DATA AUGMENTATION BASED ON PIXEL-LEVEL IMAGE BLEND AND DOMAIN ADAPTATION

Di LIU[1],Xiao-Chun HOU[2],Yan-Bo LIU[3] , Lei Liu[4] , Yan-Cheng Wang[5]

[12345]School of Information and Software Engineering, University of Electronic Science and Technology of China, ChengDu,China

ABSTRACT

Object detection typically requires a large amount of data to ensure detection accuracy. However, it is often impossible to ensure sufficient data in practice. This paper presents a new data augmentation method based on pixel-level image blend and domain adaptation. This method consists of two steps: 1.Image blend using a labeled dataset as object instances and an unlabeled dataset as background images.2. Domain adaptation based on Cycle Generative Adversarial Networks (Cycle GAN).A neural network will be trained to transform samples from step 1 to approximate the original dataset. Statistical consistency between new dataset generated by different data augmentation methods and original dataset will be measured by metrics such as generator loss and hellinger distance. Furthermore, a detection/segmentation network for diabetic retinopathy based on Mask R-CNN will be built and trained by the generated dataset. The effect of data augmentation method on the detection accuracy will be presented.

KEYWORDS

Data Augmentation, Object Detection, Image Blend, Domain Adaptation, Diabetic Retinopathy

1. INTRODUCTION

In the task of object detection, data augmentation of training samples is of great significance, which can reduce over-fitting and improve the generalization performance of the detection models. Traditional data augmentation methods, such as cropping, flipping, and colour jittering, are able to obtain a certain degree of detection accuracy. However, object detection needs to not only recognize different kinds of instances but also distinguish the same instance in different contexts. Therefore, using image blend to expand context information is an effective data augmentation method [1].Image blend is to cut and paste object instances into other background images to obtain new samples which contains object instances. Nevertheless, if object instances are blended randomly with background images, it's possible to generate unreasonable image contexts, which can even degrading the detection accuracy [2], [3]. Related work presents a method of training a deep learning network to predict whether the background image is suitable for image blend with the object instances [4].But when it comes to the object detection of medical images, which has less information, it only needs to judge whether the scale and location of an instances are correct or not.

Samples generated by image blend often have different styles from original object instances. Domain adaptation is a effective method to transform blended image to be close to the original

Natarajan Meghanathan et al. (Eds) : CCSEA, CLOUD, SIPRO, AIFU, SEA, DKMP, NCOM - 2019
pp. 275-285, 2019. © CS & IT-CSCP 2019 DOI: 10.5121/csit.2019.90923

object instances, so as to improve the quality of generated samples. Generative Adversarial Networks (GAN) is a common domain adaptation method recently [5], 6]. On the basis of GAN, a cyclic network called Cycle GAN which is composed of two mirror GAN was presented [8].Cycle GAN is able to transform and reconstruct samples cyclically between source domain and target domain, thus improving the consistency between the real samples and generated samples [9].This paper presents a new data augmentation method combined pixel-level image blend and domain adaptation. And a object detection model of diabetic retinopathy will be established to verify the validity and applicability of this method.

2. Partial Algorithm Principle

2.1 Object Detection Algorithm

Object detection is an important branch of computer vision field. Which is mainly used to locate and recognize object instances with specific features in the image. Traditional methods, such as SIFT [10], SURF [11], DPM [12], mainly devoted to extract local features and match these features to retrieve instances. Few instance samples are needed when using traditional methods. But at the same time, local features extracted by these methods are not 'rich' enough to obtain better detection accuracy. The recent object detection algorithm are based on convolutional neural network (CNN)[13] andregion proposal algorithm to obtain better detection accuracy [14], [15], [16].Mask R-CNN is a representative of such algorithms. It presents a new structure based on feature pyramid network (FPN) [17] and micro Fully Convolutional Networks (FCN) [18]for each region of interest (RoI). Mask R-CNN [19] has excellent detection accuracy and can segment the object instances at the pixel level at the same time. But these algorithms based on CNN require a large amount of labeled dataset to train the detection model. Otherwise, Problems such as over-fitting, low detection accuracy will comes to these algorithms. In conclusion, it is necessary to search a suitable data augmentation method to expand dataset in practice.

2.2 Data Augmentation Algorithm

Data augmentation is to expand datasets by generate new samples with a certain methods. There are some traditional data augmentation methods below:

1. PCA Jittering: Applinga transformation to each pixel $I_{xy} = [I_{xy}^R, I_{xy}^G, I_{xy}^B]^T$of the image. The transformation is defined as :

$$[p_1, p_2, p_3][\alpha_1\lambda_1, \alpha_2\lambda_2, \alpha_3\lambda_3]^T.$$

p_i, λ_i are the eigenvectors and eigenvalues of the covariance matrix of I_{xy}, α_i is a random variable.

