Detection of exudates in fundus photographs with imbalanced learning using conditional generative adversarial network

RUI ZHENG,[1,5] **LEI LIU,** [2,5] **SHULIN ZHANG,**[1] **CHUN ZHENG,**[3] **FILIZ BUNYAK,**[4] **RONALD XU,**[1] **BIN LI,**[2] **AND MINGZHAI SUN**[1,*]

[1]*Department of Precision Machinery and Instrumentation, University of Science and Technology of China, Hefei, Anhui 230022, China*
[2]*Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, Anhui 230022, China*
[3]*The 105 Hospital of PLA, Hefei, Anhui 230031, China*
[4]*Department of Computer Science, University of Missouri, Columbia, MO 65211, USA*
[5]*These two authors contribute equally to the study*
*mingzhai@ustc.edu.cn

Abstract: Diabetic retinopathy (DR) is a leading cause of blindness worldwide. However, 90% of DR caused blindness can be prevented if diagnosed and intervened early. Retinal exudates can be observed at the early stage of DR and can be used as signs for early DR diagnosis. Deep convolutional neural networks (DCNNs) have been applied for exudate detection with promising results. However, there exist two main challenges when applying the DCNN based methods for exudate detection. One is the very limited number of labeled data available from medical experts, and another is the severely imbalanced distribution of data of different classes. First, there are many more images of normal eyes than those of eyes with exudates, particularly for screening datasets. Second, the number of normal pixels (non-exudates) is much greater than the number of abnormal pixels (exudates) in images containing exudates. To tackle the small sample set problem, an ensemble convolutional neural network (MU-net) based on a U-net structure is presented in this paper. To alleviate the imbalance data problem, the conditional generative adversarial network (cGAN) is adopted to generate label-preserving minority class data specifically to implement the data augmentation. The network was trained on one dataset (e_ophtha_EX) and tested on the other three public datasets (DiaReTDB1, HEI-MED and MESSIDOR). CGAN, as a data augmentation method, significantly improves network robustness and generalization properties, achieving F1-scores of 92.79%, 92.46%, 91.27%, and 94.34%, respectively, as measured at the lesion level. While without cGAN, the corresponding F1-scores were 92.66%, 91.41%, 90.72%, and 90.58%, respectively. When measured at the image level, with cGAN we achieved the accuracy of 95.45%, 92.13%, 88.76%, and 89.58%, compared with the values achieved without cGAN of 86.36%, 87.64%, 76.33%, and 86.42%, respectively.

1. Introduction

According to 2016 WHO report, from 1980 to 2014 the number of adults living with diabetes has risen from 108 million to 422 million with the global prevalence increasing from 4.7% to 8.5% [1]. Among many complications that diabetes leads to, diabetic retinopathy (DR) is a significant cause of blindness. In 2010, DR caused 2.6% of blindness [2]. DR occurs as a result of long-term accumulated damages to the small blood vessels in the retina. After 20 years of diabetes, nearly all patients with Type I diabetes and >60% of patients with Type II diabetes have some degree of retinopathy [3]. At the early stage of DR, patients may not have any symptoms of vision problems. However, when it develops to the late stage, it may permanently cause vision

loss and blindness. Therefore it is important for diabetic patients to have a comprehensive retina screening at least once a year. It is shown that blindness due to DR can be prevented in 90% of the cases by early detection through regular screening [4].

Fundus photography is the most commonly used and effective way of screening for diabetic retinopathy. Ophthalmologists look for early signs of DR such as exudates. Exudates are lipid and lipoprotein deposits that appear near leaking capillaries within the retina. They develop at the early stage of DR and may appear as yellow areas with variable sizes from a few pixels to as large as the optic disc (Fig. 1).

Fundus images are usually examined by ophthalmologists. However, due to the limited number of the eye doctors and the massive screening population, many computer-aided diagnosis systems (CAD) have been developed to automatically detect the typical pathological signs of DR in the hope of improving the screening efficiency and releasing the expensive medical resources. The sensitivity and accuracy of the CADs are critical for the early diagnosis. However, the low contrast, irregular shapes, and sparsity of the early signs of DR pose enormous challenges to the analysis methods.
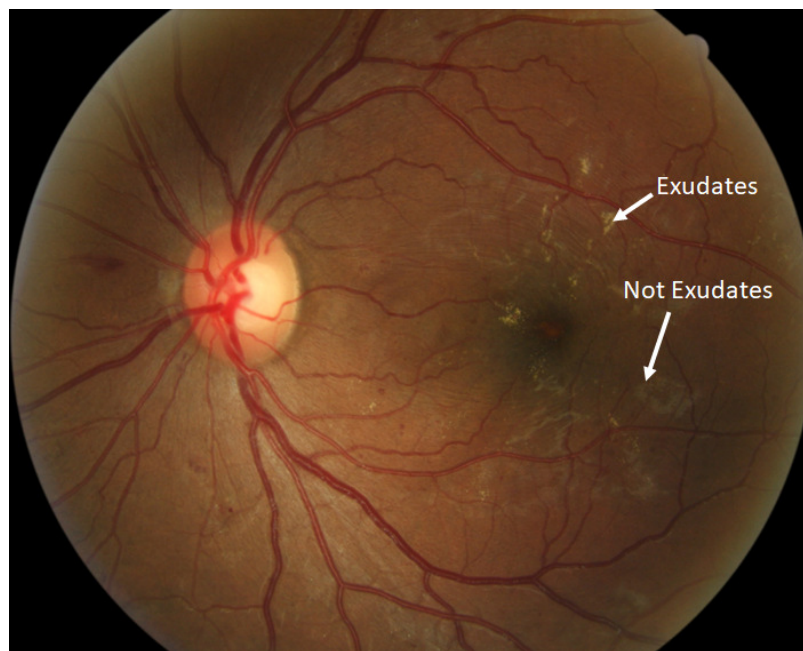

Fig. 1. Retinal image with exudates.

Deep convolutional neural networks have been proven to outperform conventional image analysis methods in many aspects, especially because they do not require explicit feature extraction. Recently, deep learning methods were applied to detect exudates in color fundus photographs and showed promising results although the highest reported sensitivity and accuracy were still lower than those using the state-of-the-art conventional image analysis methods [5–8]. For example, the deep learning method in [5] obtained a maximum sensitivity of 81.35% while the conventional method in [8] achieved a sensitivity of 92.42% both on the DIARETDB1 dataset [9]. The performance of the deep learning methods is mainly limited by two challenges. One is the limited expert labeled training data to train network, and the other is the severely imbalanced representations of different classes.

