2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Cycle-Consistent Adversarial Networks for Smoke Detection and Removal in Endoscopic Images

Zhisen Hu and Xiyuan Hu

Abstract— **During endoscopic surgery, smoke removal is important and meaningful for increasing the visual quality of endoscopic images. However, unlike natural image dehaze, it is practical impossible to build a large paired endoscopic image training dataset with/without smoke. Therefore, in this paper, we propose a new approach, called Desmoke-CycleGAN, which combined detection and removal of smoke together, to improve the CycleGAN model for endoscopic image smoke removal. The detector can provide information about smoke locations and densities, which helps the GAN model to be more stable and efficient for smoke removal. Although some imperfections still exist, the experimental results have demonstrated that this method outperforms other state-of-the-art smoke removal approaches with unpaired real endoscopic images.**

Clinical Relevance— **This can help improve the visibility in endoscopic surgery and to get smoke-free endoscopic images with better quality.**

I. Introduction

In endoscopic surgery, with the development of medical imaging technology, doctors can see the internal tissue clearly just as they are looking at the tissue directly. However, some smoke may be generated while performing a surgery, which will heavily degrade the clarity of some details of tissue. Therefore, smoke removal task for endoscopic images is necessary.

Smoggy images can be enhanced by traditional image enhancement methods. Many classic models have been used for de-smoking, such as atmospheric scattering model [1] and dark channel prior [2]. Besides traditional image dehazing approaches, deep learning algorithms have achieved great success these years in many applications, including image dehazing, whether training with paired images or unpaired images. For example, Chen *et al.* proposed a generative cooperative network for endoscopic smoke segmentation and removal (De-smoke GCN) [5]. Sidorov *et al.* propose a method based on Perceptual Adversarial Networks [4] and SSIM-Loss (SSIM-PAN) [3]. However, these methods are supervised learning models which relies on many paired training images. An unsupervised method based on CycleGAN [6] and perceptual loss (CycleDehaze) [8] is proposed by Engin *et al.*. They use a pre-trained VGG [12] network to extract features in smoggy images to improve the quality of de-smoked images. However, for real endoscopic image de-smoke, it still has limitations for avoiding semantic errors. For instance, without a reference image, they may remove some devices or generate strange new devices by mistake since the devices are kind of similar to smoke in color.

And thus, in this paper, we add a smoke detection network into the CycleGAN-based framework for detecting the locations and densities (i.e., a single channel mask with values between 0 and 1) of smoke. This extra information can improve the training procedure of CycleGAN and achieve a better de-smoke result. Furthermore, during the training process, the detector is updated synchronized with its output from an initial binary mask (can only reflect smoke location) to a [0,1] normalized grayscale image (can reflect both location and density of smoke). In addition, a cycle-consistency perceptual loss and newly proposed cyclic detection loss have been incorporated to improve the stability of detector and generator. Through experimental evaluations, this method can be proved more effective compared to other state-of-the-art approaches trained with unpaired images.

This paper is structured as follows. In section II, we discuss our proposed method. Then in section III, we present the experiments and evaluation results. Finally, the conclusions are drawn in section IV.

II. Proposed Methods

The basic GAN model [10] is composed of a generator and a discriminator. The generator aims to find a mapping $\hat{x} = G(z;\theta_G)$ that maps latent random variables to generated data. To optimize this mapping, the discriminator is trained to recognize fake and real samples. CycleGAN [6] provides a solution of unsupervised image-to-image translation. The framework is comprised of two GANs to find two-directional mappings between two domains. Besides the GAN loss, a cyclic loss function forces the original image and the reconstructed image to be consistent with each other.

We adopt the CycleGAN [6] model as our basic model for endoscopic image de-smoke. Unlike general image-to-image translation, we incorporate some prior knowledge of smoke (e.g., location and density), derived by a detection sub-network, into the CycleGAN framework to improve endoscopic image de-smoke. In addition, two cyclic consistency loss and a smooth L1 loss have been added to improve the robustness and stability of the detector and generator networks.

Zhisen Hu is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (email: huzhisen1117@gmail.com).

Xiyuan Hu is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (corresponding author to provide email: huxy@njust.edu.cn).