2. Noise: such as filter image with Gaussian Blur.
3. Random Scale, Random Crop, Horizontal/Vertical Flip, Shift and Rotation/Reflection, and Color Jittering etc.

Traditional data augmentation method will cause distortion and distortion to the original image. Due to the translation invariance in Mask R-CNN, methods such as shift have no significant effect on the detection accuracy. In contrast, pixel-level image blendis an effective method to generate new image samples.

2.3 Pixel-level Image Blend

According to recent research I, a single instance which is placed in different views, scales, directions, or lighting conditions extracts different features in object detection algorithm based on

CNN. Therefore, image blend is an effective methodto improve the coverage of various context conditions in a dataset [1]. It is necessary to ensure the global and local consistency of the image when generating new image samples. Therefore, it is reasonable to blend background image with pixel-level segmentation mask of the instance object instead of RoI about it. Pixel-level image blend includes the following steps:

1. Collecting object instances: labeled dataset is necessary for object detection. In order to extract object instances in images, pixel-level mask of each image is necessary. These foreground masks can be used to cut and paste object instances to the background image.
2. Collect background images: Background images must not contain the object instances and be similar to the original background of the instances. Otherwise, the differences between instances and background images may lead to useless context information [4].
3. Cut and paste the object instances: Different blending methods can be chosen to paste the object instances into the background images which is randomly selected. By this way, it is possible to ensure that the blending images covers different context information.

However, due to the difference between the object instances and the background images, artifacts may appeared at the edge of the object instances, which cause to a decrease in the global consistency of the blending image.

2.4 Cycle GAN Domain Adaptation

GAN is a generation model based on deep networks that distinguishes the distribution of input data and generates new data samples [7]. GAN usually consists of two sub-networks: one called generator, denoted by $G(z)$, another one called discriminator, denoted by $D(x)$. The generator takes noise data as input and provides the generated data to the discriminator. The discriminator takes real data or generated data as input, then predicts whetherthe input data is real or not. Training this networks corresponds to a minimax two-player game. The generator generates samples closer to real data in the process. The process can be denoted as:

$$\min_{G} \max_{D} V(G,D) \tag{1}$$

$$V(G,D) = \mathbb{E}_{x\sim p_{data}(x)}[log\, D(x)] + \mathbb{E}_{z\sim p_z(z)}\left[log\left(1 - D\big(G(z)\big)\right)\right] \tag{2}$$

$\mathrm{V}(G,D)$ denotes the loss function of the generator and the discriminator. p_{data} is the distribution of the original samples, p_z is the random noise distribution.

Based on GAN, Cycle GAN is used for establishing a mapping from the source domain to the target domain without additional information [8]. Cycle GAN contains two GANs, denoted by (G_{AB}, D_B) and (G_{BA}, D_A), which denotes a cyclic mapping between source domain A to target domain B. Mapping follows the following rules:

$$a \approx G_{BA}\big(G_{AB}(a)\big), b \approx G_{AB}\big(G_{BA}(b)\big) \tag{3}$$

The losses of two GANs can be expressed as:

$$L^{B}_{GAN}(G_{AB}, D_B) = \mathbb{E}_{b\sim p_d(b)}[log\, D_B(b)] + \mathbb{E}_{a\sim p_d(a)}\left[log\left(1 - D_B\big(G_{AB}(a)\big)\right)\right] \tag{4}$$

$$L^{A}_{GAN}(G_{BA}, D_A) = \mathbb{E}_{a\sim p_d(a)}[log\, D_A(a)] + \mathbb{E}_{b\sim p_d(b)}\left[log\left(1 - D_A\big(G_{BA}(b)\big)\right)\right] \tag{5}$$

In order to achieve cyclic consistency and prevent the GAN from mapping the source domain to a single picture in the target domain, cyclic consistent loss is defined as:

$$L_{CYC}(G_{AB}, G_{BA}) = \mathbb{E}_{a\sim p_d(a)}\left\|G_{BA}\big(G_{AB}(a)\big) - a\right\|_1 + \mathbb{E}_{b\sim p_d(b)}\left\|G_{AB}\big(G_{BA}(b)\big) - b\right\|_1 \tag{6}$$