Limited expert-labeled data is a general problem in the medical image analysis field since it

is usually costly to acquire the ground truth data. Imbalanced dataset manifests itself in two aspects. One is that there are much fewer images of pathological conditions than those of normal conditions. This is particularly true for datasets acquired during the massive screening, such as the DR screening. In these datasets majority of the images do not contain any pathological conditions. The other aspect of the imbalanced data is that the number of abnormal pixels is much less than that of the normal pixels. For example, in the image with only a few exudates, more than 98% of the pixels are normal with about 2% are abnormal pixels. This severely imbalanced representation of different classes renders great challenge for constructing a robust and reliable neural network with good generalization property.

To address the problems discussed hereinbefore, we developed an ensemble deep convolutional neural network (MU-net) based on the U-net framework that was mainly designed for medical applications with limited ground truth data. Furthermore, to solve the imbalanced class problem, we applied the conditional generative adversarial network (cGAN) to augment the dataset and specifically upsample the minority class. Our method significantly outperformed previous methods and had good generalization property over different datasets.

The paper is organized as follows. In section 2, we provide a short overview of different approaches used for exudate detection and discuss methods for imbalanced dataset learning. In section 3, we present detailed explanations of our approach. We illustrate the results in section 4 with discussions and end the paper in section 5 with a short conclusion.

2. Related works

Automatic exudate detection in fundus images has been studied before, and many methods have been developed, including traditional image processing methods and recent approaches based on machine learning and neural networks. Thresholding methods based on global or local image gray-level [10–13], and clustering-based image thresholding, such as Otsu thresholding [11, 13], have been applied for exudate segmentation. Methods based on morphological operations [8, 14–18] have also been explored. In these methods, dominant structures, such as blood vessels and optic disc, were usually first identified and removed from the images to reduce the interference with the exudate segmentation. Region growing method [19–22] has also been demonstrated for automatic segmentation of exudates, particularly in combination with the artificial neural network [22].

Machine learning based methods, such as linear discriminant classifiers [23, 24], support vector machine (SVM) [25, 26], Naive Bayes classifier [27] and random forest algorithm [28], have been studied. Most of these methods consisted of building a feature vector for each pixel or pixel cluster that was to be classified with a machine learning approach into exudates or non-exudates. The features were usually based on the color, brightness, size, shape, edge strength, texture, contextual information, etc. of the pixel clusters.

CNN has shown excellent performance in various applications including medical image processing. Pavle et al. [29] constructed a 10-layer of a neural network for exudates detection. In [5], Feng et al. applied a Fully Convolutional Neural Networks (FCN) to segment optic disc and exudates. Fujita et al. [7] used a single convolutional neural network to detect exudates, hemorrhages, and microaneurysms simultaneously. CNN based methods demonstrated great promise in the fundus image analysis. However, compared to the conventional state-of-the-art methods, the performance of CNN based approach, particularly with the best-reported sensitivity of 81.35% on the DiaReTDB1 dataset [5], still has much room for improvement. The weaker performance of the CNN based methods partially comes from the limited training data and the significantly imbalanced class data.

Imbalanced data poses enormous challenges to many classifiers, most of which are optimized to reduce global error rate without taking into consideration of data distribution. One straightforward method to alleviate the class imbalance problem is to upsample the minority class or to downsample

the majority class. Synthetic minority over-sampling technique (SMOTE) [30] is an excellent method from this perspective. In this paper, we applied the cGAN framework to augment the training dataset and to upsample the minority class specifically. Our results demonstrated that this was an effective way to improve network robustness and generalization properties.

3. Proposed methodology

In this section, we presented the modified U-net (MU-Net) structure for exudate detection and demonstrated applying conditional generative adversarial network(cGAN) to generate synthetic images as a new method of data augmentation and minority class upsampling.

3.1. Image pre-processing

The images from different datasets were acquired with different types of fundus cameras, which rendered different resolutions and signal-to-noise ratios (SNR). To reduce the image variation from different datasets, we applied a pre-processing step before feeding the images to the neural network. First, all the images were resized to 580×580 pixels. Second, only the green channel from the original RGB images were processed since the green channels showed the highest contrast of exudates from the background [31]. However, for the GAN training we used the original RGB images. Third, an adaptive contrast enhancement technique [32] was applied to improve the contrast of exudates on the retinal surface.

3.2. U-net and modified U-net (MU-net)

U-net was first proposed by Ronneberger [33] for biomedical image segmentation. It is a supervised method based on Convolutional Neural Networks (CNNs). It was designed to produce precise segmentation with a low number of training images, which was of particular importance for medical image applications since it was usually costly to label a large number of the medical images by medical experts.
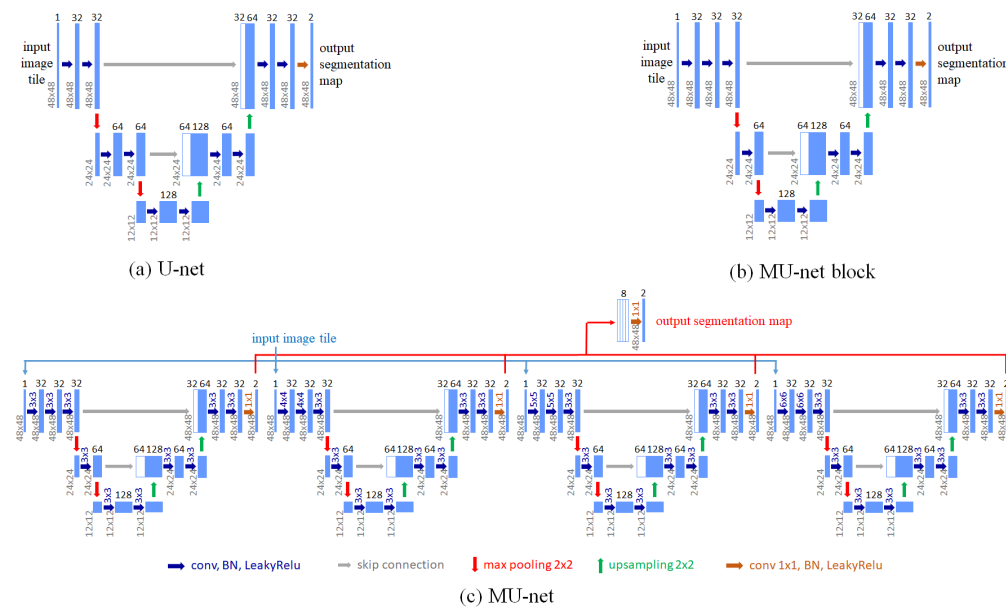

Fig. 2. (a) Structure of the U-net; (b) A building block of the MU-net(MU-net block); (c) MU-net consisting of 4 building blocks with different filter sizes.

