Supplementary: Encoding Structure-Texture Relation with P-Net for Anomaly Detection in Retinal Images

Kang Zhou[1,⋆], Yuting Xiao[1,⋆], Jianlong Yang[2], Jun Cheng[3], Wen Liu[1], Weixin Luo[1], Zaiwang Gu[3], Jiang Liu[2,4], and Shenghua Gao[1,5,⋆⋆]

[1] School of Information Science and Technology, ShanghaiTech University, China
{zhoukang, xiaoyt, liuwen, luowx, gaoshh}@shanghaitech.edu.cn
[2] Cixi Institute of Biomedical Engineering, Chinese Academy of Sciences, China
yangjianlong@nimte.ac.cn
[3] UBTech Research, China
juncheng@ieee.org, guzaiwang01@gmail.com
[4] Southern University of Science and Technology, China
liuj@sustech.edu.cn
[5] Shanghai Engineering Research Center of Intelligent Vision and Imaging, China

In this supplementary material, i) we first introduce the network architecture detail of our P-Net; ii) then we analyze the impact of the training samples number for our P-Net; iii) and we validate that the proposed method can generalize well to novel class discovery in retinal images; iv) finally, we show the qualitative results on the publicly available real-world dataset, i.e., MVTec AD [1].

1 Network Architectures of P-Net

Let C_k denote a Convolution-BatchNorm-ReLU layer with k filters. All convolutional operations are 3×3 spatial filters applied with stride 1, which will produce a feature map that has the same spatial size with the input feature map. The scale factor of both the max-pooling layer (denoted as Mp) in the encoder, and the upsampling layer (denoted as Up) in the decoder is 2.

1.1 Structure Extraction Network with Domain Adaptation

Structure extraction network architecture. For implementation, we adopt the U-Net [2] with four max-pooling layers as our structure extraction network. The architecture consists of:

encoder: $C_{64} - C_{64} - Mp - C_{128} - C_{128} - Mp - C_{256} - C_{256} - Mp - C_{512} - C_{512} - Mp - C_{512} - C_{512}$;

decoder: $C_{256} - C_{256} - Up - C_{128} - C_{128} - Up - C_{64} - C_{64} - Up - C_{64} - C_{64} - Up - C_1$.

Discriminator architecture. The architecture of the discriminator for domain adaptation is adopted from [3].

⋆ Kang Zhou and Yuting Xiao equally contribute to this work.
⋆⋆ Shenghua Gao is the corresponding author.

1.2 Image Reconstruction Module

Image and structure encoder architecture. The architecture of both the image encoder and structure encoder is $C_{64} - C_{64} - Mp - C_{128} - C_{128} - Mp - C_{256} - C_{256} - Mp - C_{512} - C_{512} - Mp - C_{512} - C_{512}$.
Decoder architecture. The decoder architecture is $C_{256} - C_{256} - Up - C_{128} - C_{128} - Up - C_{96} - C_{96} - Up - C_{64} - C_{64} - Up - C_m$, where m denotes the channel number of original image.

2 Analysis of The Number of Training Samples in iSee Dataset

We conduct experiments to analyze the impact of the training sample number. The total number of training samples in iSee dataset [4] is 4000, and we adopt different rates to random select the training samples. The results are shown in Table 1, in which we can see that as the number of training samples decreases, the performance is slightly declining. Particularly, we use 1% (i.e., 400) samples of the whole training set to train our P-Net and the model achieves 0.698 AUC, which is 96% (0.698/0.725) of the performance achieved with all training set. These results demonstrate that our model is not sensitive to the number of normal training samples. This is because the diversity of structure in normal fundus images is small. Although we use a small sample for training, the model can still capture the structure in the normal image.

Table 1. The results of different training samples in iSee dataset.

Rate	1%	5%	10%	20%	40%	60%	80%	100%
AUC	0.698	0.699	0.709	0.707	0.705	0.711	0.713	0.725

3 Novel Class Discovery in iSee Dataset

Clinicians can recognize the images with diseases that they have never seen before. It is desirable to achieve this capability for an intelligent diagnosis system. Such problem corresponds to the novel concept discovery setting, *i.e.*, the testing set contains the images falling out the categories in the training set. To evaluate the performance of our method under such setting, we use 4000 normal images and half of the images with AMD, PM, glaucoma, and DR in our iSee dataset as the training set. We use the remaining images for outlier detection in testing phase, which contains other categories of diseases that different from that in training set. We define normal, AMD, PM, glaucoma and DR as inlier and the other diseases as outlier.

We report the performance of different methods under such setting in Table 2. We can see that our method outperforms other methods. The reason for the success of our method is that our network can capture the structure and image contents correlation. Even for the images with certain diseases, such correlation still exists and if they are included in the training set, the network can learn such correlation. While for the unseen types of diseases, the correlation between

structure and image content is different, which would lead to a large reconstruction error and such reconstruction error would help the new disease discovery in testing phase.