In summary, the total loss of the Cycle GAN can be expressed as:

$$L = L^{B}_{GAN}(G_{AB}, D_B) + L^{A}_{GAN}(G_{BA}, D_A) + L_{CYC}(G_{AB}, G_{BA}) \tag{7}$$

3. EXPERIMENT AND RESULTS ANALYSIS

The paper's approach to treating diabetic retinopathy images uses the method described above and consists of three steps: 1. Data pre-processing, using existing data sets to fuse new data. 2. Data domain conversion, using Cycle GAN transformation to generate data to the target domain of the sample data. 3. Data validity test, using the generated data to train the Mask R-CNN detection network, and test the detection accuracy on the original data set, then analyze the detection metrics. On the basis of the experiment, this paper will compare the similarities and differences between the data generated by this program and the traditional data augmentation method, and analyze the validity and applicability of the method through statistical metrics such as the Intersection over Union (IoU), precision and recall.

3.1 Data

Diabetic retinopathy (DR) is a complication of diabetes that threatens vision and even leading to blindness. DR can be clinically divided into non-proliferative diabetic retinopathy and proliferative diabetic retinopathy. Due to the differences in medical equipment, datasets of DR often have different style. Therefore, it is difficult to obtain a large amount of available data.

All the object instances data in this paper is collected from West China hospital, Sichuan University. The dataset have 547 cases in total. DR manifests as retina hemorrhage, which has irregular texture, boundary and scale. Therefore, this paper mainly focus on instances larger than 20x20 pixels to extract more representative features. The background image data in this paper is from the Diabetic Retinopathy Detection dataset of kaggle 2018. Random combination of object instances and background images is adopted to ensure the diversity of generated data.

3.2 Cut and Paste Blend

It is mentioned above that pixel-level mask is necessary to blend the object instances and background images. But there are different methods to use the masks, such as Cut and Paste or Poisson Blend [7].

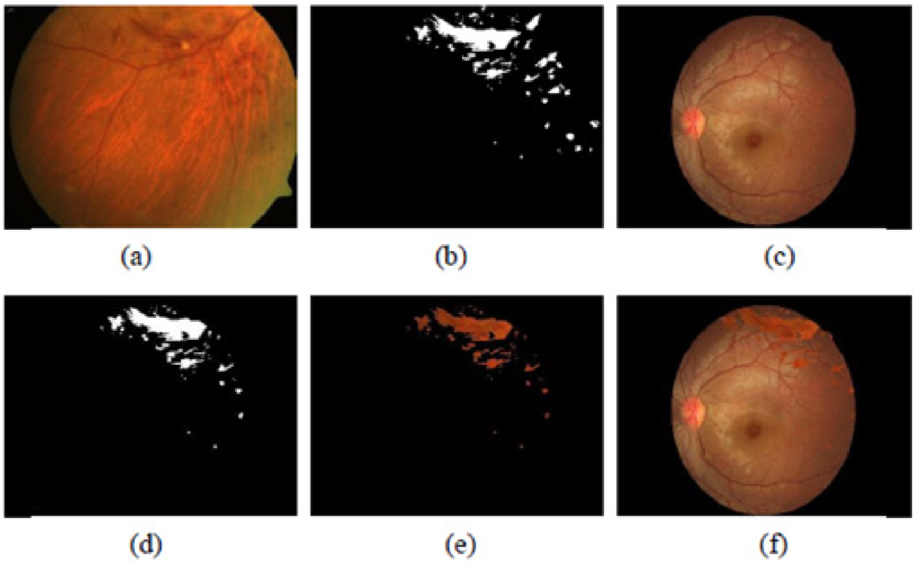


Figure 1. (a) original image (b) mask image (c) background image (d)mask after removing useless region (e) object instance after removing useless region(f) result of Cut and Paste Blend

Cut and Paste method means directly extracting the object instances area and replaces the corresponding pixel in the background images. The experiment is as follows:

Figure 1 (a) is a sample of 547 cases of object instances. Figure 1 (b) is a corresponding mask image, and Figure 1 (c) is a sample of the background image for blending. In order to prevent the object instance from appearing in an unrelated region, the first step is to find out the contour of the eyeball region of Figure 1 (c), and remove the mask where is beyond the eyeball region in Figure 1 (b). The mask after modify is shown in Figure 1 (d). And then, the corresponding object instance pixels in Figure 1 (a) are extracted as Figure. 1 (e) according to Figure 1 (d). Finally, Figure 1 (e) is used to replace the pixels of the corresponding region in Figure 1 (c), so that the blending image Figure 1 (f) is obtained.