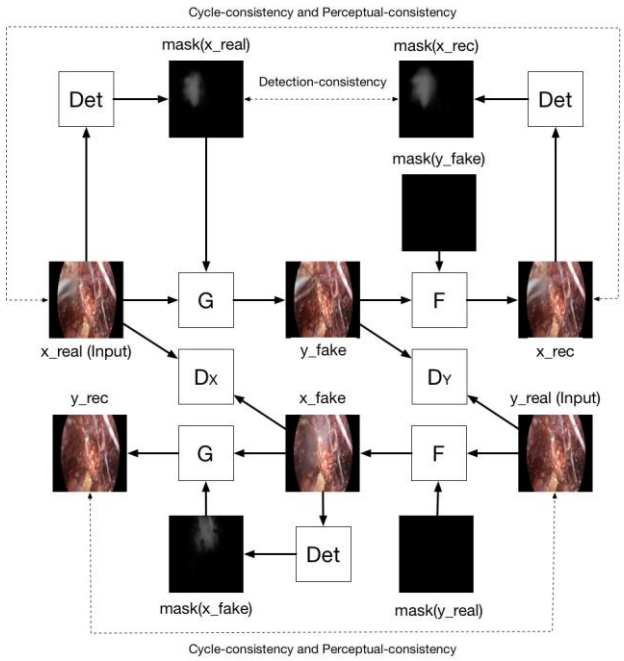

Figure 1. The architecture of our method, where G & F refer to the generators, D_X & D_Y to the discriminators, and Det to the detector. This architecture can be split into two parts: smoggy to smoke-free image ($X \to Y$), and smoke-free to smoggy image ($Y \to X$).

A. Our De-smoke Networks

The overall architecture of our proposed endoscopic image de-smoke model is shown in Fig.1, where G & F refer to the generators for translating images between smoggy domain and smoke-free domain, D_X & D_Y to the discriminators for discriminating whether the generated images are real or fake, and *Det* to the detector for segmenting smoke mask from smoke-free images. Each image is a 3-channel RGB image, and each mask is a 1-channel gray-scale image. Because we incorporate a detector for providing extra smoke information, the input data is a four-channel image including RGB and an additional smoke mask. Smoggy image and its mask are put into the generator together as a 4-channel tensor and then output a 3-channel smoke-free image. As for the smoke-free input, we set its mask as a zero matrix in the same size as the input, and they will be put in another generator together to output a 3-channel smoggy image. A pre-trained U-Net [9] structure is employed to provide an initial binary smoke segmentation mask from original images, which can represent the location of smoke. To provide extra information about the density of smoke, we modified the binary mask into a gray-scale image mask, which is in zero-to-one closed continuous interval value, during the training process. Moreover, the continuous smoke detector is optimized synchronously during the training of GAN model. A cyclic detection-consistency loss is proposed in our work, which can add regularizations to the differences between the masks detected from real smoggy images and similarly reconstructed smoggy images.

B. Loss Functions

There are different functions for different parts of the method.

Adversarial Loss: We use the adversarial loss [10] for two mapping functions ($X \to Y$ and $Y \to X$). As for one of the mappings, such as $X \to Y$, the loss function can be expressed as:

$$L_{GAN}\left(G, D_Y, X, Y\right) = E_{y\sim p_{data}(y)}\left[\log D_Y\left(y\right)\right] + E_{x\sim p_{data}(x)}\left[1-\log D_Y\left(G\left(x\right)\right)\right], \quad (1)$$

where x and y , separately, are real images in domain X and Y , and G is the generator that transfer the images in domain X to images which look similar in domain Y., while D_Y is the discriminator that tries to differentiate between the generated samples and real samples in domain Y . Similarly, we can derive the other function $L_{GAN}\left(F, D_X, Y, X\right)$ in the same way.

Cycle-Consistency Loss: we apply the cycle-consistency loss [6] to bring the reconstructed image back to original image, which can improve the accuracy of the mapping from input to output in target domain. This can be described as:

$$L_{cyc}\left(G,F\right) = E_{x\sim p_{data}(x)}\left[\left\|F\left(G\left(x\right)\right)-x\right\|_1\right] + E_{y\sim p_{data}(y)}\left[\left\|G\left(F\left(y\right)\right)-y\right\|_1\right], \quad (2)$$

where $F\left(G(x)\right)$ and $G\left(F(y)\right)$ are reconstructed images in domain X and Y .

Cyclic Perceptual-Consistency Loss: Both adversarial loss [10] and cycle-consistency loss [6] use only pixel-level feature, which cannot recover all the missed information when the smoggy region in the image is heavily blurry. With perceptual-consistency loss [8], high-level and low-level features can be extracted properly. We use VGG16 [12] architecture as extractor, to extract features from its $2nd$ and $5th$ pooling layers. Then, this loss [8] can be described as:

$$L_{per}\left(G,F\right) = E_{x\sim p_{data}(x)}\left[\left\|\phi\left(x\right)-\phi\left(F\left(G\left(x\right)\right)\right)\right\|_2^2\right] + E_{y\sim p_{data}(y)}\left[\left\|\phi\left(y\right)-\phi\left(G\left(F\left(y\right)\right)\right)\right\|_2^2\right], \quad (3)$$

where ϕ is a VGG16 [12] feature extractor.

