Pathological Evidence Exploration in Deep Retinal Image Diagnosis

Yuhao Niu,[1,2,*] Lin Gu,[4,*] Feng Lu,[1,2,3,†] Feifan Lv,[1,3] Zongji Wang,[1] Imari Sato,[4]
Zijian Zhang,[5] Yangyan Xiao,[6] Xunzhang Dai,[5] Tingting Cheng[5]
[1]State Key Laboratory of VR Technology and Systems, School of CSE, Beihang University, Beijing, China
[2]Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, Beijing, China
[3]Peng Cheng Laboratory, Shenzhen, China
[4]National Institute of Informatics, Japan
[5]Xiangya Hospital Central South University, China
[6]The Second Xiangya Hospital of Central South University, China

Abstract

Though deep learning has shown successful performance in classifying the label and severity stage of certain disease, most of them give few evidence on how to make prediction. Here, we propose to exploit the interpretability of deep learning application in medical diagnosis. Inspired by Koch's Postulates, a well-known strategy in medical research to identify the property of pathogen, we define a pathological descriptor that can be extracted from the activated neurons of a diabetic retinopathy detector. To visualize the symptom and feature encoded in this descriptor, we propose a GAN based method to synthesize pathological retinal image given the descriptor and a binary vessel segmentation. Besides, with this descriptor, we can arbitrarily manipulate the position and quantity of lesions. As verified by a panel of 5 licensed ophthalmologists, our synthesized images carry the symptoms that are directly related to diabetic retinopathy diagnosis. The panel survey also shows that our generated images is both qualitatively and quantitatively superior to existing methods.

Introduction

Deep learning has become a popular methodology in analyzing medical images such as diabetic retinopathy detection (Gulshan et al. 2016), classifying skin cancer (Esteva et al. 2017). Though these algorithms have proven quite accurate in classifying specific disease label and severity stage, most of them lack the ability to explain its decision, a common problem that haunts deep learning community. Lacking interpretability is especially imperative for medical image application, as physicians or doctors relies on medical evidence to determine whether to trust the machine prediction or not.

In this paper, we propose a novel technique inspired by Koch's Postulates to give some insights into how convolutional neural network (CNN) based pathology detector makes decision. In particular, we take the diabetic retinopathy detection network (Antony 2016) for example. Noted that not limited for (Antony 2016), this strategy could also be extended to interpret more general deep learning model.

We at first apply (Antony 2016) on the reference image (Fig 1.(a)) and extract the pathological descriptor (Fig 1.(b)) that encodes the neuron activation directly related to prediction. Picking thousand out of millions of parameters in neuron network is like separating the potential pathogen. Koch's Postulates claims that the property of pathogen, though invisible for naked eye, could be determined by observing the arose symptom after injecting it into subject. Similarly, we *inject* the pathologic descriptor into the binary vessel segmentation (Fig 1.(c)) to synthesize the retinal image. We achieve this with a GAN based network as illustrated in Fig 5. Given pathologic descriptor and binary vessel segmentation, our generated image (Fig 1.(d)) exhibits the expected symptom such as microaneurysms and hard exudates that appear in the target image. Since our descriptor is lesion-based and spatial independent, we could arbitrarily manipulate the position and number of lesions. Evaluated With a panel of 5 licensed ophthalmologists, our generated retinal images are qualitatively and quantitatively superior to existing methods.

Specifically, we encode a series of pathological descriptor as illustrated in Fig 2. According to our analysis, the diabetic retinopathy detection network (Antony 2016) predicts the diabetic level through a few dimensions (6 among 1024, lighted with colors) of bottleneck feature in Fig 2. We then identify the neurons that directly contribute to these 6 dimension bottleneck features in the activation net and record their position and activation value as the pathologic descriptor. Since neuron activation is spatially correlated with individual lesion, our descriptor is defined as lesion-based that allows us to manipulate its position and quantity.

Our main contributions are mainly three-folds:

1. We define a pathological descriptor that encodes the key parameters of CNN which is directly related to disease prediction. This descriptor is associated with individual

*These two authors contributed equally to the paper.
†Corresponding Author: Feng Lu (lufeng@buaa.edu.cn)

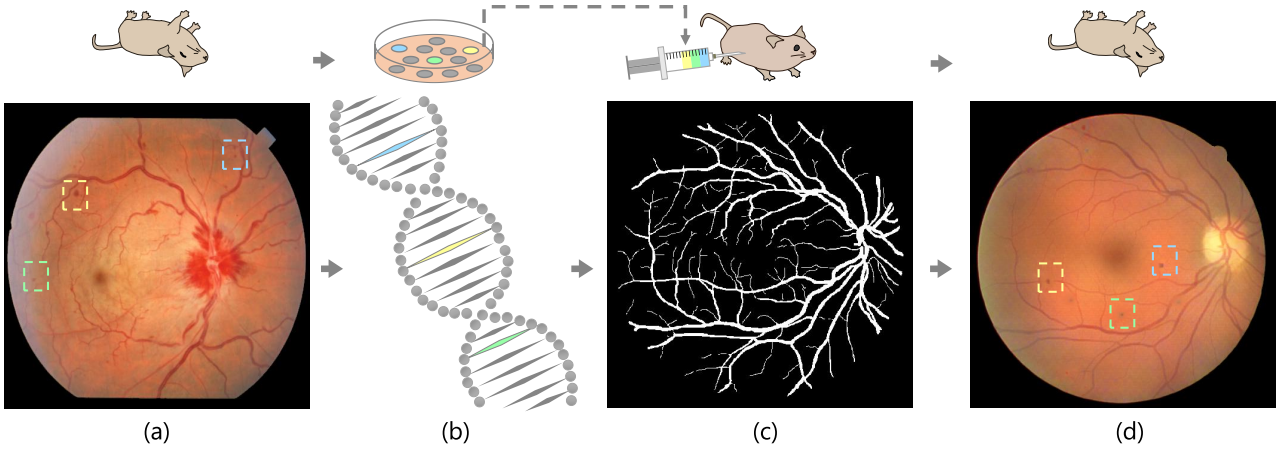

Figure 1: Koch's Postulates are criteria in Evidence Based Medicine (EBM) to determine the pathogen for a certain disease. They state that the pathogen must be found in diseased subjects but not in healthy ones; the pathogen must be isolated and grown in pure culture; the cultured pathogen should cause disease after injected into healthy subject; the pathogen isolated again is the same as the injected one. The methodology of this paper is an analogy to Koch's Postulates. (a) Reference retinal image with disease. (b) Extract pathological descriptor from an image like separating pathogen. (c) Apply the descriptor on a binary vessel segmentation like injecting purified pathogen into the subject. (d) The synthesized image or subject with same symptom.