We constructed an U-net model with eleven convolutional layers (Fig. 2(a)). It had a downsampling and an upsampling paths as shown in Fig. 2(a) and Table 1. In the downsampling path, two convolutional layers both with a 3×3 kernel were followed by normalization and dropout. A leaky rectified linear unit (LeakyReLU) instead of a rectified linear unit (ReLU) as in the original study [33], was implemented as the activation function. LeakyReLUs allow a small, non-zero gradient when a unit is not active [34]. A 2×2 max pooling with a stride of two at each dimension was applied. In the upsampling path, similar to the downsampling path, two 3×3 convolutional layers were followed by a LeakyReLU activation. An upsampling layer with a kernel size of 2×2 and stride of two in each dimension was applied and the output was fused with the corresponding layers in the downsampling path. In the last layer a 1×1 convolution reduced the number of output channels to 2. The input of the network was 48×48 image patches and the output image size was the same as the input.

Table 1. U-net block Configuration

Downsampling path	Output size	Upsampling path	Output size
Conv1 32 3×3 filters	48×48		
Conv2 32 3×3 filters	48×48	Conv11 2 1×1 filters	48×48
Max pooling 2×2	24×24	Conv10 32 3×3 filters	48×48
Conv3 64 3×3 filters	24×24	Conv9 32 3×3 filters	48×48
Conv4 64 3×3 filters	24×24	Upsampling 2×2	48×48
Max pooling 2×2	12×12	Conv8 64 3×3 filters	24×24
Conv5 128 3×3 filters	12×12	Conv7 64 3×3 filters	24×24
Conv6 128 3×3 filters	12×12	Upsampling 2×2	24×24

Table 2. MU-net block Configuration

Downsampling path	Output size	Upsampling path	Output size
Conv1 32 3×3 filters	48×48		
Conv2 32 3×3 filters	48×48	Conv11 2 1×1 filters	48×48
Conv3 64 3×3 filters	48×48	Conv10 32 3×3 filters	48×48
Max pooling 2×2	24×24	Conv9 32 3×3 filters	48×48
Conv4 64 3×3 filters	24×24	Upsampling 2×2	48×48
Max pooling 2×2	12×12	Conv8 64 3×3 filters	24×24
Conv5 128 3×3 filters	12×12	Conv7 64 3×3 filters	24×24
Conv6 128 3×3 filters	12×12	Upsampling 2×2	24×24

To improve the performance of the original U-net, we constructed a modified U-net (MU-net)(Fig. 2(c)), which was an ensemble of four MU-net blocks (Fig. 2(b)). The detailed configuration of the MU-net block was listed in (Table 2). The four MU-net blocks differed in the convolution kernel sizes (3×3, 4×4, 5×5 and 6×6)(Fig. 2(c)).

3.3. Conditional Generative Adversarial Network (cGAN) as a data augmentation and minority class upsampling method
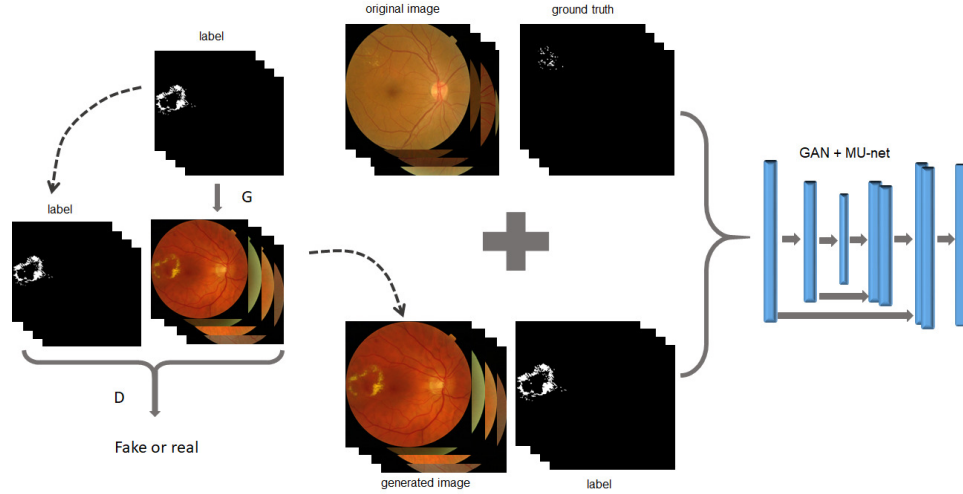

Fig. 3. GAN+MU-net. The cGAN network was first trained to generate label preserving synthetic images (with details of the training process described in the main context). The synthetic image pair together with the real image pair from the training dataset were used to train the GAN+MU-net.

GANs are generative adversarial networks proposed in [35–39]. In the unconditioned GAN, the generator G is trained to map from a random noise vector z to the output vector $y : G : z \rightarrow y$ so that the synthetic vector y is indistinguishable from the real vector as tested by an adversarially trained discriminator D. In conditional GANs (cGAN) [39], besides the random noise vector z, an observed image x is also supplied as the inputs. CGANs learn a mapping from the observed image x and the random noise vector z, to $y : G : \{x, z\} \rightarrow y$ by optimizing the following objective function:

$$G^* = arg \min_{G} \max_{D} \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L_1}(G) \tag{1}$$

where

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{(x,y) \sim p_{data}(x,y)}[log D(x, y)] \\ + \mathbb{E}_{x \sim P_{data}(x), z \sim p_z(z)}[log(1 - D(x, G(x, z)))] \tag{2}$$

$$\mathcal{L}_{L_1}(G) = \mathbb{E}_{(x,y) \sim p_{data}(x,y),(z) \sim p_z(z)}[\|y - G(x, z)\|_1] \tag{3}$$

We adapted the cGAN architecture from pix2pix in [39]. Two key features of the pix2pix framework were U-net [33] based generator and a Markovian discriminator (PatchGAN). The U-net based generator allowed low-level image information to shortcut across the network to achieve more realistic images.

During the training phase, the input of the pix2pix network was an image pair of the original RGB image with the corresponding binary ground truth image from the training set in e_ophtha_EX (Fig. 3). The training set for the pix2pix is the same as the training set of the U-net, MU-net block and MU-net. All the weights were initialized from a Gaussian distribution with mean 0 and standard deviation 0.02. To optimize the network, we alternated between one gradient descent step on the discriminator D, then one step on the generator G [39]. We applied the Adam solver [40], with learning rate of 0.0002, and the momentum parameters $\beta_1 = 0.5, \beta_2 = 0.999$.

