Multi-Task Learning for Diabetic Retinopathy Grading and Lesion Segmentation

Alex Foo,[1] Wynne Hsu,[1] Mong Li Lee,[1] Gilbert Lim,[1,2] Tien Yin Wong[2]
[1]School of Computing [2]Duke-NUS Medical School
National University of Singapore

Abstract

Although deep learning for Diabetic Retinopathy (DR) screening has shown great success in achieving clinically acceptable accuracy for referable versus non-referable DR, there remains a need to provide more fine-grained grading of the DR severity level as well as automated segmentation of lesions (if any) in the retina images. We observe that the DR severity level of an image is dependent on the presence of different types of lesions and their prevalence. In this work, we adopt a multi-task learning approach to perform the DR grading and lesion segmentation tasks. In light of the lack of lesion segmentation mask ground-truths, we further propose a semi-supervised learning process to obtain the segmentation masks for the various datasets. Experiments results on publicly available datasets and a real world dataset obtained from population screening demonstrate the effectiveness of the multi-task solution over state-of-the-art networks.

Introduction

Automated screening for Diabetic Retinopathy (DR) using deep learning has achieved significant progress and demonstrated excellent performance in terms of having accurate classifications for referable versus non-referable DR (Ting et al. 2019). The top performing deep learning networks have obtained scores of above 90% for the area under the Receiver Operating Characteristics curve (ROC) for moderate and above DR severity level (Sayres et al. 2019; Ting et al. 2017). However, there remains a need to provide more fine-grained grading of the DR severity level.

A diabetic retina image can be graded into different levels of severity according to the International Clinical Diabetic Retinopathy Scale (ICDRS) (Gulshan et al. 2016). There are 5 levels in this scale: 0 (no apparent DR), 1 (mild DR), 2 (moderate DR), 3 (severe DR), 4 (poliferative DR). Figure 1 shows sample retinal images and their corresponding DR severity level based on ICDRS. Learning a model to classify retina images into the various groups is a challenge due to limited labeled data. This is further worsened by the situation where the labels can be noisy as a result of the subjective labeling by different human graders.

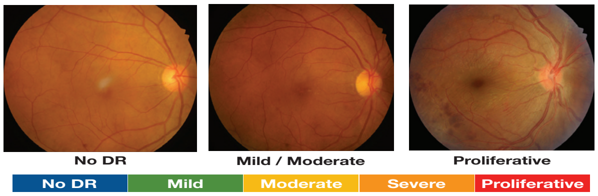

Figure 1: Sample retina images and their ICDRS grade taken from (Gulshan et al. 2016).

We observe that human graders examine the retina image for the presence of one or more lesions such as microaneurysms, hemorrhages, hard exudates and soft exudates before they assign the DR severity level to the image. The ability to automate the segmentation of lesions in a retina image can facilitate this grading process, and serve to validate the DR severity level assigned by the grader. In this work, we explore the feasibility of adopting a multi-task learning approach to perform these two closely related tasks: DR grading and lesion segmentation. Specifically, we extend the well-known UNet architecture (Ronneberger, Fischer, and Brox 2015) by replacing its encoder with a classic VGG-16 network (Simonyan and Zisserman 2014). This creates a dual-stream network, where the VGG-16 encoder outputs the DR grading and the decoder outputs the lesion segmentation. Both streams thus complement each other. The pixel level lesion segmentations provide justification for the DR severity level grading, while the severity level grading guides the lesion segmentation. Additionally, the streams regularize each other during training which prevent overfitting of the individual tasks.

Obtaining precise pixel level annotation of abnormalities associated with DR like microaneurysms, soft exudates, hard exudates and hemorrhages is labour intensive. Given the limited availability of this invaluable resource for model learning and performance evaluation, we propose a semi-supervised approach for effective multi-task learning. Experiment results on several benchmark datasets demonstrate that our approach leads to a significant performance boost over existing networks, notably in differentiating retina images having moderate DR from mild DR.