lesion.

2. Inspired by Koch's Postulates, we propose a novel interpretability strategy to visualize the pathological descriptor by synthesizing fully controllable pathological images. The synthesized images are verified by a group of licensed ophthalmologists.
3. With our pathological descriptors, we could generate medical plausible pathology retinal image where the position and quantity of lesion could arbitrarily manipulated.

Related Works

Diabetic Retinopathy Detection

With the fast development of deep learning, this technique has achieved success on several medical image analysis applications, such as computer-aided diagnosis of skin cancer, lung node, breast cancer, *etc.* In the case of diabetic retinopathy (DR), automatic detection is particularly needed to reduce the workload of ophthalmologists, and slow down the progress of DR by performing early diagnose on diabetic patients (Gulshan et al. 2016).

In 2015, a Kaggle competition (Kaggle 2016) was organized to automatically classify retinal images into five stages according to *International Clinical Diabetic Retinopathy Disease Severity Scale* (American Academy of Ophthalmology 2002). Not surprisingly, all of the top-ranking methods were based on deep learning. Then, another deep learning method (Gulshan et al. 2016), trained on 128175 images, achieved a high sensitivity and specificity for detecting diabetic retinopathy. Noting that image-level grading lacks intuitive explanation, the recent methods (Yang et al. 2017; Wang et al. 2017) shifted the focus to locate the lesion position. However, these methods often relied on a large training set of lesion annotations from professional experts.

In this paper, we propose a novel strategy to encode the descriptor from the DR detector's activated neurons directly related to the pathology. For the sake of generality, we select the o_O(Antony 2016), a CNN based method within the top-3 entries on Kaggle's challenge. Even now, the performance of o_O is still equivalent to the latest method (Wang et al. 2017). This method is trained and tested on the image level DR severity stage.

Generative Adversarial Networks

Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) were first proposed in 2014, adopting the idea of zero-sum game. Subsequently, CGANs (Mirza and Osindero 2014) attempted to use additional information to make the GAN controllable. DCGAN (Radford, Metz, and Chintala 2015) combined CNN with traditional GAN to achieve a shocking effect. Pix2pix (Isola et al. 2017) used the U-Net (Ronneberger, Fischer, and Brox 2015) combined with adversarial training and achieved amazing results. CycleGAN (Zhu et al. 2017) used two sets of GANs and added cycle loss to achieve style transfer on unpaired data.

Style Transfer

Recently, neural style transfer using deep CNNs becomes popular. A typical method was proposed in (Gatys, Ecker, and Bethge 2016), in which they directly optimized the input pixels to restrict both content and style features extracted by CNNs. Later in (Luan et al. 2017) and (Gatys et al. 2017), semantic masks were introduced to improve image style transfer. To speed up the transfer procedure, (Johnson, Alahi, and Fei-Fei 2016) added a network to synthesize image. Once trained by a given style reference image, it can finish style

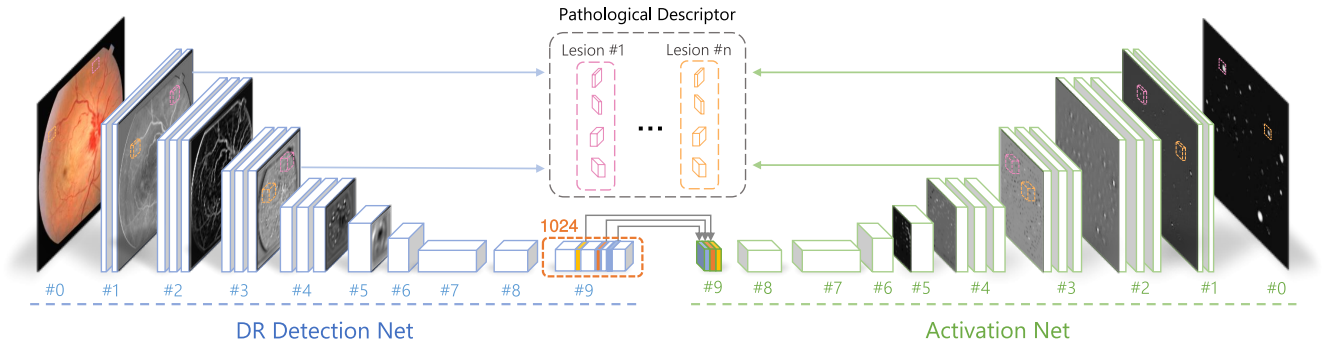

Figure 2: The process for extracting pathological descriptors. First, a pathological reference image is fed into the DR detection net. Next, the extracted features are mapped to the input pixel space through the activation net to get activation projections, which indicate the locations and appearance of most lesions. Finally, the features and related activation projections are cropped into small patches around the found lesions, which are recognized as pathological descriptors.

transfer by one feed-forward propagation.