The pix2pix network was trained for 350 epochs with batch size of 1. During the image generation phase, the input of the pix2pix network was a binary ground truth image from the training set, and the output was a synthetic image that was similar to but different from the input of the network. We applied the trained pix2pix network to all the 60 binary ground truth images in the e_ophtha_EX dataset to generate 60 synthetic images. Then we train the pix2pix for a second time independently without using the first trained weights. After that, with the second time trained pix2pix we generate another 60 synthetic images. Majority of the synthetic images were very similar to the real images in the sense of general looking and appearance of exudates, but some with obvious artifacts as shown in Fig. 4. This is consistent with previous studies that some GAN generated images are far from realistic images, which was noted before and thought as a general problem of GAN [39]. In total, 120 synthetic images were generated, among which 86 images with no artifacts were selected.
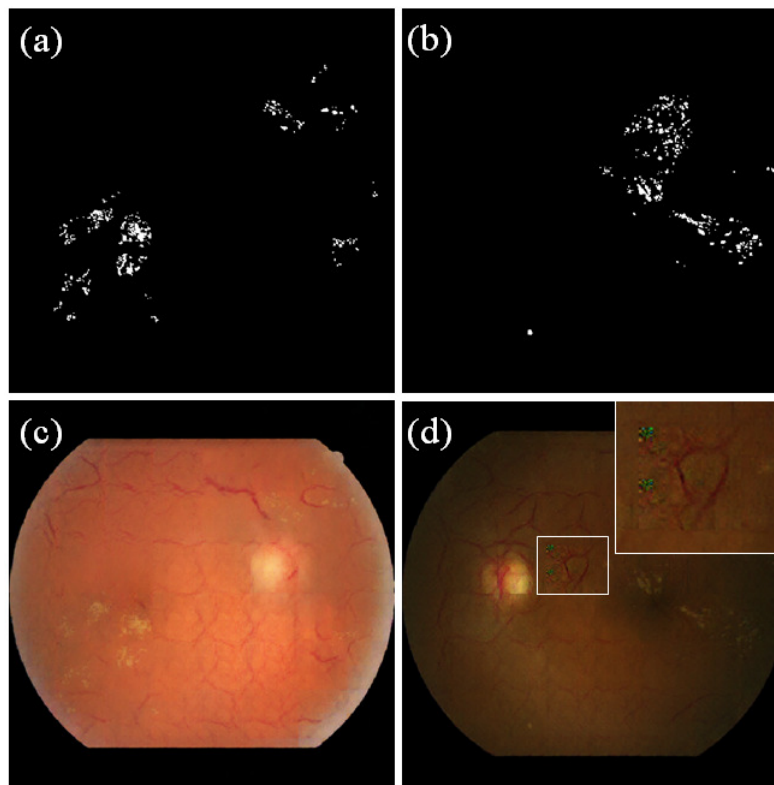

Fig. 4. Synthetic images generated using cGAN. (a) and (b) were the labeled binary images with the corresponding synthetic images (c) and (d) respectively. (c) was a good synthetic image that was similar to the real image in the sense of general looking and appearance of exudates, while (d) showed some artifacts. The inset in (d) was the enlarged version of the white rectangle area. The green spots were due to the falsely generated colors in the RGB channels.

3.4. Training and testing of the networks

It is worth to note that training of the networks was done only with the e_ophtha_EX dataset. All the other three datasets were to evaluate the network performance only.

In the e_ophtha_EX dataset, there are 82 labeled images of which 47 images consisting of exudates. During the training phase, 60 images (35 with and 25 without exudates) were used as

the training images, and the other 22 images (12 with and 10 without exudates) were used as the testing images.

To fully utilize the limited data, we applied a patch-based training process. We partitioned the image pair (the green channel of original RGB image and the corresponding ground truth image) into patches of size of 48×48. We used two different partitioning approaches: selective partitioning and uniform partitioning, and we named the generated patches as "selective patches" and "uniform patches". For the selective partitioning, we randomly selected a pixel from the set of the exudate pixels of an image. With the selected exudate pixel (x_i, y_i) as the reference point, we got a 48×48 block of the image $\{(x_{i-23}, y_{i-23}), (x_{i-22}, y_{i-22}), ..., (x_{i+24}, y_{i+24})\}$. This approach ensured that the selected patch contained exudate pixels. Patches selected this way may overlap. Similarly, to do the uniform partition, the images were uniformly spliced into patches of size of 48×48. Since there was only a small percentage of the image pixels were exudate pixels, the majority of the uniform patches did not consist of exudate pixels.

We applied a two-stage training process for the U-net, MU-net block, MU-net and the GAN+MU-net.

For the U-net, MU-net block and the MU-net, at the first stage, 67,200 patches were used to train the networks. 75% of these patches (50,400 patches) were selective patches(with exudates), and 25% of them were uniform patches. We experimented the ratio between the selective and the uniform patches and empirically found this one rendered the best results. Obviously, at this stage, the minority class (exudate pixel) was upsampled to alleviate the imbalanced data problem. At the second stage, 47,040 uniform patches were generated and used to fine tune the network. The purpose of the second stage training was to enhance the diversity of the samples and to reduce the false positive detections.

For the GAN+MU-net, besides the 60 training images from the e_ophtha_EX dataset, we also included 86 synthetic images without artifacts as described previously. During the training, 61,320 (75%) selective and 20,440 uniform patches were applied at the first training stage, and 114,464 uniform patches were applied at the second training stage.

For all the training at both stages, we applied stochastic gradient descent with a batch size of 64, a momentum of 0.3, and weight decay of 10^{-6}. We initialized the weights in each layer from a zero-mean Gaussian distribution with a standard deviation of 0.01. The learning rate was initialized to 0.01. Dropout was applied to alleviate overfitting, and the rate was set to 0.2. We trained 400 epochs at the first stage and fine-tuned the networks at the second stage for another 400 epochs.

All the trained networks were tested using the left 22 images in the e_ophtha_EX dataset to evaluate the performance. Besides, MU-net and GAN+MU-net were further tested on the other three datasets (DiaReTDB1, HEI-MED, and MESSIDOR) to assess the generalization property of the networks.

4.   Experimental results

4.1.   Datasets

We used four public datasets in the study: e_ophtha_EX [28], DiaReTDB1 [9], HEI-MED [26] and MESSIDOR [41]. However, training of the networks was done only with the e_ophtha_EX dataset, while the other three were used to test the performance of the networks.