Related Work

Traditional DR screening systems typically first identify regions of interests for feature extraction before using the extracted features for classification (Lim et al. 2014). Subsequently, advances in deep learning enabled the entire retina image to be passed as input through some neural network which classifies whether the image has referable DR (Zhou et al. 2018; Ting et al. 2017).

The lack of interpretability of such deep learning systems for DR screening has led to research in semantic segmentation which classify pixels into various retinal features, namely - optic disc, microaneurysms, haemorrahges, soft and hard exudates (Saha, Sathish, and Sheet 2019; Yan et al. 2019). More recent developments explore the utilization of multi-task learning for deep neural networks. This is achieved by having two outputs at the end of a network, each with a different loss function (Zhou et al. 2018). A weighted sum of the losses is then used for training.

The work in (Samala et al. 2017) investigates the use of auxiliary tasks. This involves having outputs at various points of the network with the goal of sharpening that section of the network. Specific loss functions, which are more characteristic of that section and its output task, are defined and added to the overall training loss.

Our proposed approach combines multi-task learning with auxiliary tasks by using the VGG16 encoder section to output the image classification while the decoder section outputs the image segmentation. Loss functions are defined for each task and combined for training. More details will be elaborated in the next section.

Proposed Network

Our proposed network is inspired by two widely used deep learning architectures - UNet and VGG-16. VGG-16 has shown great performance for both general image classification and medical image grading tasks. Using VGG-16 enables us to leverage on its feature extraction strengths to obtain the grading classification of retinal images. Further, the availability of pretrained weights of VGG16 on ImageNet and Places365 datasets allows us to leverage on the successful improvements that transfer learning has displayed for image classification and segmentation.

The simplest and most commonly employed use of multi-task learning has an added output beside the original output. In the case of a UNet architecture, this involves adding a DR grading output at the end of the decoder section of the network as shown in Figure 2.

Another way is to have a two-step network where the UNet outputs the segmentation mask which is then used to process the grading prediction (see Figure 3). However, from a training perspective, this is not ideal because of the need to separate the training of the network into two steps - first for the segmentation output, then for DR grading output.

Figure 4 describes a network by which the encoder outputs a grading prediction, while the decoder outputs the usual segmentation prediction. This network is based on the functional similarities between VGG-16 and the encoder portion of UNet. We replace the encoder portion with a VGG-16 network, and this VGG-16 encoder section consequently outputs the DR grading classification of the retina image, while the decoder utilises the features extracted by the VGG-16 network to obtain the lesion segmentation result. Figure 5 gives the details of each layer in this network.

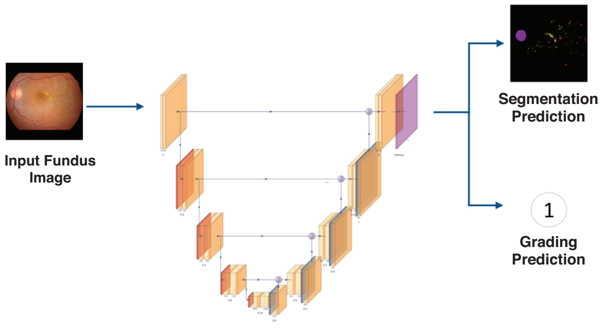

Figure 2: Variant A. Multiple outputs at the decoder section of UNet.

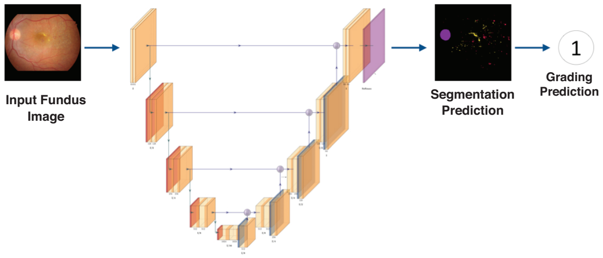

Figure 3: Variant B. Two-step network.