Synthesizing Biomedical Images

The traditional biomedical imaging synthesizing uses the medical and biological prior knowledge accumulated by humans, combined with complex simulation methods to produce realistic results. Probably most well-known efforts are the work (Fiorini et al. 2014), the work (Bonaldi et al. 2016), GENESIS (Bower, Cornelis, and Beeman 2015), NEURON (Carnevale and Hines 2006), L-Neuron (Ascoli and Krichmar 2000) *etc.* With the development of deep learning, some methods like (Zhao et al. 2018) began to synthesize realistic retinal and neuronal images in a data-driven way. Tub-sGAN, a variant of (Zhao et al. 2018), synthesized image given a binary tubular annotation and a reference image. Although their generated images could show pleasant visual appearance, the diabetic retinopathy symptoms and retina physiological details are either lost or incorrect as verified by the ophthalmologists. In this paper, we propose a pathologically controllable method that can generate realistic retinal image with medical plausible symptoms.

Pathological Descriptor

In this section, we would describe how to extract lesion based pathological descriptor from the activated neurons of Diabetic Retinopathy (DR) detector (Antony 2016).

Diabetic Retinopathy Detection

Here, we briefly introduces DR detector (Antony 2016) used in this paper. It takes retinal fundus image of shape $448 \times 448 \times 3$ as input and outputs the 5 grades (0-4) diabetic retinopathy severity. As shown in the left part of Fig 2, the DR detection network is stacked with several blocks, each of which consists of 2-3 convolutional layers and a pooling layer. As the number of layers increases, the network merges into a $1 \times 1 \times 1024$ bottleneck feature. To add nonlinearity to the net and to avoid neuronal death, a leaky ReLU (Maas, Hannun, and Ng 2013) with negative slope 0.01 is applied following each convolutional and dense layer.

The bottleneck feature is fed into a dense layer (not shown in the figure) to predict the severity labels provided in the *DR Detection Challenge* (Kaggle 2016). The network is trained with Nesterov momentum over 250 epochs. Data augmentation methods, such as dynamic data re-sampling, random stretching, rotation, flipping, and color augmentation, are all applied. We refered to (Antony 2016) for details.

Key Bottleneck Features

We further identify a few *key features* (colored one in the middle of Fig 2) from 1024 dimension bottleneck features. After the training stage of network, we are able to generate the bottleneck features for individual sample in the training set. Then we train a random forest (Dollár and Zitnick 2015) classifier to predict the severity label on these bottleneck features. Following (Gu et al. 2017), we could identify the important features by counting the frequency of each feature that contributes to prediction. In the current setting, we find that, with random forest, only 6 of 1024 bottleneck features could deliver equivalent performance as o_O (Antony 2016).

Activation Network

Among millions of neurons in the network, only thousands of neurons actually contribute to the bottleneck feature's activation and the final prediction. To explore the activity of these neurons, we introduce an activation network (Zeiler and Fergus 2014). We perform a back-propagation-liked procedure from the 6 dimension key bottleneck features to get *activation projections* for detector feature layers.

As shown in the right part of Fig 2, our activation net is a reverse version of the DR detector. For individual layer in detector, there is a corresponding reverse layer with the same configuration of strides and kernel size. For a convolutional layer, the corresponding layer performs transposed convolution, which shares same weights, except that the kernel is flipped vertically and horizontally. For each max pooling layer, there is an unpooling layer that conducts a partially inverse operation, where the max elements are located through a skip connection and non-maximum elements are

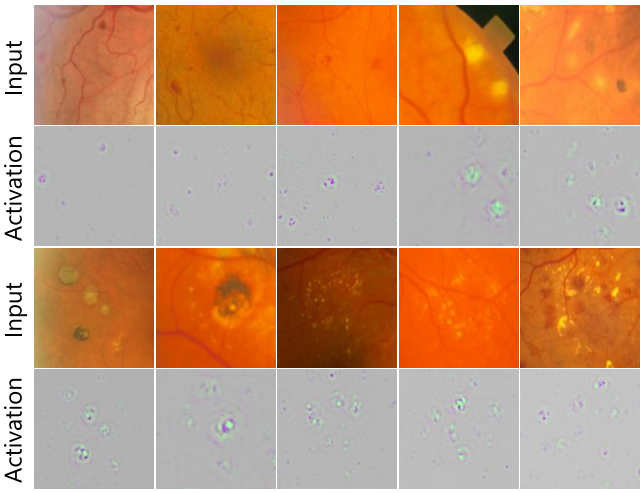

Figure 3: We input fundus with different lesions into the pipeline in Fig 2 and extract their related activation projections in layer #0. Some results are cropped and shown here.

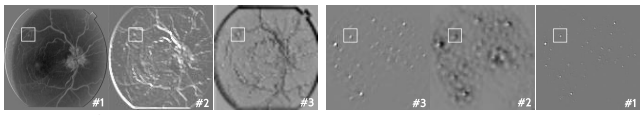

Figure 4: Feature maps and related activation projections. In each layer, only one channel is shown.

filled with zeros. For a ReLU function, there is also a ReLU in the activation net, which drops out the negative projection. We treat the fully connection as 1×1 convolution. In the implementation we use auto-differentiation provided in Tensorflow (Abadi et al. 2016) to reverse each layer.

Here, we demonstrate some neuron activation examples in Fig 3&4. It shows that, though our DR detector is trained with the image-level labels, its neuronal activity is sensitive to and therefore can locate a variety of DR lesions such as microaneurysms, soft exudates and hard exudates.

Retinal Pathological Descriptor

Now, we define a *pathological descriptor* to encode lesion feature and activation, which should serve as the evidence when doctors make diagnose. As the neuron activation is spatially correlated with the retinal lesions, we could associate the descriptor with individual lesion. Fig 4 shows how the lesion feature changes across different layers. A descriptor contains patches cropped from these maps.