3.3 Poisson Blend and Gaussian Blur

Poisson Blend is an image blend method based on the Poisson equation. Laplacian convolution kernel is used to obtain the divergence of each pixel in the image. Poisson Blend establishes a Poisson equation according to the divergence, and calculates the pixel value of the blend image.Poisson Blend can make the difference between the object instances and the background images smoothly diffused into the blend image and finally obtain seamless blend image [20].

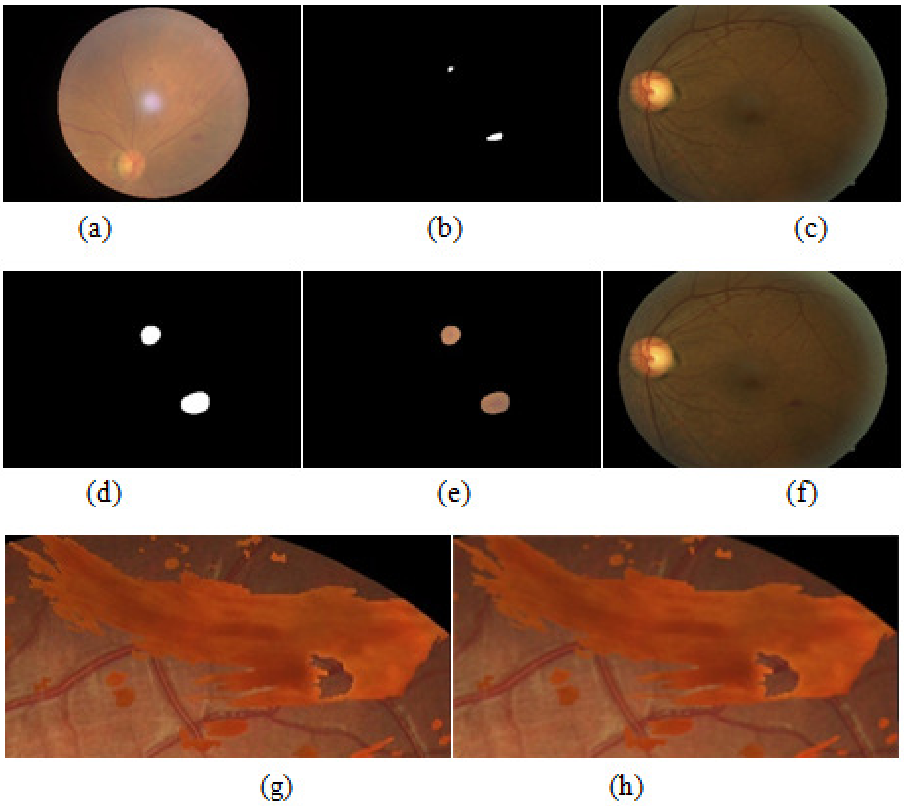


Figure 2. (a) original image (b) mask image (c) background image (d)mask after morphological dilation (e) object instance after morphological dilation (f) result of Poisson Blend (g) instance detail of Cut and Paste Blend (h) instance detail of Gaussian Blur

Poisson Blend need to cut pixels other than object instances edges. Therefore, mask image Figure 2 (b) should be processed with morphological dilation first. And then extracting the object instance pixels according to the mask Figure 2 (d). Figure 2 (e) is the extracted object instance.

Finally, the Poisson reconstruction equation of Figure 2 (c) and Figure 2 (e) is established and solved, so that a blend image obtained. The blending image is shown in Figure 2 (f).

Analyzing the results of the two blend method, it is obvious that Cut and Paste Blend retains the features of the object instance, but there are artifacts at the edge of the object instance, which decrease the global consistency of the blend image. Poisson fusion can achieve seamless blend, but the object instance is hard to recognize in the blend image. In order to maintain the features of the object instance, this paper chooses to use the Cut and Paste Blend and add a Gaussian Blur filter on the edge of the object instance to smooth the image. The result is shown in Figure 2 (h).

3.4 Domain Adaptation

Through the image blend processing above, a new instance sample containing object instance and background image has been obtained. However, due to the difference in distribution between the object instance and the background image, the blending image is still inconsistent with the original image, which may results in inaccurate feature extraction of the detection model. Domain adaptation is used to solve this problem. In this section, the blend images are defined as source domain and the original images are defined as target domain. And a Cycle GAN is trained learn the mapping between source domain and target domain. After training Cycle GAN, the generator G_{AB} can be used to transform blending images to new samples close to the original images.