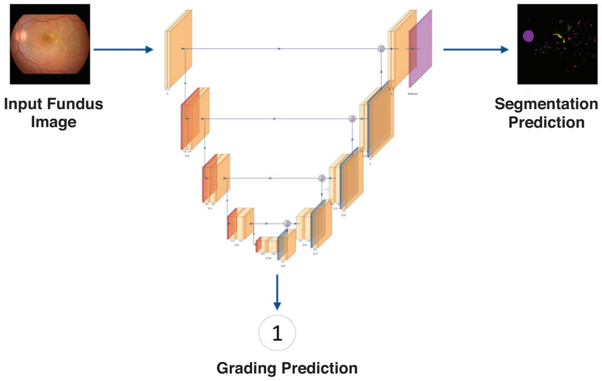

Figure 4: Variant C. Proposed network.

Encoder Section

As mentioned, we replaced the conventional UNet encoder with a VGG16 network, passing skip connections to the decoder section before every Max-Pool operation. At the end of the 5th VGG-16 convolutional block, the fully connected layers are replaced by a Global-Max-Pooling operation to extract the final features before proceeding to complete the image classification task with the Dense operation.

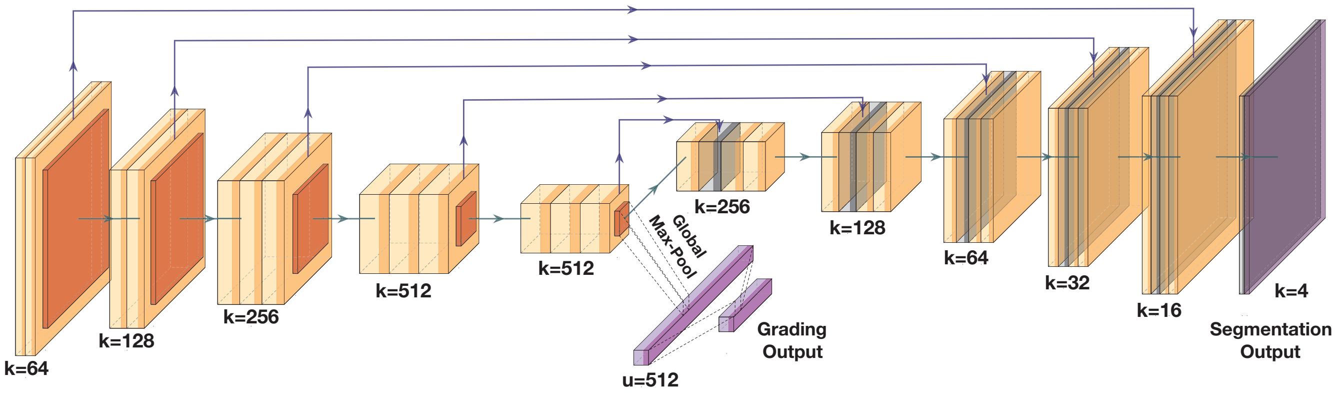

Figure 5: Details of the layers in the proposed network, where k and u is the number of kernels and dense units respectively.

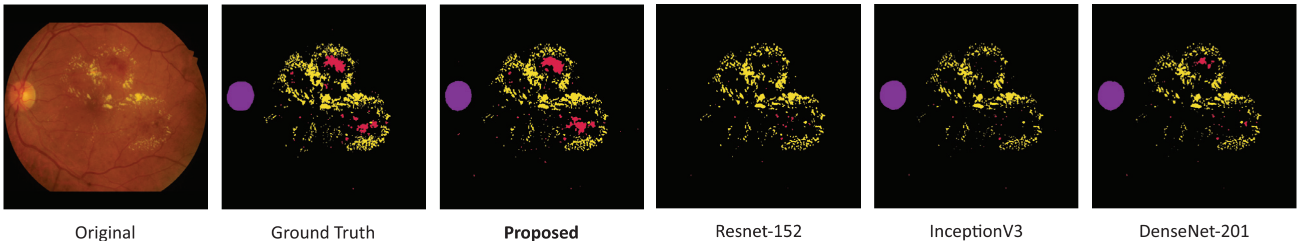

Figure 6: Retinal image segmentation by different encoder architectures. Red and yellow markings indicate red lesions and yellow lesions respectively, while purple markings indicate optic disc.