When a retinal fundus x_s is fed into the pipeline in Fig 2, the features and activation projections in layer l are denoted as F_l and A_l, respectively. In order to indicate position and boundary of an individual lesion, the last layer activation A_0 is thresholded into a binary mask M_0. To describe a lesion area, we define a rectangle region r that covers a connected conponent (a lesion) in M_0. Thus different lesions could be denoted as different r. We then use the multi-layer information to construct our retinal pathological descriptor for each lesion. In particular, we first down-sample M_0 into each layer l to generate the binary lesion mask M_l. With r, we cropped feature patch F_{lr} from F_l, activation patch A_{lr} from A_l and mask patch M_{lr} from M_l. The pathological descriptor for lesion r consists of the information from multiple layers, written as $d = \{d_l | l \in \Lambda\}$, where $d_l = \langle M_{lr}, A_{lr}, F_{lr} \rangle$.

Visualizing Pathological Descriptor

According to Koch's Postulates, though pathogen is invisible (at least for naked eye), its property could be observed on the subject after injecting the purified pathogen. Similarly, we evaluate and visualize the interpretative medical meaning of this descriptor by using a GAN based method to generate fully controllable DR fundus images. Our goal is to synthesize the diabetic retinopathy fundus images(Fig 1.(d)) that carry the lesions that appear on a pathological reference one (Fig 1.(a)). Since our descriptor is lesion based, we could even arbitrarily manipulate the number and position of symptom. As shown in Fig 5, with given descriptors $\mathcal{D}$, we design a novel conditional GAN to achieve this.

Our whole network structure consists of four sub-nets: the generator net, the discriminator net, the retina detail net, and the DR detection net. Given the vessel segmentation image and a noise code as input, the generator tries to synthesize a tubular structured phantom. The discriminator net tries to distinguish the synthesized images from the real ones. To further enhance the physiological details during generation, we use the retina detail net to constrain the detail reconstruction. The DR detection net is the *key part* of our proposed architecture, which constrains the synthesized images with the user-specified pathological descriptors in feature level. After training, one can easily obtain synthesized fundus from vessel segmentations using the generator net.

Generator and Discriminator

We use a U-Net (Ronneberger, Fischer, and Brox 2015) network structure for our generator network.Taking a segmentation image $y \in \{0,1\}^{W \times H}$ with a noise code $z \in \mathbb{R}^Z$ as input, the network outputs a synthesized diabetic retinopathy fundus RGB image $\hat{x} \in \mathbb{R}^{W \times H \times 3}$. The entire image synthesis process can be expressed as $G_\theta : (y, z) \mapsto \hat{x}$. Similarly, we can also define discriminant function $D_\gamma : (X, y) \mapsto p \in [0, 1]$. When X is the real image x, p should tend to 1 and when X is the composite image $\hat{x}$, p should tend to 0. We follow the GAN's strategy and solve the following optimization problem that characterizes the interplay between G and D:

$$\max_\theta \min_\gamma L(G_\theta, D_\gamma) = \mathbb{E}_i[L_{\text{adv}}(i, \theta, \gamma) + L_{\text{retina}}(i, \theta) + L_{\text{patho}}(i, \theta)], \quad (1)$$

where $L_{\text{adv}} = \log D_\gamma(x_i, y_i) + \log(1 - D_\gamma(G_\theta(y_i, z_i), y_i))$, is the adversarial loss, with L_{retina} and L_{patho} being retina detail loss and pathological loss. To be more specific, learning the discriminator D amounts to maximizing $-L_D = L_{\text{adv}}$ and the generator G is learned by minimizing a loss $L_G = \tilde{L}_{\text{adv}} + L_{\text{retina}} + L_{\text{patho}}$ with a simpler adversarial

$$\tilde{L}_{\text{adv}} = -\log D_\gamma(G_\theta(y_i, z_i), y_i). \quad (2)$$

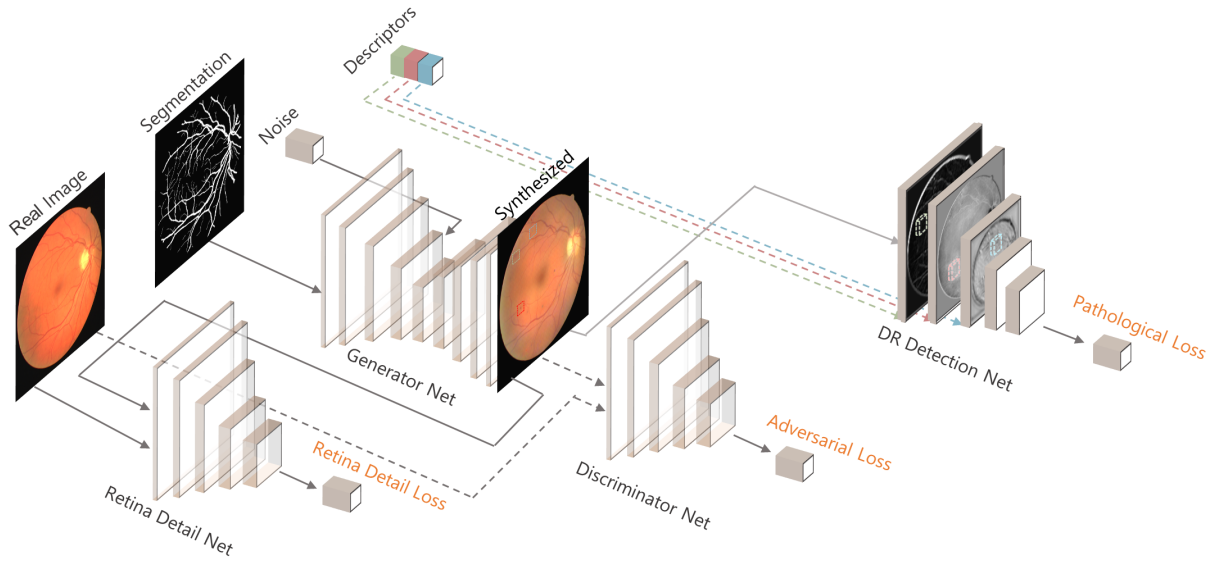

Figure 5: The architecture and data flow of our symptom transfer GAN, which contains four nets in training phase. After training, the generator itself is able to synthesize retinal fundus with lesions on specific locations.