Table 2. The results of novel concept discovery on the iSee dataset

Method	Deep SVDD	Auto-Encoder	Pix2Pix	VAE-GAN	GANomaly	Our Method
AUC	0.5794	0.5244	0.5249	0.6172	0.6240	**0.6593**

4 Qualitative Results on MVTec AD Dataset

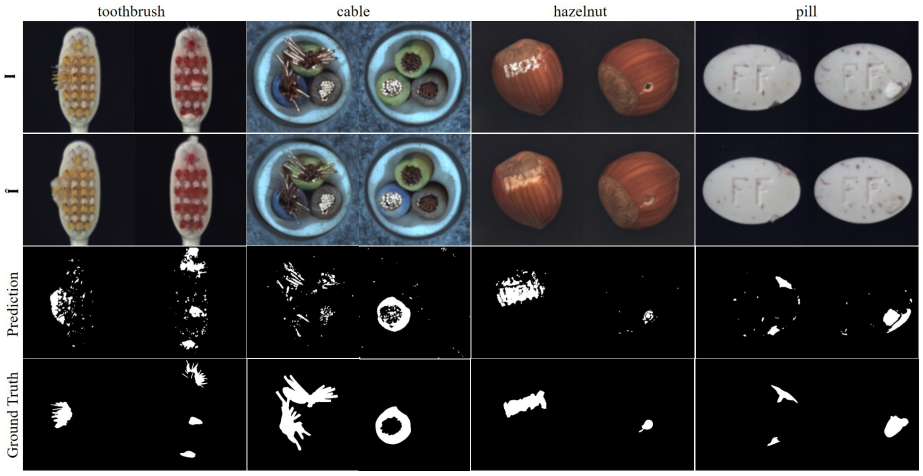
In this section, we analyze the reconstruction performance on different categories of images in MVTec AD dataset. As shown in Fig. 1, we can observe that the abnormal region can not be reconstructed reasonably. For example, the "graffiti" and the "hole" on the hazelnut is poorly reconstructed.

A common existing problem is that the reconstruction of high frequency signal is distorted. We can find this problem at some region such as metal conductors in the cable and the "rough surface" of the pill. This phenomenon is more obvious in texture images than in object images because the surface of texture images is pretty rough. Much more high frequency signal makes reconstruction error higher such as the reconstructed carpet and grid images. Thus, for the texture images, the performance of our method is relatively worse than for the object images.

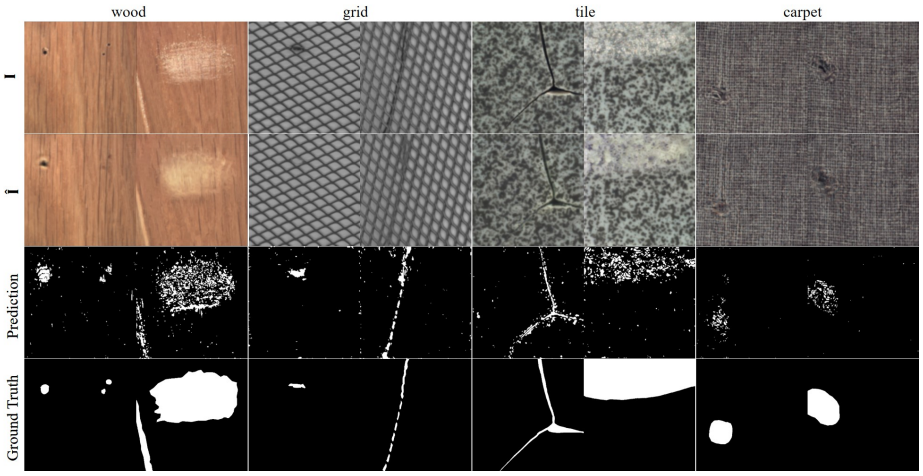
To get the pixel-level prediction of abnormal regions, we follow [1] and define a minimum defect area for normal class data. Then we segment the difference map (i.e., the absolute value of the residual map between $\mathbf{I}$ and $\hat{\mathbf{I}}$) of normal class samples with increased threshold. This process is not stopped until the area of anomaly region is just below the defect area we defined and this threshold is utilized for segmentation anomaly region in testing phase.

References

1. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 9592–9600
2. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer (2015) 234–241
3. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 7472–7481
4. Yan, Y., Tan, M., Xu, Y., Cao, J., Ng, M., Min, H., Wu, Q.: Oversampling for imbalanced data via optimal transport. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 5605–5612



(a) Object categories



(b) Texture categories

Fig. 1. Qualitative reconstruction results of different categories of images. $\mathbf{I}$ and $\hat{\mathbf{I}}$ denotes original image and reconstructed image, respectively. "Ground Truth" denotes the pixel-level abnormal annotation in MVTec AD Dataset, and "Prediction" is the pixel-level abnormal region predictions. To get the "Prediction", we follow [1], and compare the difference between original image $\mathbf{I}$ and reconstructed image $\hat{\mathbf{I}}$.