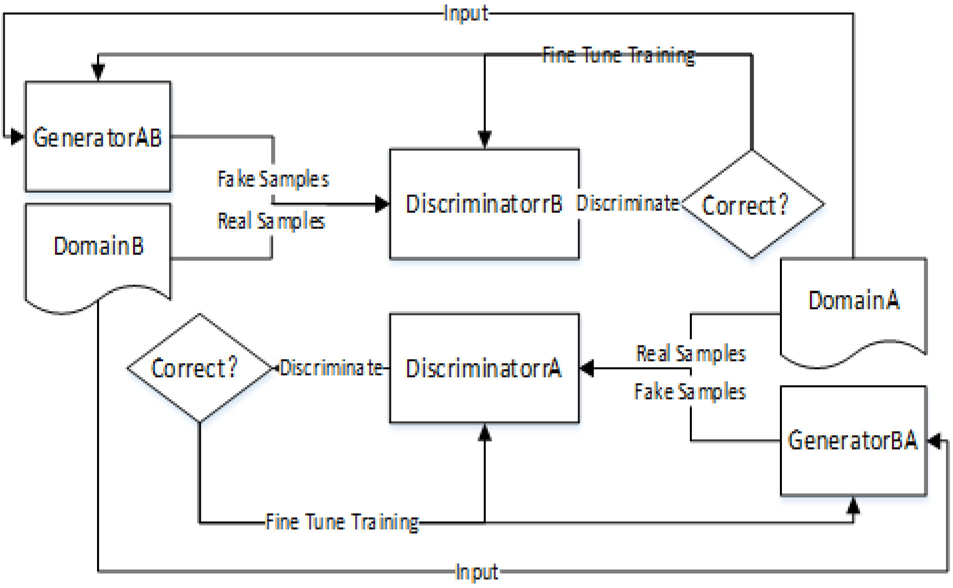


Figure 3 Structure of Cycle GAN Model. Cycle GAN works by training two transformations G_{AB} and G_{BA} between source domain A and target domain B in parallel.

Figure 3 shows the basic structure of the Cycle GAN. Cycle GAN uses the symmetric GANs to perform the same training on the source domain and target domain. But this paper focuses on the transformation of the source domain to the target domain. Therefore, G_{BA} is changed to takes the sample generated by G_{AB} as input instead of samples from domain B, so that two generators can be trained together. In addition, since the generated samples are used for object detection, hellinger distance is combined into cyclic consistent loss to improve the statistical consistency of the generated samples with the original images. The new loss function is defined as:

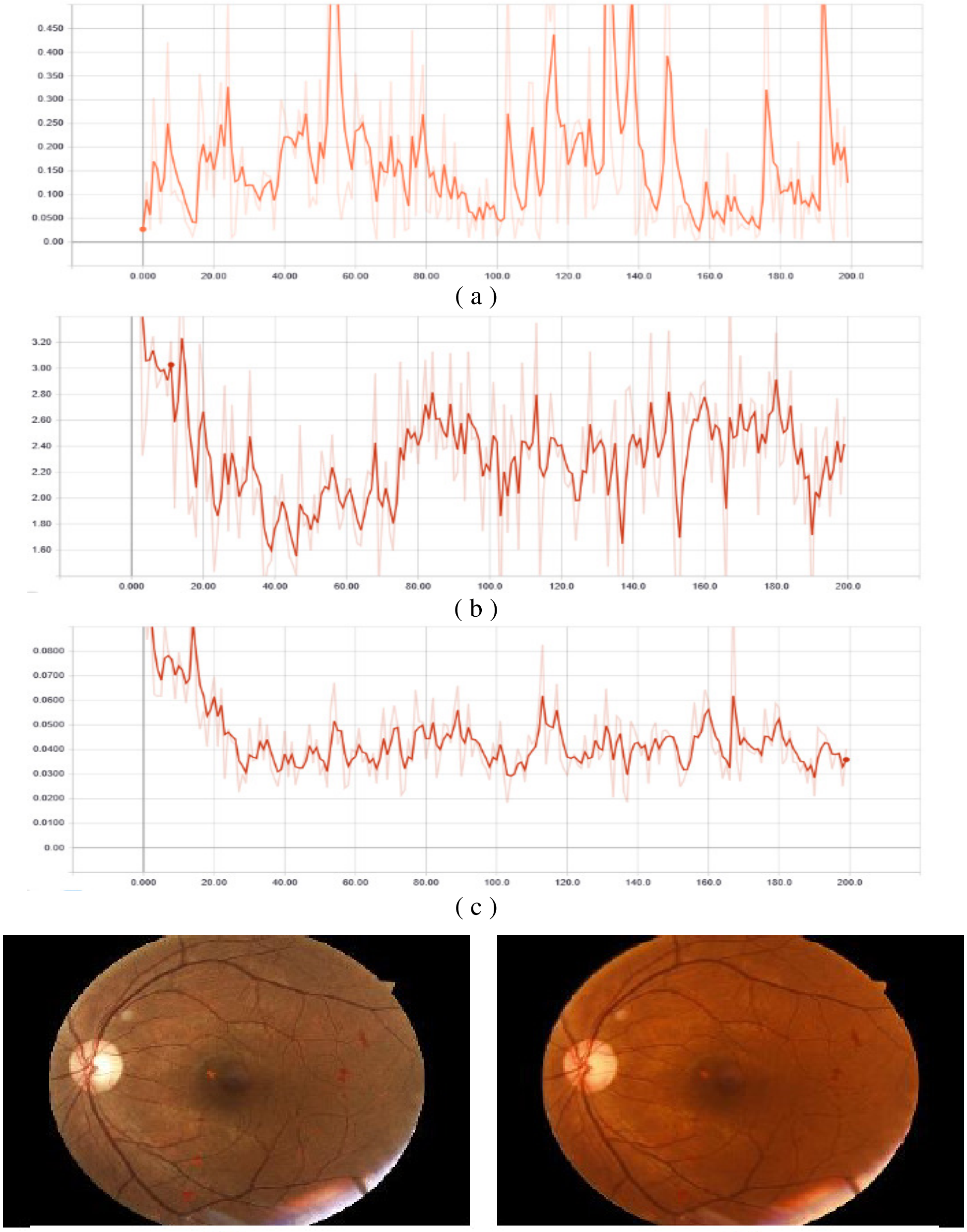
$$L = L_D + L_G + L_C \tag{8}$$

It consists of discriminator loss, generator loss and cyclic loop consistent loss:

$$L_D = \mathbb{E}_{b\sim p_d(b)}[\log D_B(b)] + \mathbb{E}_{a\sim p_d(a)}[\log D_A(a)] \tag{9}$$

$$L_G = \mathbb{E}_{a\sim p_d(a)}\left[log\left(1 - D_B\left(G_{AB}(a)\right)\right)\right] + \mathbb{E}_{a\sim p_d(a)}\left[log\left(1 - D_A\left(G_{BA}\left(G_{AB}(a)\right)\right)\right)\right] \tag{10}$$

$$L_C = \mathbb{E}_{a\sim p_d(a)}\left\|G_{BA}\left(G_{AB}(a)\right) - a\right\|_1 + \mathbb{E}_{a\sim p_d(a)}\frac{1}{\sqrt{2}}\left\|\sqrt{G_{BA}\left(G_{AB}(a)\right)} - \sqrt{a}\right\|_2 \tag{11}$$



(a)

(b)

(c)

(d) (e)

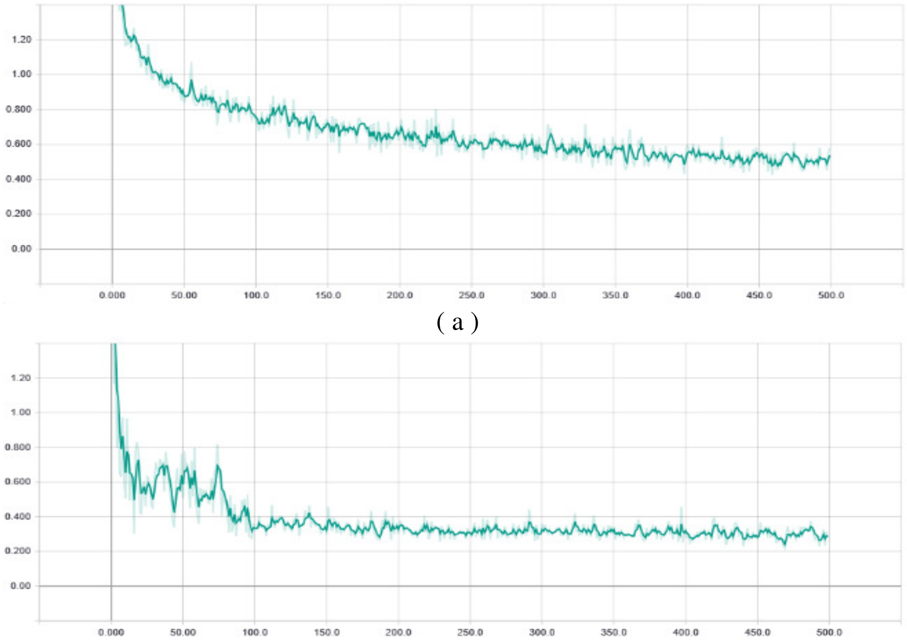
Figure 4. (a) discriminator loss (b) generator loss (c) cyclic consistency loss (d)blending image sample from source domain (e) generated image sample from target domain

Result of training this domain adaptation model is shown in Figure 4 (a) to Figure 4 (c). According to the loss, model in 50th iteration has better performance. Transform the blend image with the trained model, new samples shown in Figure. 4 (e) is obtained.