We have also considered and compared alternative architectures such as ResNet (He et al. 2016), InceptionV3 (Szegedy et al. 2016) and DenseNet (Huang et al. 2017). However, despite outperforming VGG-16 in most image classification tasks, notably in (Russakovsky et al. 2015), these networks fair poorly when implemented as the encoder network for image segmentation tasks.

Figure 6 shows the segmentations by the various encoder architectures. We observe the ResNet and InceptionV3 do not pick up the red lesions, while DenseNet detects only partial regions that correspond to the red lesions. This is probably due to the use of batchnorm operation in these alternative networks that produces a less accurate representation of features for this task.

Decoder Section

We introduce three enhancements to the conventional UNet decoder section in our proposed network. First, we use transposed convolutions (Dumoulin and Visin 2016) instead of the usual up-sampling approaches. Transposed convolutions has been shown to be a more accurate means of decoding the features extracted for image segmentation tasks, achieved by mathematically transposing the initial convolutional kernel. Second, the decoder is modified to mirror that of a VGG-16 network, which provides better decoding. Finally, we replace the last sigmoid activation function with a softmax operation since a pixel cannot belong to multiple classes. Our experiments show that these enhancements lead to notable improvements for the lesion segmentation task.

Loss Functions

Our proposed network utilizes two loss functions from the segmentation task and the DR grading task which provide mutually beneficial sharpening and regularization of the network weights during backpropagation. Specifically, the loss function of the segmentation task signals back to the entire network (encoder and decoder) its segmentation loss between the predicted result and the ground truth, while the loss function of the grading task signals back to the encoder section its grading classification loss. Moreover, the network is refined through the direct relation that both loss signals have on each other - an image with a DR severity level of 1 only has red lesions, while an image with a DR severity level of 4 has both red and yellow lesions.

The lesion segmentation task outputs a prediction S of dimension h x w x 5 from an input image of dimension h x w x c, where h and w are the height and width of the input image, c is the number of channels and we have a total of 5 DR classes. The vector S is then compared against the ground-truth segmentation mask Y. For this, we use the binary cross-entropy loss as shown in Equation (1):

$$\begin{aligned} L_B(S,Y) = -\sum_{i=1}^{h}\sum_{j=1}^{w}\sum_{k=1}^{5} Y_{i,j,k}\log S_{i,j,k} \\ + (1-Y_{i,j,k})\log(1-S_{i,j,k}) \end{aligned} \tag{1}$$

Note that although previous work has shown success for using Dice-loss or Jaccard-loss to improve image segmentation results due to their higher sensitivity to class imbalanced

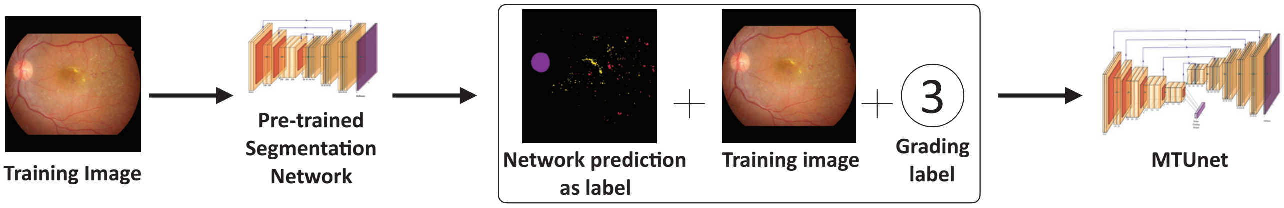

Figure 7: Semi-supervised training process.

Table 1: Performance of the network variants for the lesion segmentation and DR grading classification tasks

	Dice (%)	ROC (%)
Variant A	55.37	86.27
Variant B	73.89	82.05
Variant C	**76.12**	**89.45**

Table 2: Dice score of networks for lesion segmentation

Network	Red lesions	Yellow lesions
UNet	62.25 ±7.646	76.03 ±1.415
LinkNet	73.56 ±1.026	83.54 ±0.19
Feature Pyramid	72.66 ±0.541	77.28 ±0.429
DeepLabv3	56.73 ±1.217	72.01 ±1.661
MTUnet	**75.15 ±0.778**	**84.27 ±0.456**

datasets, however, our initial experiments involving various combinations of either Dice Loss, generalized Dice Loss or Jaccard loss did not yield good performance for our segmentation task.