Cyclic Detection-Consistency Loss: This loss is used for training smoke detector network. Because the quality of images generated by the generator is usually poor at the initial stage, the detector network is not being trained during this stage. We use a pre-trained smoke detector instead. When the reconstructed smoggy images are similarly consistent with the real smoggy images, we start to train the detector with a detection consistency loss. We assume that the smoke detected from reconstructed images and real images should be consistent, which requires the results of detector should be similar for the original image and reconstructed image. This

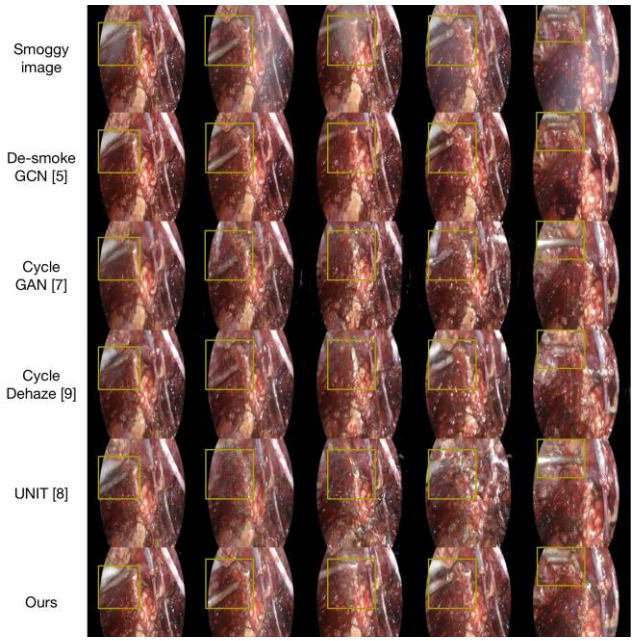

Figure 2. Qualitative results on real testing dataset.

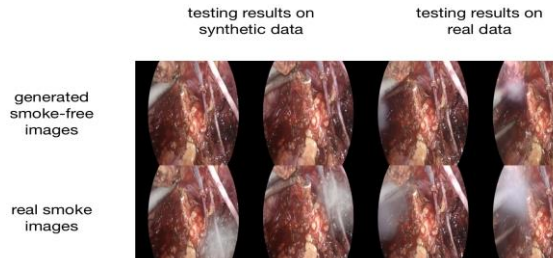

Figure 3. Comparisons on synthetic and real testing dataset with De-smoke GCN.

improves the stability of the detector while training the model. This newly proposed cyclic detection-consistency loss is with the following form:

$$L_{de}(G,F) = E_{x\sim p_{data}(x)}\left[\left\|d(x) - d\left(F\left(G(x)\right)\right)\right\|_1\right], \quad (4)$$

where, d is the smoke mask detected from smoggy images by detector.

Smoothness Loss: In fact, the smoke is supposed to be continuous and smooth, taking penalties for discontinuity can be reasonable. Therefore, we take the L1 norms of the detected smoke masks' gradients along x_d and y_d directions for image smoothing. This function can be expressed as:

$$\begin{aligned} L_{smooth}(d) = &\sum_{x_d,y_d}\left|d(x_d+1,y_d) - d(x_d,y_d)\right| \\ &+ \sum_{x_d,y_d}\left|d(x_d,y_d+1) - d(x_d,y_d)\right|. \end{aligned} \quad (5)$$

Full Objective Function: Thus, the full objective of loss function in our model can be described as follows:

$$\begin{aligned} L = &L_{GAN}(G,D_Y,X,Y) + L_{GAN}(F,D_X,Y,X) \\ &+\lambda_c L_{cyc}(G,F) + \lambda_p L_{per}(G,F) + \lambda_d L_{de}(G,F) \\ &+\lambda_s\left(L_{smooth}\left(d(x)\right) + L_{smooth}\left(d\left(F(y)\right)\right)\right), \end{aligned} \quad (6)$$

where L is the full objective, G and F stand for two generators, D_X and D_Y stand for the discriminators, and λ s are the hyper-parameters for each loss function.