3.5 Object Detection

To examine the detection accuracy of the generated images. This paper trains the Mask R-CNN detection model with 2000 images generated by Cycle GAN. The original 547 cases of images are used as the test dataset to verify the accuracy of the trained model. For object detection, it is generally considered that the RoI is positive when the IoU is greater than 0.5. This paper use IoU over 0.5 and IoU over 0.75 to calculate the metrics include average precision and average recall. The average precision is denoted by $AP_{0.5}$and $AP_{0.75}$, and the average recall is denoted by $AR_{0.5}$ and $AR_{0.75}$.

The training loss and verification loss of the detection model are shown in Figure 5 (a) and Figure 5 (b). The annotated images and detection results of the test dataset are shown in Figure 5 (c) to Figure 5 (f). After calculation, The IoUs of instances between original image Figures 5 (c) and detection result Figures 5 (d) are 0.81 and 0.79, which results in precision 1.0 and recall 1.0.The IoUs of instances between original image Figures 5 (e) and detection result Figures 5 (f) are 0.77, 0 and 0.75, which results in precision 1.0 and recall 0.67. In order to compare and analyse the quality of the data augmentation methods above, 400 cases of original dataset and 2000 images generated by the traditional data augmentation method are also used to train Mask R-CNN As an experimental comparison. Similar to Figure 5, the metrics of the detection models obtained by the three datasets are calculated, and the results are shown in Table 1.



(a)

(b)

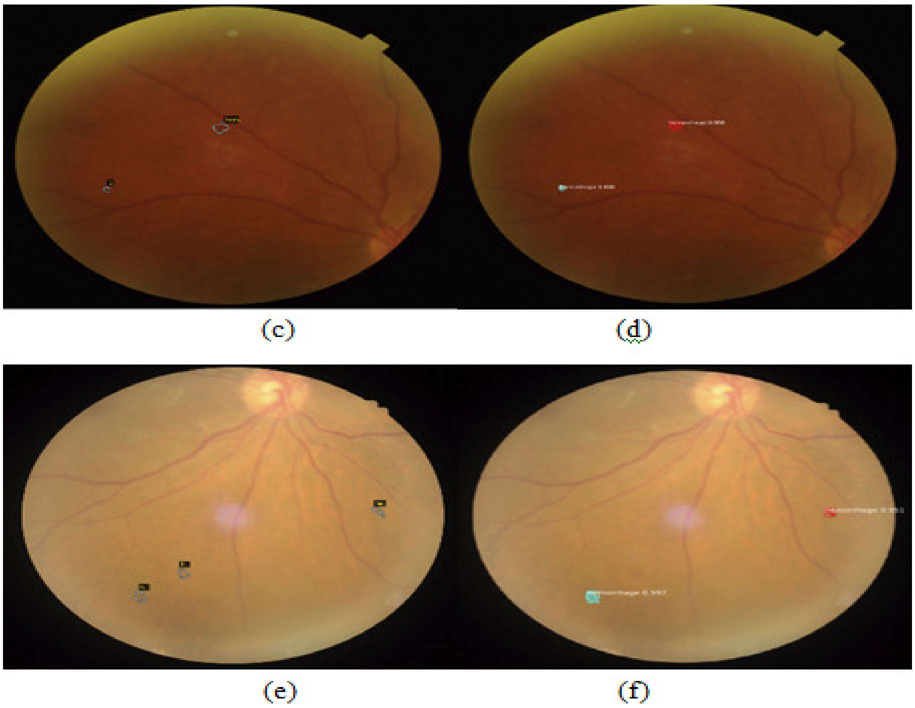


Figure 5. (a) train loss (b) val loss (c) labeled image from dataset (d) detected instance of Figure 5 (c) (e) labeled image from dataset (f) detected instance of Figure 5 (e)

Table 1. Detection results

Data Augmentation Method	$\boldsymbol{AP_{0.5}}$	$\boldsymbol{AP_{0.75}}$	$\boldsymbol{AR_{0.5}}$	$\boldsymbol{AR_{0.75}}$
No data augmentation	0.61	0.46	0.67	0.52
Traditional method	0.72	0.58	0.75	0.59
Image blend & Domain adaptation	0.74	0.61	0.84	0.79

4. CONCLUSION

In this paper, a data augmentation method combining pixel-level image blend and domain adaptation is proposed. By using the augmented data for the detection model training, the effectiveness of different data augmentation is compared.