The DR grading task outputs a prediction vector P, in one-hot encoding format for the 5 DR classes. P is then compared against the grading ground-truth G. We use the multi-category cross-entropy loss as shown in Equation (2):

$$L_C(P,G) = -\sum_{k=1}^{5} G_{i,j,k,} \log P_{i,j,k} \tag{2}$$

Finally, we weight and combine these two losses above into a multi-task network loss L_M as follows:

$$L_M = L_B + \alpha L_C \tag{3}$$

where α is a hyper-parameter. We experimented with different values of α and found that α = 1 gives the best result for both tasks.

Semi-Supervised Training

Given the limited number of images that have ground-truth labels for both retinal grading and segmentation mask, we propose a semi-supervised approach to increase the number of images to train the proposed network. Figure 7 shows the training processs.

We first use a small dataset where the ground truth segmentation mask is available to obtain an initial network for the lesion segmentation task. With this initial network, we obtain the segmentation masks for the training dataset where the images are only labeled with the DR severity levels. Note that if an image in the training dataset has a severity level of 0, then we ignore the red and yellow lesion outputs from the initial network. We also increase the accuracy of this initial network by applying data augmentation using random rotations, width/height shifts, shear, zoom, vertical/horizontal flips and channel shifts on-the-fly during training.

Performance Study

We implemented the three network variants in Keras and carried out experiments to evaluate their performance. We adopt Adam algorithm to optimise the networks, with an initial learning rate of 1e-5. The input size of each network is 512 x 512 x 3. ImageNet pretrained weights are used as the initial weights of the encoder section, while weights of the decoder section are initialized by the He Normal algorithm. A 4X Tesla V100 GPU server is utilized to train the network. We use the following datasets in our experiments.

1. Indian Diabetic Retinopathy Image Segmentation Dataset (IDRiD-S) (Porwal et al. 2018). This is a publicly available dataset with 81 retinal images for the segmentation task where each image has a manually labelled segmentation mask. We split these images into training and test sets of 54 images and 27 images respectively.
2. Indian Diabetic Retinopathy Image Dataset (IDRiD-D) (Porwal et al. 2018). This publicly available dataset contains 413 images which have been manually classified into the 5 DR severity levels.
3. Kaggle DR Competition Dataset[1] (Kaggle). This dataset consists of 35126 training images, 10906 validation images and 42670 test images provided by EyePACS[2], a free platform for DR screening. Each image has a manually labelled grading of 0 to 4. For the training set of the DR grading classification task, we combine the training and test images and downsample the images by the number of images with DR severity level of 4. This gives us 1914 images per severity level.
4. Singapore National Diabetic Retinopathy Screening Program (SiDRP14-15) (Ting et al. 2017). This dataset consists of retinal images obtained from the DR screening of a multi-ethnic population with ground-truth labels for

[1] https://www.kaggle.com/c/diabetic-retinopathy-detection

[2] http://www.eyepacs.com/data-analysis

Table 3: ROC (%) of networks for the different DR severity levels on IDRiD

Network	No DR	Mild	Moderate	Severe	Proliferative
NASNet	93.00 ±0.784	79.07 ±3.881	71.75 ±5.342	72.63 ±4.239	83.95 ±4.729
DenseNet-201	62.25 ±13.49	38.64 ±5.872	40.49 ±5.590	36.98 ±6.060	36.83 ±12.78
ResNeXt-101	92.09 ±3.223	79.28 ±4.391	67.22 ±2.499	72.26 ±9.102	66.62 ±4.879
InceptionResNetV2	94.86 ±3.286	77.92 ±10.03	67.91 ±13.00	75.74 ±5.868	67.17 ±14.09
MTUnet	**99.00 ±0.330**	**89.94 ±2.093**	**84.13 ±3.011**	**82.86 ±3.110**	**87.20 ±2.859**