Retina Detail Loss

Though a L1 loss (or MAE) between synthetic image could deliver a satisfactory result for style transfer application on common images, it fails to preserve the physiological details in the fundus. We will elaborate this in the experiment section. Therefore, we define *retina detail loss* as:

$$L_{\text{retina}} = w_{\text{dd}} L_{\text{dd}} + w_{\text{tv}} L_{\text{tv}}, \tag{3}$$

where diverge of details L_{dd} is meant to preserve physiological details in the fundus, while total variance loss L_{tv} is the global smoothing term.

We choose to measure diverge of details in VGG-19 (Simonyan and Zisserman 2014) feature space. For specific layer λ and VGG feature extraction function F_V^{λ},

$$L_{\text{dd}} = \|F_V^{\lambda}(x_i) - F_V^{\lambda}(\hat{x}_i)\|. \tag{4}$$

In addition, to ensure the overall smoothness, we also regulate image gradients to encourage spatial smoothness L_{tv}:

$$\sum_{w,h} \|\hat{x}_i^{(w,h+1)} - \hat{x}_i^{(w,h)}\| + \|\hat{x}_i^{(w+1,h)} - \hat{x}_i^{(w,h)}\|. \tag{5}$$

Pathological Loss

To constrain the synthesized image to carry pathological features, we enforce it to have the similar detector neuron activation on lesion regions to the reference image. In this way, the synthesized image should be equivalent to the reference from DR detector point of view. We regulate the neuron activation to be close to the ones recorded in pathological descriptors $\mathcal{D}$ with a *pathological loss* :

$$L_{\text{patho}} = w_{\text{dp}} L_{\text{dp}} + w_{\text{mv}} L_{\text{mv}}, \tag{6}$$

where the pathological diverge L_{dp} is the differences on features summed over lesion regions and layers, while masked variance loss L_{mv} represents the local smoothing term.

As shown in Fig 5, we input the synthesized image $\hat{x}_i$ into the pre-trained DR detector. To ensure the extracted feature pathes $\mathcal{F}_{l\rho}$ in fixed pre-specified regions $\rho(d)$ are close to those F_{lr} in given descriptors $d \in \mathcal{D}$ across all layers $l \in \Lambda$, we define L_{dp} as ($\otimes$ means elementwise multiply)

$$L_{\text{dp}} = \frac{1}{|\mathcal{D}|} \sum_{d\in\mathcal{D}} \frac{1}{|\Lambda|} \sum_{l\in\Lambda} \frac{w_{\text{gram}}}{W_\rho H_\rho} \cdot ||\mathbb{G}(\mathcal{M}_{ld} \otimes \mathcal{F}_{l\rho}(\hat{x}_i)) - \mathbb{G}(\mathcal{M}_{ld} \otimes F_{lr}(d))||, \tag{7}$$

where $\mathcal{M}_{ld} = M_{lr} \otimes \text{normalize}(|A_{lr}|)$ is computed with M_{lr}, A_{lr}, the elements of d_l, and used as a feature mask. The binary mask M_{lr} restricts loss in pixel-level, and A_{lr} stresses the lesion region as a soft mask. In this definition, we measure the diverge by the symmetric Gram matrix $\mathbb{G}_{K\times K}$, which represents the covariance of different channels in feature maps $F \in \mathbb{R}^{W\times H\times K}$:

$$\mathbb{G}(F)_{i,j} = \sum_{w,h} F_{whi} F_{whj}. \tag{8}$$

At the same time, our network would integrate the synthetic lesion features into the current background. First of all, a lesion redistribution mask M_{rd}, which covers lesion regions in the synthesized image, is computed based on all ρ and M_{0r}. Then the mask is dilated to M_{grd} by a gauss kernel to softly expand the boundary. With masked synthetic image $\tilde{x}_i = M_{\text{grd}} \hat{x}_i$, we define masked variance loss L_{mv} as

$$\sum_{w,h} \|\tilde{x}_i^{(w,h+1)} - \tilde{x}_i^{(w,h)}\| + \|\tilde{x}_i^{(w+1,h)} - \tilde{x}_i^{(w,h)}\|. \tag{9}$$

Implementation Details

The chosen norm in above equations is L1. Weights for different losses are $w_{\text{dd}} = 1, w_{\text{tv}} = 100, w_{\text{dp}} = 10, w_{\text{mv}} = 5w_{\text{tv}}, w_{\text{gram}} = 10^6$. Based on experience, we set λ to be the second convolutional layer in the fourth block of VGG.

The batch size is set to 1. Before each training step, the input image values are scaled to $[-1, 1]$, and a random rotation is performed on the input. The training is done using the ADAM optimizer (Kingma and Ba 2014) and the learning rate is set to 0.0002 for the generator and 0.0001 for the discriminator. In order to ensure that generator and discriminator are adapted, we update generator twice then update discriminator once. During training, the noise code is sampled element-wise from zero-mean Gaussian with standard deviation 0.001; At testing run, it is sampled in the same manner but with a different standard deviation of 0.1. The training finishes after 20000 mini-batches. In addition, we find the result more robust if the whole model is initialized to a trained one with no pathological loss.

Experiment Results

Dataset and preparation

In this paper, we select three datasets: DRIVE (Staal et al. 2004), STARE (Hoover, Kouznetsova, and Goldbaum 2000) and Kaggle (Kaggle 2016). DRIVE contains 20 training images and 20 test images, with each of size 584×565×3. STARE contains 40 images and the size is 700×605×3. The Kaggle dataset contains 53576 training images and 35118 test images of various size. The DR detector is trained on Kaggle dataset following (Antony 2016). When training generator, we uses binary image and its corresponding retinal image in the DRIVE dataset. Before processing, we change all of the image into size of 512 × 512×3 following (Zhao et al. 2018).