Analysis of the results of Table 1 shows that the use of data augmentation can effectively reduce over-fitting of the detection model, improve the precision and recall. All data augmentation methods improve the $AP_{0.5}$ and $AR_{0.5}$ of the detection model to more than 0.7 when the training dataset was expanded from 400 cases to 2000 cases. Compared to the detection model without data augmentation. The precision was improved by more than 0.1. And the data augmentation method u combining pixel-level image blend and domain adaptation has a greater improvement on the recall rather than traditional method. The $AR_{0.5}$ reached 0.84 and $AR_{0.75}$ reached 0.79. The improvement of the precision is relatively lower than recall, the $AP_{0.5}$ reached 0.74 and $AP_{0.75}$ reached 0.61.

These results means that the additional context information generated by this method is more effective than the traditional method, so that the model can extract more features of the same object instance, which is beneficial to recognizing the object instance, thereby improving the recall. On the other hand, the fact that the features of the dataset are not obvious enough resulted in some false detection such as Figures 5(e). As a result, the precision is relatively low.

In summary, data augmentation can effectively expand the dataset of the object detection, and solve the over-fitting problem of the object detection model trained by small dataset. The data augmentation method based on pixel-level image blend and domain adaptation has better performance than the traditional method. At the same time, the validation of this method is accomplished in the dataset of medical images which has less contextual information. The validity and applicability of the method still need to be tested and optimized on other datasets. The subsequent work will also continue to optimize and improve the algorithm on the basis of this method.

ACKNOWLEDGEMENTS

The research work was supported by the Fundamental Research Funds for the Central Universities under Grant No. ZYGX2016J092, the Sichuan Science and Technology Project under Grant No. 2017GZ0318, and the Fundamental Research Funds for the Central Universities under Grant No. ZYGX2015J068.

REFERENCES

[1] Dwibedi, D., Misra, I., & Hebert, M. (2017, October). Cut, paste and learn: Surprisingly easy synthesis for instance detection. In The IEEE international conference on computer vision (ICCV).

[2] Karsch, K., Hedau, V., Forsyth, D., &Hoiem, D. (2011). Rendering synthetic objects into legacy photographs. ACM Transactions on Graphics (TOG), 30(6), 157.

[3] Ros, G., Sellart, L., Materzynska, J., Vazquez, D., & Lopez, A. M. (2016). The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3234-3243).

[4] Dvornik, N., Mairal, J., &Schmid, C. (2018). Modeling visual context is key to augmenting object detection datasets. arXiv preprint arXiv:1807.07428.

[5] Chen, Y., Li, W., Sakaridis, C., Dai, D., & Van Gool, L. (2018, March). Domain adaptive faster r-cnn for object detection in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3339-3348).

[6] Volpi, R., Morerio, P., Savarese, S., &Murino, V. (2018, June). Adversarial feature augmentation for unsupervised domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5495-5504).

[7] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ...&Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).

[8] Chu, C., Zhmoginov, A., & Sandler, M. (2017). CycleGAN: a Master of Steganography. arXiv preprint arXiv:1712.02950.

[9] Almahairi, A., Rajeswar, S., Sordoni, A., Bachman, P., &Courville, A. (2018). Augmented CycleGAN: Learning Many-to-Many Mappings from Unpaired Data. arXiv preprint arXiv:1802.10151.

[10] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110.

[11] Bay, H., Tuytelaars, T., & Van Gool, L. (2006, May). Surf: Speeded up robust features. In European conference on computer vision (pp. 404-417). Springer, Berlin, Heidelberg.

[12] Felzenszwalb, P., McAllester, D., &Ramanan, D. (2008, June). A discriminatively trained, multiscale, deformable part model. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on (pp. 1-8). IEEE.

[13] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[14] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[15] Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[16] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[17] Lin, T. Y., Dollár, P., Girshick, R. B., He, K., Hariharan, B., &Belongie, S. J. (2017, July). Feature Pyramid Networks for Object Detection. In CVPR (Vol. 1, No. 2, p. 4).

[18] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[19] He, K., Gkioxari, G., Dollár, P., &Girshick, R. (2017, October). Mask r-cnn. In Computer Vision (ICCV), 2017 IEEE International Conference on (pp. 2980-2988). IEEE.

[20] Pérez, P., Gangnet, M., & Blake, A. (2003). Poisson image editing. ACM Transactions on graphics (TOG), 22(3), 313-318.