Table 4: ROC (%) of networks for the different DR severity levels on SiDRP14-15

Network	No DR	Mild	Moderate	Severe	Proliferative
NASNet	61.37 ±0.569	47.14 ±1.547	59.81 ±3.224	81.88 ±8.410	87.64 ±0.538
DenseNet-201	51.00 ±3.190	48.11 ±2.053	43.44 ±9.126	57.47 ±16.385	66.33 ±14.374
ResNeXt-101	55.55 ±3.921	52.10 ±1.461	58.69 ±4.049	75.59 ±12.481	42.76 ±5.103
InceptionResNetV2	66.79 ±4.272	46.84 ±5.985	59.16 ±10.337	83.23 ±7.013	83.52 ±2.653
MTUnet	**78.56 ±0.122**	**58.44 ±0.606**	**84.75 ±1.875**	**97.77 ±0.745**	**98.11 ±0.867**

grading classification. This dataset consists of 71896 images from 14880 patients.

We evaluate the effectiveness of the networks for both the lesion segmentation and DR grading tasks. For the lesion segmentation task, let X be the set of pixels extracted, and Y be the set of annotated pixels in the ground truth. We use the Dice similarity coefficient which measures the similarity between the two sets X and Y as follows:

$$Dice(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}$$

For the DR grading task, we use the area under the Receiver Operating Characteristic (ROC) curve to measure the classification accuracy for the different severity levels.

Experiments on Network Variants

We first evaluate the three network variants. Each network is trained for 50 epochs using the downsampled Kaggle dataset with segmentation labels obtained from the semi-supervised training procedure using IDRiD-S as the training set.

Table 1 shows the results. We see that variant C outperforms the other two networks for both lesion segmentation and DR grading. As such, we will use variant C as the network for the rest of our experiments. We call our proposed network MTUnet.

Experiments on Lesion Segmentation

In this set of experiments, we compare the performance of our proposed MTUnet for the lesion segmentation task with the following networks:

1. UNet (Ronneberger, Fischer, and Brox 2015). This is the original UNet architecture and serves as our baseline.
2. LinkNet (Chaurasia and Culurciello 2017). This is similar to UNet except that concatenation is replaced by addition for skip connections.
3. Feature Pyramid Network (Lin et al. 2017). This network extracts features of the image at different scales and combines them into a joint prediction.
4. DeepLabv3 (Chen et al. 2017). This approach uses Atrous convolutions to retain context.

We train all these methods using the 54 training images in IDRiD-S, and test them on 27 testing images. For MTUnet, we use the semi-supervised process in Figure 7 to obtain the DR severity level for the images in IDRiD-S.

Table 2 shows the average results of three training runs. We observe that our MTUnet outperforms all the other networks for both the segmentation of red and yellow lesions. Notably, under the paired t-test at 10% significance level, the improvement of MTUnet over LinkNet is statistically significant, suggesting that having a DR grading as an auxiliary task helps to sharpen the lesion segmentation task.

Experiments on DR Grading Classification

Finally, we evaluate the grading classification performance of our proposed MTUnet. We compare MTUnet with the following networks:

1. NASNet (Zoph et al. 2018). NasNet leverages on the use of reduction cells which are convolutional cells that return a feature map where the feature map height and width is reduced by a factor of two.
2. DenseNet-201 (Huang et al. 2017). This network connects each layer to every other layer in a feed-forward fashion.
3. ResNeXt-101 (He et al. 2016). This is an improved variation of ResNet where neurons at one path are no longer connected to the neurons at other paths.
4. InceptionResNetV2 (Szegedy et al. 2017). This is an improvement from InceptionV3 by harnessing the effectiveness of residual connections as in ResNet.

All the networks are trained using the down-sampled Kaggle dataset. Since the images in Kaggle do not have the ground truth lesion segmentation, we use the semi-supervised process to obtain the segmentation mask labels for these images. We test the networks on the IDRiD-D and SiDRP14-15 datasets.

Table 3 shows the ROC scores of the various networks for the IDRiD-D dataset. We observe that MTUnet is able to achieve higher scores for all severity levels as compared to the alternative networks. Notably, MTUnet obtains laudable scores of at least 82.86% and at most 99% across all severity levels.