Visualization of Pathological Descriptor

We use images in STARE and Kaggle as pathological references to extract the pathological descriptor. The position of individual lesion is randomly chosen. We have organized a group of 5 ophthalmologists to evaluate our results. After training the generator, we test it on binary vessel images from DRIVE test set. The exemplary result in Fig 6 shows that our pathological descriptors contain appearance and color features of lesions in different types. Microaneurysms in (a) and (b) looks very plausible. The laser scars in (c) are consistent with the fundus of treated RD patients. However, our generated hard exudates in (d) are of some artifacts.

In Fig 7, we compare Tub-sGAN's results with ours. Our method generates images with realistic lesion details where Tub-sGAN fails. For example, in images synthesized by our method (row 2), we can clearly spot the microaneurysms appeared at the far end of vessels. However, the lesions of Tub-sGAN (row 1) could not be classified into any known symptom. In addition, our method could output images with clear vessels, and the optic disc is better with concave appearance.

Quantitative Comparison

To further strengthen our method, we organized a peer review by a board of 5 professional ophthalmologists. The ophthalmologists were asked to double-blindly evaluate the randomly shuffled fundus images synthesized by our method and Fila-sGAN. For each image, they gave three scores

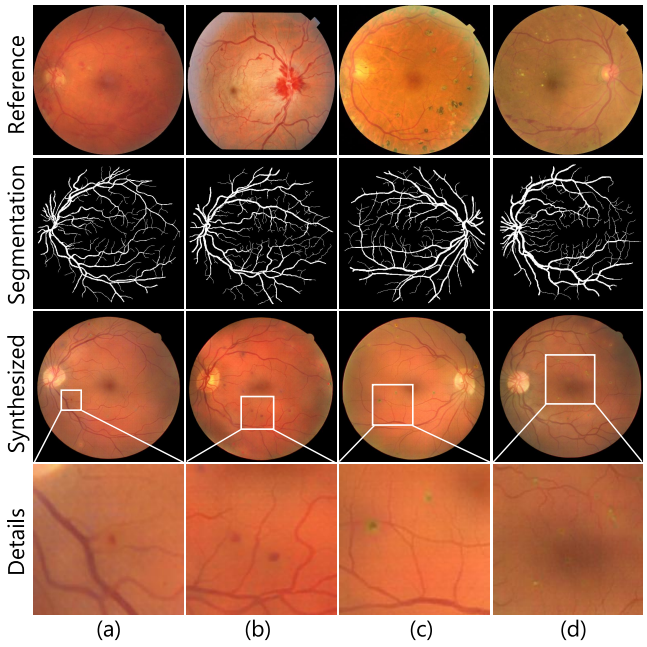

Figure 6: Results of our experiment. We use reference images in row 1 and vessel segmentations in row 2, to generate synthesized retinal fundus, as shown in row 3-4.

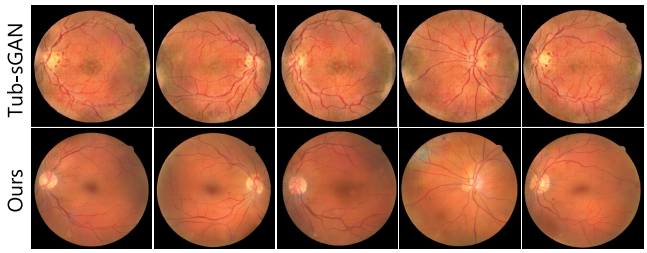

Figure 7: Pathological details comparison between Tub-sGAN and ours.

(ranged 1-10, higher indicates better): 1. realness of the fundus image, 2. realness of the lesions, 3. severity of the DR. Finally, we collected valid scores on 560 images, of which the average scores are show in Table 1. The p value of T-test is 9.80e-10 and 7.95e-5 for fundus and lesion realness respectively that our mean score is higher than Fila-sGAN.

	Score 1	Score 2	Score 3
Tub-sGAN	2.91	2.53	2.53
Ours	4.21	3.37	3.08

Table 1: Average scores from the ophthalmologists.

Lesion Manipulation

As mentioned above, our method is lesion-based, which makes lesion-wise manipulation possible. As shown in

Fig 8, we trained two generators from the same reference image (row 1, col 1) but distribute the lesion to the upper region (row1, col 2&4) and lower region(row1, col 3&5) respectively. On the other hand, we can also control the number of lesions. For example, drop out some of descriptors to generate less lesions, or clone some of descriptors to get more lesions. Row 2 in Fig 8 shows a series of synthesized pictures with lesion number increasing from zero to 3 times of that in the reference image, and the varying severity of the DR symptom could be observed in the fundus images.

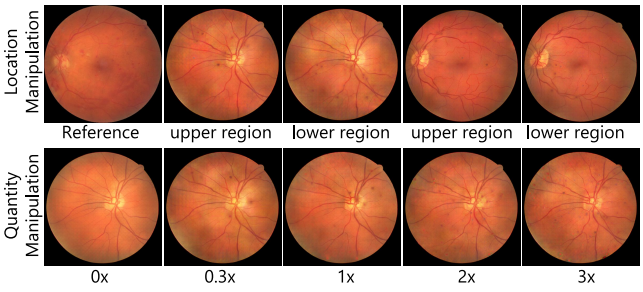

Figure 8: Results of lesion manipulation.

We also evaluate the above synthesized images on the Diabetic Retinopathy Detector (Antony 2016). According to (American Academy of Ophthalmology 2002), number of microaneurysms is an important criteria for the severity diagnose. Here we manipulate the number of microaneurysms and count its severity prediction. For each lesion number, we synthesize 10 images on each segmentation in DRIVE test set which has 20 testing image. Thus, we count the predicted severity for 200 images over each number of lesions. As reported in Fig 9, by manipulating the lesion number, we receive consistency result from DR detector.

