetic retinopathy based on anti-noise detection and attention-based fusion. In *MICCAI*, pages 74–82, Cham, 2018. Springer. 1, 2, 6

[18] L. Liu and et al. Multi-Task Deep Model With Margin Ranking Loss for Lung Nodule Analysis. *IEEE Trans Med Imaging*, 39(3):718–728, 2020. 7, 8

[19] Shaoteng Liu, Lijun Gong, Kai Ma, and Yefeng Zheng. Green: a graph residual re-ranking network for grading diabetic retinopathy. In *MICCAI*, pages 585–594, Cham, 2020. Springer International Publishing. 6

[20] Ze Liu and et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. IEEE Int. Conf. Comput. Vis.(ICCV)*, pages 10012–10022, 2021. 2

[21] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*, pages 11976–11986, 2022. 2, 4, 5

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*. 5

[23] Li Lu, Xuhao Pan, Panji Jin, and Ye Ding. Swin-mmc: Swin-based model for myopic maculopathy classification in fundus images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 18–30. Springer, 2023. 3, 7

[24] Samiksha Pachade and et al. Retinal fundus multi-disease image dataset (rfmid). 2020. 6

[25] Prasanna Porwal and et al. Indian diabetic retinopathy image dataset (idrid). 2018. 6, 7

[26] Alexander Rakhlin. Diabetic retinopathy detection through integration of deep learning classification framework. *BioRxiv*, page 225508, 2017. 6

[27] Clara I Sánchez and et al. Evaluation of a computer-aided diagnosis system for diabetic retinopathy screening on public data. *Investigative ophthalmology & visual science*, 52(7): 4866–4871, 2011. 1, 2, 6

[28] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pat- tern Recognit*, pages 4510–4520, 2018. 2, 3

[29] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 7, 8

[30] Rui Sun, Yihao Li, Tianzhu Zhang, Zhendong Mao, Feng Wu, and Yongdong Zhang. Lesion-aware transformers for diabetic retinopathy grading. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pat- tern Recognit*, pages 10938–10947, 2021. 1, 2, 4, 6, 7

[31] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 2019. 6

[32] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656, 2015. 5

[33] Enes Sadi Uysal, M Şafak Bilici, B Selin Zaza, M Yiğit Özgenç, and Onur Boyar. Exploring the limits of data augmentation for retinal vessel segmentation. *arXiv preprint arXiv:2105.09365*, 2021. 4

[34] Holly H. Vo and Abhishek Verma. New deep neural nets for fine-grained diabetic retinopathy recognition on hybrid color space. In *2016 IEEE International Symposium on Multimedia (ISM)*, pages 209–215, 2016. 6, 7

[35] Wenxiao Wang, Lu Yao, Long Chen, Binbin Lin, Deng Cai, Xiaofei He, and Wei Liu. Crossformer: A versatile vision transformer hinging on cross-scale attention. *arXiv preprint arXiv:2108.00154*, 2021. 1, 6, 7

[36] Zhe Wang, Yanxin Yin, Jianping Shi, Wei Fang, Hongsheng Li, and Xiaogang Wang. Zoom-in-net: Deep mining lesions for diabetic retinopathy detection. In *MICCAI*, pages 267–275, 2017. 1, 2, 6, 7

[37] David Yorston. Retinal diseases and vision 2020. *Community Eye Health*, 16(46):19–20, 2003. 1

[38] Shuang Yu and et al. Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification. In *MICCAI*, pages 45–54. Springer, 2021. 1, 2, 3, 6, 7

[39] Yi Zhou, Xiaodong He, Lei Huang, Li Liu, Fan Zhu, Shanshan Cui, and Ling Shao. Collaborative learning of semi-supervised segmentation and classification for medical images. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*, 2019. 1, 2, 6, 7

[40] Wenhui Zhu and et al. Self-supervised equivariant regularization reconciles multiple instance learning: Joint referable diabetic retinopathy classification and lesion segmentation. *18th International Symposium on Medical Information Processing and Analysis (SIPAIM)*, 2022. 1

[41] Wenhui Zhu, Peijie Qiu, Xiwen Chen, Huayu Li, Hao Wang, Natasha Lepore, Oana M Dumitrascu, and Yalin Wang. Beyond mobilenet: An improved mobilenet for retinal diseases. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 56–65. Springer, 2023. 3

[42] Wenhui Zhu, Peijie Qiu, Oana M Dumitrascu, Jacob M Sobczak, Mohammad Farazi, Zhangsihao Yang, Keshav Nandakumar, and Yalin Wang. Otre: Where optimal transport guided unpaired image-to-image translation meets regularization by enhancing. In *International Conference on Information Processing in Medical Imaging*, pages 415–427. Springer, 2023. 1

[43] Wenhui Zhu, Peijie Qiu, Mohammad Farazi, Keshav Nandakumar, Oana M Dumitrascu, and Yalin Wang. Optimal transport guided unsupervised learning for enhancing low-quality retinal images. *arXiv preprint arXiv:2302.02991*, 2023. 1