Table 4 gives The ROC scores of the various networks for the SiDRP14-15 dataset. Again, we see that MTUnet outperforms all the alternative networks for all severity levels. Importantly, while the alternative networks fare poorly in identifying moderate DR images, MTUnet performs well with an ROC score of 84.75%, which is at least a 24% performance difference. This is significant because from a clinical perspective, moderate DR triggers the commencement of medical treatment.

Conclusion

In this work, we have described a multi-task deep learning system called MTUnet which leverages on the close relation between DR grading and lesion segmentation tasks. Specifically, we have proposed an extended UNet architecture with VGG-16 substituted as the encoder section of the network. This network accepts a retinal fundus image input and outputs both the DR grading and the lesion segmentation. To deal with the lack of lesion segmentation ground-truth masks, we have additionally proposed a semi-supervised approach to obtain the segmentation masks required for multi-task training. Experiment results on various benchmark datasets demonstrate the effectiveness of our multi-task approach over state-of-the-art networks. Crucially, the 25% performance difference obtained in the identification of moderate DR on SiDRP14-15 illustrates the clinical significance of our approach.

Acknowledgements

This research is supported by the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG-RP-2018-008).

References

Chaurasia, A., and Culurciello, E. 2017. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *IEEE Visual Communications and Image Processing (VCIP)*.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Dumoulin, V., and Visin, F. 2016. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*.

Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M. C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* 316(22):2402–2410.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *CVPR*.

Lim, G.; Lee, M. L.; Hsu, W.; and Wong, T. Y. 2014. Transformed representations for convolutional neural networks in diabetic retinopathy screening. In *AAAI Workshop on Modern Artificial Intelligence for Health Analytics (MAIHA)*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *CVPR*.

Porwal, P.; Pachade, S.; Kamble, R.; Kokare, M.; Deshmukh, G.; Sahasrabuddhe, V.; and Meriaudeau, F. 2018. Indian diabetic retinopathy image dataset (idrid). *IEEE Dataport*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3).

Saha, O.; Sathish, R.; and Sheet, D. 2019. Fully convolutional neural network for semantic segmentation of anatomical structure and pathologies in colour fundus images associated with diabetic retinopathy. *arXiv preprint arXiv:1902.03122*.

Samala, R. K.; Chan, H.-P.; Hadjiiski, L. M.; Helvie, M. A.; Cha, K. H.; and Richter, C. D. 2017. Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms. *Physics in Medicine & Biology* 62(23).

Sayres, R.; Taly, A.; Rahimy, E.; Blumer, K.; Coz, D.; Hammel, N.; Krause, J.; Narayanaswamy, A.; Rastegar, Z.; Wu, D.; et al. 2019. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology* 126(4).

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *CVPR*.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Ting, D. S. W.; Cheung, C. Y.-L.; Lim, G.; Tan, G. S. W.; Quang, N. D.; Gan, A.; Hamzah, H.; Garcia-Franco, R.; San Yeo, I. Y.; Lee, S. Y.; et al. 2017. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *Jama* 318(22):2211–2223.

Ting, D. S. W.; Pasquale, L. R.; Peng, L.; Campbell, J. P.; Lee, A. Y.; Raman, R.; Tan, G. S. W.; Schmetterer, L.; Keane, P. A.; and Wong, T. Y. 2019. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology* 103(2):167–175.

Yan, Z.; Han, X.; Wang, C.; Qiu, Y.; Xiong, Z.; and Cui, S. 2019. Learning mutually local-global u-nets for high-resolution retinal lesion segmentation in fundus images. *arXiv preprint arXiv:1901.06047*.

Zhou, K.; Gu, Z.; Liu, W.; Luo, W.; Cheng, J.; Gao, S.; and Liu, J. 2018. Multi-cell multi-task convolutional neural networks for diabetic retinopathy grading. In *IEEE Engineering in Medicine and Biology Society (EMBC)*, 2724–2727.

Zoph, B.; Vasudevan, V.; Shlens, J.; and Le, Q. V. 2018. Learning transferable architectures for scalable image recognition. In *CVPR*.