Detail Preservation

As we discussed above, it is intuitive to regulate the difference between real and synthesized images on L1 loss as most of style transfer application (Gatys, Ecker, and Bethge 2016). However, this kind of metrics such as MAE, MSE, and SSIM only focus on low-level information in the image. In our practice on retinal image, we find the images generated with L1 loss rather than our retina detail loss fail to preserve some important physiological details such as optic disc boundary and choroid. Here we compare the images generated by Tub-sGAN (Zhao et al. 2018), method with L1 loss and our current application in Fig 10. We can see our method with a retina detail loss is appropriate for synthesizing photorealistic fundus images.

Ablation study

We evaluate the effects of individual components by adjusting their weights.

Retina Detail Loss The retina detail loss serves to preserve physiological detail. When reducing the weight of retina detail loss, the synthesized optic disc blurs, and noises in the image increase, as shown in Fig 11.

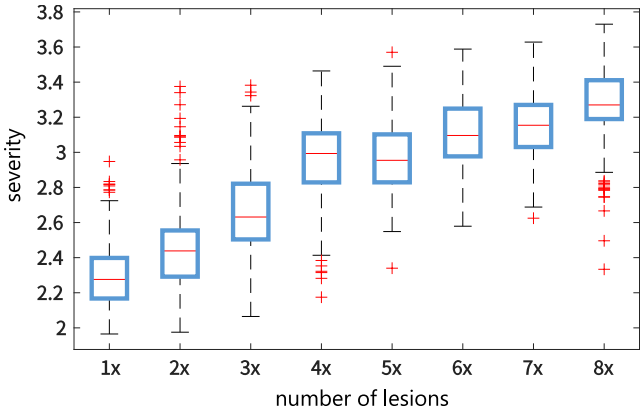

Figure 9: The severity score with increasing lesions number.

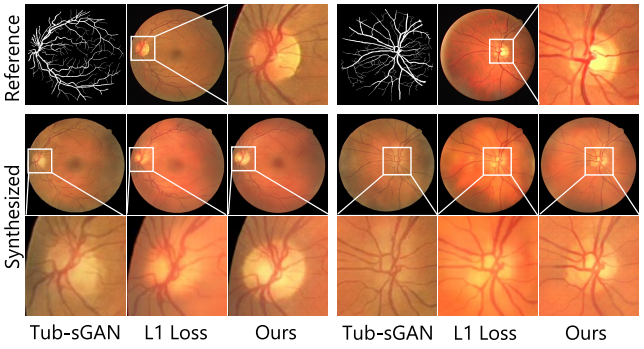

Figure 10: The bright area of optic disc. Row 1 are two real fundus reference images with a clear boundary around each optic disc. Below are fundus synthesized by different methods, among which ours is best in realism.

Pathological Loss The pathological loss controls the synthesized lesions. When reducing the weights of pathological loss, we observe that lesions become weaker and weaker before they disappear. To further confirm this point, we evaluate the severity score of the generated images by DR detector. As Fig 12 shows, the severity increases with the weight of pathological loss. It worth pointing out that the severity score is 0 when the constraints of pathological loss is absent.

Color Transfer

Unlike traditional style transfer, our method focuses on transferring pathological feature rather than color or brightness. Here, we transfer the appearance by applying Deep Photo Style Transfer (Luan et al. 2017) (DPST) after our synthesized image. As shown in Fig 13, compared to the direct synthesized image at row 2, the image after DPST possess a higher color consistency while preserving the pathological lesions such as microneurysms.

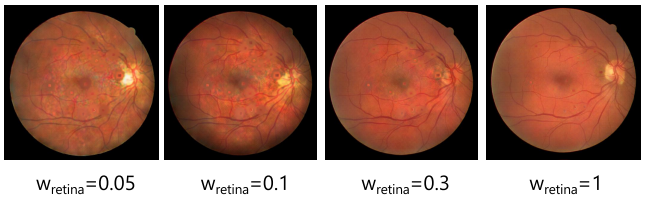

Figure 11: The generated image with increasing weights of retina detail loss.

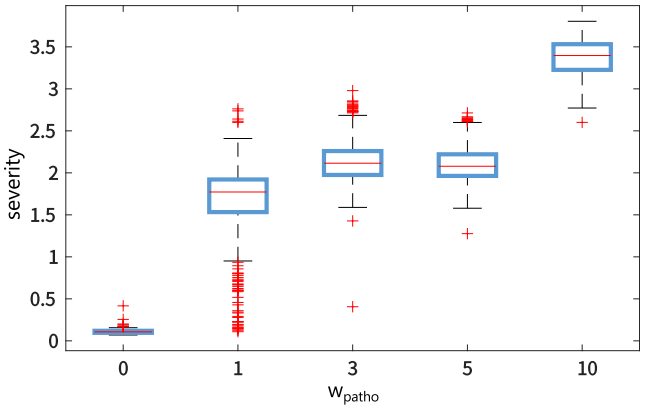

Figure 12: The severity score with increasing weights of pathological loss.

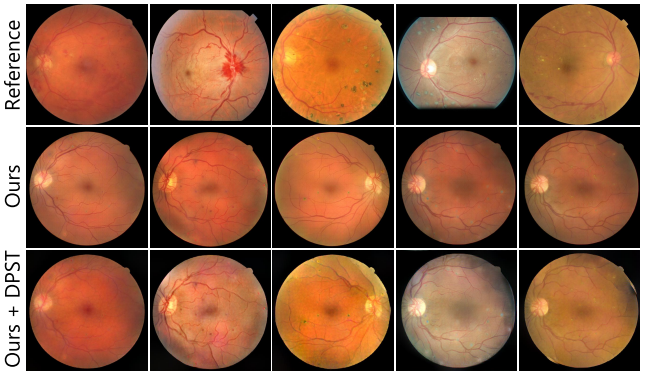

Figure 13: Color transfer results. Our method with a post process achieves a better visual effect.