The e_ophtha_EX dataset consists of 82 fundus images with four different image sizes of 1440×960 or 1504×1000 or 2544×1696 or 2048×1360 pixels. All images were acquired with a 45-degree field of view. Forty-seven of the images have exudates that were carefully marked by two ophthalmologists. The other 35 images do not contain exudates.

The DiaReTDB1 dataset contains 89 images with an image size of 1500×1152 pixels. All the images were captured using the same 50-degree field-of-view digital fundus camera with varying imaging settings. 26 images contain exudates in the dataset with labled groundtruth [42].

The HEI-MED dataset consists of 169 fundus images that are representative of a various degree of diabetic macular edema (DME) with a resolution of 2196×1958 pixels and with a 45 Field of View (FOV). All the images were captured with a Zeiss Visucam PRO fundus camera. 115 images are considered as normal retinal images and 54 retinal images are diagnosed with diabetic macular edema. 54 images contains exudates with labeled groundtruth while 115 images have no exudates [26].

The MESSIDOR dataset is the largest fundus image database, which contains 1200 TIFF images with three different image sizes, 1440×960, 2240×1488 and 2304×1536 pixels. All the images were acquired using a color video 3CDD camera on a TopCon TRC NW6 with a 45-degree field of view. 800 images were acquired with pupil dilation and 400 without dilation. There are 226 images consisting of exudates and 974 images without exudates [26]. However, labeled exudates were not provided. To perform the detailed evaluation of the proposed system for exudate detection, we manually annotated the exudate pixels of the 226 images with the help of two ophthalmologists. Labels at the image level, i.e., whether an image containing exudate or not, were labeled for all the 1200 images. Our manually annotated dataset is available upon request.

4.2. Evaluation metrics

We evaluated the performance of the proposed methods based on both lesion level [43] and image level measurements [44]. For the lesion level evaluation, we compared the predicted candidates with the truly labeled exudates provided in the datasets. Instead of directly counting the number of pixels that were correctly classified or misclassified, we applied the approach based on set operations [28, 43]. Briefly, if the two sets had enough overlap as controlled by an overlapping factor σ, which was set to 0.2 in our study, the same as in [28], the candidates were considered correctly classified. We computed the sensitivity, precision and the F1-score according to the equations defined in Table 3.

Table 3. Definitions of the evaluation metrics

Performance Measure	Mathematical Formula
Accuracy	(TP+TN)/(TP+TN+FN+FP)
Sensitivity	TP/(FN+TP)
Specificity	TN/(FP+TN)
Precision	TP/(FP+TP)
F1-Score	2*TP/(2*TP+FP+FN)

[1] TP stands for true positive; FP: false positve; TN: true negtive; FN: false negative.

For the image level evaluation, we applied the method proposed in [44]. One of the goals of the study is to develop an exudate detector, indicating whether or not an image is normal, i.e., free of exudates, or abnormal, i.e., containing one or more exudates. Thus we define a true negative image as an image without any exudates and the program does not detect any suspicious candidates. A true positive image is an image with exudates and the program detects candidate lesions, among which at least one candidate lesion is correctly detected as measured with the previously described lesion level criteria. A false positive image is defined as an image that has no exudates, but the program falsely detects candidate lesions. A false negative image is an image that consists of exudates but the program does not detect any candidate; or the program

detects candidates but none of the candidate lesions has enough overlap with the corresponding ground truth labels as evaluated using the lesion level criteria [28].

4.3. MU-net significantly improved exudate segmentation

Table 4. Evaluation of the performance of U-net, MU-net block, MU-net and GAN+MU-net at the lesion level. It was evaluated using the 22 test images (12 images with and 10 without exudates) in the e_ophtha_EX dataset.

	Accuracy	Specificity	Sensitivity	Precision	F1-score
U-net	99.94%	99.97%	83.49%	85.16%	84.32%
MU-net block	99.95%	99.97%	92.60%	89.67%	91.11%
MU-net	99.96%	99.98%	**94.12%**	91.25%	92.66%
GAN+MU-net	**99.97%**	**99.99%**	90.94%	**94.72%**	**92.79%**

Compared with the original U-net, our MU-net block clearly outperformed the original U-net by a large margin as tested both on the lesion level and image level (Table 4 and 5). F1-score of MU-net was 1.5% higher than MU-net block, demonstrating that the combination of four MU-net blocks could improve the performance of the network. The sensitivity of MU-net was more than 10% higher than that of the original U-net, and the precision and F1-score were more than 6% higher. On the image level, the accuracy of MU-net was more than 22% higher than the original U-net.
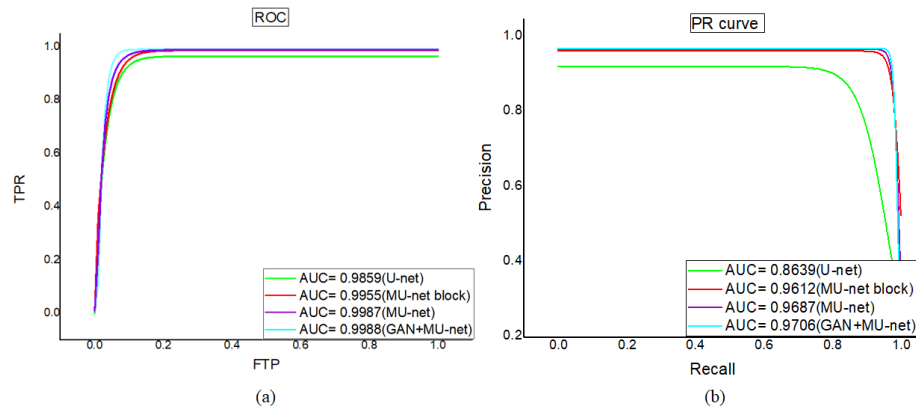

Fig. 5. (a) ROC curves of the U-net, MU-net block, MU-net and GAN+MU-net. (b) Precision-Recall curves of the four networks. GAN+MU-net had the highest AUCs as calculated either with the ROC curves or the Precision-Recall curves.

Figure 5 showed the ROC and the Precision-Recall(P-R) curves with the AUCs of the U-net, MU-net block, and MU-net. The AUCs of the MU-net block calculated both using the ROC and the P-R curves were higher than those of the U-net. The AUCs of the MU-net were also higher than MU-net block. The curves of the MU-net completely enclose of the U-net, demonstrating that the MU-net outperforms the U-net by a large margin.

Table 5. Evaluation of the performance of the U-net, MU-net block, MU-net and GAN+MU-net at the image level using the 22 test images in the e_ophtha_EX dataset.