Computation Time

All the experiments are tested out on a sever with Intel Xeon E5-2643 CPU, 256GB memory and Titan-Xp GPU. Training time on DRIVE and testing time are shown in table 2. Compared to Tub-sGAN (Zhao et al. 2018), we are faster which benefits from a more streamlined feature extraction network and descriptor-based comparison.

Conclusion

To exploit the network interpretability in medical imaging, we proposed a novel strategy to encode the descriptor from the activated neurons that directly related to the prediction. To visually illustrate the extracted pathologic descriptor, we followed the similar methodology of Koch's Postulates that aim to identify the unknown pathogen. In addition, we proposed a GAN based visualization method to visualize the pathological descriptor into a fully controllable pathology retinal image from an unseen binary vessel segmentation. The retinal images we generated have shown medical plausible symptoms as the reference image. Since pathological descriptor is associated with individual lesion and spatial independent, we could arbitrarily manipulate the position and quantity of the symptom. We verified the generated images with a group of licensed ophthalmologists and our result is shown to be both qualitatively and quantitatively superior to state-of-the-art. The feedback of doctors shows our strategy has strengthened their understanding on how deep learning makes prediction. Not limited in interpreting medical imaging, we will extend our strategy to more general interpretability problem.

	Training	Testing	Platform
Tub-sGAN	108/109 min	0.45/0.12 s	Titan-X/Xp
Ours	90 min	0.12 s	Titan-Xp

Table 2: Computation time of the different methods. The Tub-sGAN time on Titan-X is reported in their paper, and its time on Titan-Xp is mesured by us.

Acknowledgement. This work was supported by National Natural Science Foundation of China (NSFC) under Grant 61602020.

References

Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. 2016. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, 265–283.

American Academy of Ophthalmology. 2002. International clinical diabetic retinopathy disease severity scale. https://www.icoph.org/dynamic/attachments/resources/diabetic-retinopathy-detail.pdf.

Antony, M. 2016. Team o_O solution for the kaggle diabetic retinopathy detection challenge. https://www.kaggle.com/c/diabetic-retinopathy-detection/discussion/15807.

Ascoli, G. A., and Krichmar, J. L. 2000. L-neuron: a modeling tool for the efficient generation and parsimonious description of dendritic morphology. *Neurocomputing* 32:1003–1011.

Bonaldi, L.; Menti, E.; Ballerini, L.; Ruggeri, A.; and Trucco, E. 2016. Automatic generation of synthetic retinal fundus images: Vascular network. *Procedia Computer Science* 90:54–60.

Bower, J. M.; Cornelis, H.; and Beeman, D. 2015. Genesis, the general neural simulation system. *Encyclopedia of Computational Neuroscience* 1287–1293.

Carnevale, N. T., and Hines, M. L. 2006. *The NEURON book*. Cambridge University Press.

Dollár, P., and Zitnick, C. L. 2015. Fast edge detection using structured forests. *IEEE TPAMI* 37(8):1558–1570.

Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J.; Swetter, S. M.; Blau, H. M.; and Thrun, S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542:115.

Fiorini, S.; Ballerini, L.; Trucco, E.; and Ruggeri, A. 2014. Automatic generation of synthetic retinal fundus images. In *Eurographics Italian Chapter Conference*, 41–44.

Gatys, L. A.; Ecker, A. S.; Bethge, M.; Hertzmann, A.; and Shechtman, E. 2017. Controlling perceptual factors in neural style transfer. In *IEEE CVPR*.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *IEEE CVPR*, 2414–2423.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Gu, L.; Zheng, Y.; Bise, R.; Sato, I.; Imanishi, N.; and Aiso, S. 2017. Semi-supervised learning for biomedical image segmentation via forest oriented super pixels(voxels). In Descoteaux, M.; Maier-Hein, L.; Franz, A.; Jannin, P.; Collins, D. L.; and Duchesne, S., eds., *MICCAI 2017*, 702–710.

Gulshan; Peng; Coram; and 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.

Hoover, A.; Kouznetsova, V.; and Goldbaum, M. 2000. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE TMI* 19(3):203–210.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. *arXiv preprint*.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 694–711. Springer.

Kaggle. 2016. Kaggle diabetic retinopathy detection challenge. https://www.kaggle.com/c/diabetic-retinopathy-detection.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Luan, F.; Paris, S.; Shechtman, E.; and Bala, K. 2017. Deep photo style transfer. In *IEEE CVPR*, 6997–7005. IEEE.

Maas, A. L.; Hannun, A. Y.; and Ng, A. Y. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, 3.

Mirza, M., and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

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*, 234–241. Springer.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556.

Staal, J.; Abramoff, M.; Niemeijer, M.; Viergever, M.; and van Ginneken, B. 2004. Ridge based vessel segmentation in color images of the retina. *IEEE TMI* 23(4):501–509.

Wang, Z.; Yin, Y.; Shi, J.; Fang, W.; Li, H.; and Wang, X. 2017. Zoom-in-net: Deep mining lesions for diabetic retinopathy detection. In *MICCAI*, 267–275. Springer.

Yang, Y.; Li, T.; Li, W.; Wu, H.; Fan, W.; and Zhang, W. 2017. Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks. In Descoteaux, M.; Maier-Hein, L.; Franz, A.; Jannin, P.; Collins, D. L.; and Duchesne, S., eds., *MICCAI 2017*. Springer International Publishing.

Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *ECCV*, 818–833. Springer.

Zhao, H.; Li, H.; Maurer-Stroh, S.; and Cheng, L. 2018. Synthesizing retinal and neuronal images with generative adversarial nets. *Medical image analysis* 49:14–26.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*.