	True	False	Accuracy
U-net	14	8	63.64%
MU-net block	17	5	77.27%
MU-net	19	3	86.36%
GAN+MU-net	21	1	**95.45%**

[1] True contains TP and TN and False contains FP and FN.

4.4. cGAN improved the performance of MU-net in exudate segmentation

To further resolve the imbalanced data problem, we applied the cGAN framework to generate synthetic images consisting of exudates. In Fig. 4, the first row showed the ground truth binary images from the training dataset and the second row showed the generated synthetic images using the cGAN method. Majority of the generated images consist of very realistic exudates (Fig. 4(c)), but some generated exudates showed artifacts (Fig. 4(d)). This is consistent with previous studies that some GAN generated images are far from realistic images, which was noted before and thought as a general problem of GAN [39]. The synthetic images together with the true images from the training dataset(Fig. 3) were used to train the MU-net, which we named it as GAN+MU-net.

Table 6. Comparison of the performances at the lesion level of the MU-net and GAN+MU-net using different testing datasets to test the generalization property of the networks.

Database	Methodology	Accuracy	Specificity	Sensitivity	Precision	F1-Score
e_ophtha_EX	MU-net	99.96%	99.98%	**94.12%**	91.25%	92.66%
	GAN+MU-net	**99.97%**	**99.99%**	90.94%	**94.72%**	**92.79%**
DiaRetDB1	MU-net	99.96%	99.98%	91.72%	**91.10%**	91.41%
	GAN+MU-net	**99.97%**	99.98%	**93.94%**	91.02%	**92.46%**
HEI-MED	MU-net	99.94%	99.97%	91.02%	90.39%	90.72%
	GAN+MU-net	**99.95%**	**99.98%**	**90.67%**	**91.88%**	**91.27%**
MESSIDOR	MU-net	99.95%	99.97%	91.46%	89.72%	90.58%
	GAN+MU-net	**99.96%**	**99.99%**	**95.33%**	**93.38%**	**94.34%**

Table 4 showed the comparison of MU-net and GAN+MU-net. Both the precision and F1-score were significantly increased in GAN+MU-net though the sensitivity of GAN+MU-net decreased. However, on the image level, GAN+MU-net performed significantly better than MU-net with more than 9.06% of increase of the accuracy (Table 5). The AUCs of the GAN+MU-net were also higher than those of the MU-net (Fig. 5).

Table 7. Comparison of the performances of MU-net and GAN+MU-net on different testing datasets evaluated at the image level.

Database	Network	True	False	Accuracy
e_ophtha_EX	MU-net	19	3	86.36%
	GAN+MU-net	21	1	**95.45%**
DiaRetDB1	MU-net	78	11	87.64%
	GAN+MU-net	82	7	**92.13%**
HEI-MED	MU-net	129	40	76.33%
	GAN+MU-net	150	19	**88.76%**
MESSIDOR	MU-net	1037	163	86.42%
	GAN+MU-net	1075	125	**89.58%**

True contains TP and TN and False contains FP and FN.

4.5. cGAN improved the generalization property of the MU-net

Generalization property is an important measure of the performance of a network. It reflects how well the trained network can be applied to similar but different datasets. We applied the trained MU-net and GAN+MU-net on the other three public datasets.

Table 6 and Table 7 showed the comparison results of MU-net and GAN+MU-net as evaluated using lesion level and image level measurements. In general, GAN+MU-net method outperformed MU-net in segmenting exudates. Specifically, GAN+MU-net achieved segmentation accuracy on image level of 92.13% on DiaRetDB1 dataset, 88.76% on HEI-MED dataset, and 89.58% on the MESSIDOR dataset, which were more than 4%, 12%, and 3% higher than that of the MU-net as shown in Table 7, respectively. On the lesion level, GAN+MU-net also showed improved performance, especially as measured by the more comprehensive F1-score (Table 6). It is conclusive that GAN+MU-net has better generalization property than MU-net.

5. Discussion

Detection of exudates is important in preventing DR caused blindness. In the study we sought to develop a deep convolutional neural network based method for automatic exudate detection. In order to alleviate the two challenges – limited expert labeled training data and imbalanced class data – we developed an ensemble MU-net and a cGAN based data augmentation and minority class upsampling method.

Our proposed MU-net significantly improved exudate segmentation. The MU-net was an ensemble of four MU-net blocks (Fig. 2(b) and Table 2). In the MU-net blocks, the max pooling layer was rearranged to the deeper layer of the network. Max pooling layer is useful for dimension reduction and helps make the learned features more invariant to small translations of the input [45], which is important for feature detection. However, positional information of a local maximum is lost after the max pooling operation. Such information is critical for accurate localization of the features. To preserve the position information, we applied the max pooling layer after the third convolution layer instead of the second one as in the original U-net.

Besides, MU-net was an ensemble network. We applied the bootstrap aggregating (Bagging) method [46] to improve the performance of the network in detecting exudates. Bagging is a powerful and reliable method for reducing generalization errors. Differences in model initialization or convolution kernel sizes, or selections of mini-batches are often enough to cause different members of the ensemble to make partially independent errors. As a result, neural

networks usually benefit from model averaging even if all of the models are trained on the same dataset. In our case, we constructed four MU-net blocks with the only difference in convolution kernel sizes (3×3, 4×4, 5×5 and 6×6). Convolutional layers with different kernel sizes have different receptive field and obtain different features. By averaging the four MU-net blocks, more detailed information was extracted and the network generalization error was reduced [47].

Over-fitting is a common problem for neural network models, especially for deep learning method that has a large number of parameters. In the medical image analysis, in addition to the constraint of the limited labeled data, extremely imbalanced class data renders another layer of difficulty. Imbalanced data has detrimental effects on the performance of classifiers usually by introducing significant biases and reducing the generalization property of the classifier.

Data augmentation is an effective way to increase the size of the training dataset and reduce the degree of the over-fitting problem. By applying transformations, such as elastic distortion or affine transformation to the original data, data augmentation artificially increases the size of the training set [48]. However, conventional data augmentation methods do not solve the class imbalance problem. In the paper, we applied a cGAN framework not only for generating synthetic data but also as a method to upsample the minority class. GAN+MU-net significantly improved the segmentation performance and the generalization property of the MU-net. Compared with conventional data augmentation method, cGAN was able to generate synthetic images consisting of exudates (the minority class), which significantly alleviated the imbalanced class problem and improved the generalization property of the network.

Table 8. Comparison of the performances of our method with previous published methods using the four public datasets

Database	Methodology	Accuracy	Specifity	Sensitivity	Precision
e_ophtha_EX	Moazam et al. [8]	89.25%	94.60%	81.20%	90.91%
	Zhang et al. [28]	––	––	74%	79%
	Proposed	**99.97**%	**99.99**%	**90.94**%	**94.72**%
DiaRetDB1	Moazam et al. [8]	87.72%	81.25%	92.42%	87.14%
	Rajan et al. [49]	85.39%	85%	86.2%	––
	Araujo et al. [50]	99.25%	99.44%	80.83%	15.57%
	Proposed	**99.97**%	**99.98**%	**93.94**%	**91.02**%
HEI-MED	Moazam et al. [8]	95.77%	96.41%	**94.63**%	**93.72**%
	Akram et al. [51]	92.5%	94.7%	96.1%	––
	Ali et al. [52]	––	––	––	82.6%
	Proposed	**99.95**%	**99.98**%	90.67%	91.88%
MESSIDOR	Agurto [53]	85.2%	90.2%	80.9%	––
	Moazam et al. [8]	98.36%	96.41%	94.63%	92.72%
	Proposed	**99.96**%	**99.99**%	**95.33**%	**93.38**%

We compared our results with recent published literatures on exudate detection (Table 8). In the study we tested our method using four public datasets. Some of the listed methods in Table 8 only applied on one of the datasets and not all the evaluation metrics were calculated. All the listed studies applied the same lesion level evaluation [43] method as in our study. From the table, it was clear that our proposed GAN+MU-net surpassed most of the methods except the one

proposed by Moazam et al. [8] on the HEI-MED dataset, although on all the other three datasets our method showed better performance than theirs. First, the color of the images in HEI-MED was much different from the e_ophtha_EX dataset, which may deteriorate the performance of the network that was trained on the e_ophtha_EX dataset. Second, in [8], the authors trained the network using the HEI-MED dataset and tested the performance on the same dataset, which usually gave better performance.

Limited expert labeled data and imbalanced data distribution are common problems encountered in medical image analysis based on DCNN. The proposed method with cGAN was able to oversample the minority class without the requirement of heuristic modeling of the class. Therefore, although we applied the method on fundus image processing, it is readily applied to other medical image analysis. The drawback of this approach is that it requires a training phase of the cGAN. However, once it is trained, the minority class with preserved labels can be easily and effectively generated.

As we have been writing the paper, we note that recent study by Douzas [54] came to the same conclusion, demonstrating that conditional generative adversarial networks could be an effective data augmentation method for imbalanced learning.

Disclosures

The authors declare that there are no conflicts of interest related to this article.

References

1. C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," PLoS Medicine **3**, e442 (2006).
2. R. R. Bourne, G. A. Stevens, R. A. White, J. L. Smith, S. R. Flaxman, H. Price, J. B. Jonas, J. Keeffe, J. Leasher, and K. Naidoo, "Causes of vision loss worldwide, 1990–2010: a systematic analysis," Lancet Glob. Heal. **1**, e339–e349 (2013).
3. A. D. Association, *Diabetes*, 7-12 (American Diabetes Association, 1966).
4. R. J. Tapp, J. E. Shaw, C. A. Harper, M. P. De Courten, B. Balkau, D. J. McCarty, H. R. Taylor, T. A. Welborn, and P. Z. Zimmet, "The prevalence of and factors associated with diabetic retinopathy in the australian population," Diabetes Care **26**, 1731–1737 (2003).
5. Z. Feng, J. Yang, L. Yao, Y. Qiao, Q. Yu, and X. Xu, "Deep retinal image segmentation: A fcn-based architecture with short and long skip connections for retinal image segmentation," in *International Conference on Neural Information Processing,* (Springer, 2017), pp. 713–722.
6. C. Pereira, L. Gonçalves, and M. Ferreira, "Exudate segmentation in fundus images using an ant colony optimization approach," Inf. Sci. **296**, 14–24 (2015).
7. J. H. Tan, H. Fujita, S. Sivaprasad, S. V. Bhandary, A. K. Rao, K. C. Chua, and U. R. Acharya, "Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network," Inf. Sci. **420**, 66–76 (2017).
8. M. M. Fraz, W. Jahangir, S. Zahid, M. M. Hamayun, and S. A. Barman, "Multiscale segmentation of exudates in retinal images using contextual cues and ensemble classification," Biomed. Signal Process. Control. **35**, 50–62 (2017).
9. K. Kamarainen, L. Sorri, A. R. V. Pietilä, and H. K. H. Uusitalo, "The DIARETDB1 diabetic retinopathy database and evaluation protocol," in *Proceedings of British Machine Vision Conference,* (2007).
10. R. Phillips, J. Forrester, and P. Sharp, "Automated detection and quantification of retinal exudates," Graefe's Arch. for Clin. Exp. Ophthalmol. **231**, 90–94 (1993).
11. H. Yazid, H. Arof, and H. M. Isa, "Automated identification of exudates and optic disc based on inverse surface thresholding," J. Med. Syst. **36**, 1997–2004 (2012).
12. I. N. Figueiredo, S. Kumar, C. M. Oliveira, J. D. Ramos, and B. Engquist, "Automated lesion detectors in retinal fundus images," Comput. Biol. Medicine **66**, 47–65 (2015).
13. K. Wisaeng, N. Hiransakolwong, and E. Pothiruk, "Automatic detection of exudates in retinal images based on threshold moving average models," Biophysics **60**, 288–297 (2015).
14. T. Walter, J.-C. Klein, P. Massin, and A. Erginay, "A contribution of image processing to the diagnosis of diabetic retinopathy-detection of exudates in color fundus images of the human retina," IEEE Transactions on Med. Imaging **21**, 1236–1243 (2002).
15. A. Sopharak, B. Uyyanonvara, S. Barman, and T. H. Williamson, "Automatic detection of diabetic retinopathy exudates from non-dilated retinal images using mathematical morphology methods," Comput. Med. Imaging Graph. **32**, 720–727 (2008).

16. D. Welfer, J. Scharcanski, and D. R. Marinho, "A coarse-to-fine strategy for automatically detecting exudates in color eye fundus images," computerized Med. Imaging Graph. **34**, 228–235 (2010).

17. B. Harangi and A. Hajdu, "Automatic exudate detection by fusing multiple active contours and regionwise classification," Comput. Biol. Medicine **54**, 156–171 (2014).

18. E. Imani and H.-R. Pourreza, "A novel method for retinal exudate segmentation using signal separation algorithm," Comput. Methods Programs Biomed. **133**, 195–205 (2016).

19. B. M. Ege, O. K. Hejlesen, O. V. Larsen, K. Mÿller, B. Jennings, D. Kerr, and D. A. Cavan, "Screening for diabetic retinopathy using computer based image analysis and statistical classification," Comput. Methods Programs Biomed. **62**, 165–175 (2000).

20. C. Sinthanayothin, J. F. Boyce, T. H. Williamson, H. L. Cook, E. Mensah, S. Lal, and D. Usher, "Automated detection of diabetic retinopathy on digital fundus images," Diabet. Medicine **19**, 105–112 (2002).

21. H. Li and O. Chutatape, "Automated feature extraction in color retinal images by a model based approach," IEEE Transactions on Biomed. Eng. **51**, 246–254 (2004).

22. D. Usher, M. Dumskyj, M. Himaga, T. H. Williamson, S. Nussey, and J. Boyce, "Automated detection of diabetic retinopathy in digital retinal images: a tool for diabetic retinopathy screening," Diabet. Medicine **21**, 84–90 (2004).

23. C. I. Sánchez, R. Hornero, M. I. López, M. Aboy, J. Poza, and D. Abásolo, "A novel automatic image processing algorithm for detection of hard exudates based on retinal image analysis," Med. Eng. Phys. **30**, 350–357 (2008).

24. M. Niemeijer, B. van Ginneken, S. R. Russell, M. S. Suttorp-Schulten, and M. D. Abramoff, "Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis," Investig. Ophthalmol. Vis. Sci. **48**, 2260–2267 (2007).

25. A. D. Fleming, S. Philip, K. A. Goatman, G. J. Williams, J. A. Olson, and P. F. Sharp, "Automated detection of exudates for diabetic retinopathy screening," Phys. Medicine Biol. **52**, 7385 (2007).

26. L. Giancardo, F. Meriaudeau, T. P. Karnowski, Y. Li, S. Garg, K. W. Tobin Jr, and E. Chaum, "Exudate-based diabetic macular edema detection in fundus images using publicly available datasets," Med. Image Analysis **16**, 216–226 (2012).

27. B. Harangi, B. Antal, and A. Hajdu, "Automatic exudate detection with improved naïve-bayes classifier," in *Computer-Based Medical Systems (CBMS), 2012 25th International Symposium on,* (IEEE, 2012), pp. 1–4.

28. X. Zhang, G. Thibault, E. Decencière, B. Marcotegui, B. Laÿ, R. Danno, G. Cazuguel, G. Quellec, M. Lamard, P. Massin *et al.*, "Exudate detection in color retinal images for mass screening of diabetic retinopathy," Med. Image Analysis **18**, 1026–1043 (2014).

29. P. Prentašić and S. Lončarić, "Detection of exudates in fundus photographs using deep neural networks and anatomical landmark detection fusion," Comput. Methods Programs Biomed. **137**, 281–292 (2016).

30. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," J. Artif. Intell. Res. **16**, 321–357 (2002).

31. A. Gupta, A. Issac, N. Sengar, and M. K. Dutta, "An efficient automated method for exudates segmentation using image normalization and histogram analysis," in *Contemporary Computing (IC3), 2016 Ninth International Conference on,* (IEEE, 2016), pp. 1–5.

32. C. Sinthanayothin, J. F. Boyce, H. L. Cook, and T. H. Williamson, "Automated localisation of the optic disc, fovea, and retinal blood vessels from digital colour fundus images," Br. J. Ophthalmol. **83**, 902–910 (1999).

33. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention,* (Springer, 2015), pp. 234–241.

34. A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML,* vol. 30 (2013), p. 3.

35. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems,* (2014), pp. 2672–2680.

36. E. L. Denton, S. Chintala, and R. Fergus, "Deep generative image models using aï£ij laplacian pyramid of adversarial networks," in *Advances in Neural Information Processing Systems,* (2015), pp. 1486–1494.

37. A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," ArXiv Prepr. ArXiv:1511.06434 (2015).

38. T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems,* (2016), pp. 2234–2242.

39. P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," ArXiv Prepr. (2017).

40. D. Kinga and J. B. Adam, "A method for stochastic optimization," in *International Conference on Learning Representations (ICLR),* vol. 5 (2015).

41. E. Decencière, X. Zhang, G. Cazuguel, B. Laÿ, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay *et al.*, "Feedback on a publicly distributed image database: the messidor database," Image Analysis Stereol. **33**, 231–234 (2014).

42. R. Kälviäinen and H. Uusitalo, "Diaretdb1 diabetic retinopathy database and evaluation protocol," in *Medical Image Understanding and Analysis,* vol. 2007 (Citeseer, 2007), p. 61.

43. C. Wolf and J.-M. Jolion, "Object count/area graphs for the evaluation of object detection and segmentation algorithms," Int. J. Document Analysis Recognit. (IJDAR) **8**, 280–296 (2006).

44. L. Tang, M. Niemeijer, J. M. Reinhardt, M. K. Garvin, and M. D. AbrÃămoff, "Splat feature classification with application to retinal hemorrhage detection in fundus images," IEEE Transactions on Med. Imaging **32**, 364–375 (2013).
45. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436 (2015).
46. L. Breiman, "Bagging predictors," Mach. Learn. **24**, 123–140 (1996).
47. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436 (2015).
48. W. Zhu, X. Xiang, T. D. Tran, and X. Xie, "Adversarial deep structural networks for mammographic mass segmentation," ArXiv Prepr. ArXiv:1612.05970 (2016).
49. S. Rajan, T. Das, and R. Krishnakumar, "An analytical method for the detection of exudates in retinal images using invertible orientation scores," in *Proceedings of the World Congress on Engineering,* vol. 1 (2016).
50. F. Araujo, R. Veras, A. Macedo, and F. Medeiros, "Automatic detection of exudates in retinal images using neural network," Dept Comput. Fed. Univ. Braz. (2013).
51. M. U. Akram, A. Tariq, S. A. Khan, and M. Y. Javed, "Automated detection of exudates and macula for grading of diabetic macular edema," Comput. Methods Programs Biomed. **114**, 141–152 (2014).
52. S. Ali, D. Sidibé, K. M. Adal, L. Giancardo, E. Chaum, T. P. Karnowski, and F. Mériaudeau, "Statistical atlas based exudate segmentation," Comput. Med. Imaging Graph. **37**, 358–368 (2013).
53. C. Agurto, V. Murray, H. Yu, J. Wigdahl, M. Pattichis, S. Nemeth, E. S. Barriga, and P. Soliz, "A multiscale optimization approach to detect exudates in the macula," IEEE J. Of Biomed. Heal. Informatics **18**, 1328–1336 (2014).
54. G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks," Expert. Syst. with Appl. **91**, 464–471 (2018).