III. Experiments and Results

A. Implementation Details

We used PyTorch framework for all of the training and testing process. A simulated pair-matched smoggy dataset has been used for training an initial smoke detector network as the pre-trained detector. We trained the model with NVIDIA Geforce RTX 2080Ti graphics card (11GB Graphic Memory). In the experiment, we used ADAM optimizer with an initial learning rate of 0.0001, batchsize 4, and 400 training epochs.

As for the architecture, we used U-Net [9] structure for the detector. This structure is composed of five convolutional blocks as an encoder to abstract the feature from input images into 1024-channel tensors. Symmetrically, the decoder consists of five de-convolutional blocks, which is helpful to recover tensors into one-channel smoke masks. Apart from these blocks, skip connections are used to transfer high-level information to the bottom of the network. We used the same generator architecture as CycleGAN [6], which contains two stride-2 convolutions, residual blocks [13], and two other fractionally-strided convolutions with stride 1/2. Specifically, we used 9 blocks for our tasks and apply instance normalization. Similar to CycleGAN [6], we used 70×70 PatchGAN [11], which tries to distinguish whether 70×70 image patches are real or not, and upsampling layers coupled with convolutional layers were used to replace the de-convolutional layered in case of checkerboard artifacts. The VGG16 [12] network used for cyclic-perceptual consistency loss is a pre-trained model in ImageNet [14]. Finally, in order to stabilize the training process, we applied Wasserstein GAN [20] and spectral normalization [15].

B. Dataset

All of the datasets we used in this experiment is captured from Leonardo Da Vinci surgical robot video which was found on Intuitive official website[1], so there were no original paired images. As for the GAN model, we totally used hundreds of unpaired images captured from Leonardo Da Vinci surgical robot video for training and testing.

C. Experimental Results

We made comparisons between related state-of-the-art unsupervised methods, such as CycleGAN [6], CycleDehaze [8], and UNIT [7], and a supervised method, De-smoke GCN [5]. The smoke-free images generated by original CycleGAN [6] are not good enough, since it has a poor performance while processing partly smoggy endoscopic images. For instance, some semantic errors may occur after hundreds of epochs of training, and some of the results may be blurred because there

[1] https://www.intuitive.com/

TABLE I. QUANTITATIVE RESULTS COMPARED WITH STATE-OF-THE ART METHODS

Method	Evaluation Metrics		
	NIQE	PIQUE	FID
Cycle-GAN	47.59	10.67	147.33
Cycle-Dehaze	39.00	**4.50**	170.19
UNIT	39.37	5.71	181.97
De-smoke GCN	64.94	8.60	142.56
Ours	**36.81**	**4.77**	**134.15**

are too many subtle structures in endoscopic images, which are hard to recover. CycleDehaze [8] improved some details in the images, although some errors still exist. As can be shown in Figure 2, compared with those methods, the results generated by our model have the fewest semantic errors and the closest color with the original images. Although De-smoke GCN [5] framework showed satisfactory image quality while testing, it is a supervised training approach with synthetic dataset. When the real smoke images are not in the training dataset, it may suffer from incomplete smoke removal, as is shown in Figure 3. Moreover, this supervised method may suffer from abnormal image color, as can be shown in Figure 2.

As for the evaluation, the lack of paired data was kind of troublesome, because some traditional evaluation metrics cannot be used, such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [16]. Therefore, we use some no-reference image quality evaluation score e.g., Perceptual based Image Quality Evaluator (PIQUE) [17] and Natural Image Quality Evaluator (NIQE) [18] to evaluate the quality of smoke-free images generated by our model. Moreover, Frechet Inception Distance score (FID) [19], which is usually used for evaluating GAN performance, is also adopted as one of comparison metrics. Our framework outperformed these existing methods in terms of these three evaluation metrics, as shown in Table 1. The lower these metrics indicate that the generated images have less distortion and fewer differences compared with high-quality images.

IV. CONCLUSION

In this work, we propose an approach of endoscopic image smoke removal with cycle consistency GAN and smoke detection subnetwork. Besides the GAN loss and the cycle-consistency loss, we also incorporate cyclic detection loss and smoothness loss for training the detector. With the assistance of detector, semantic errors have been reduced and detailed structures have been recovered. Since the endoscopic images are filled with subtle details, such as different medical devices and diverse colors of tissue. A few details are still not generated well, which deserves our future study.

REFERENCES

[1] S. Narasimhan and S. Nayar, "Contrast restoration of weather degraded images," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 6, pp. 713-724, Jun 2003.

[2] K. He, J. Sun, and X. Tan, "Single image haze removal using dark channel prior," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 12, pp. 2341-2353, Dec 2011.

[3] O. Sidorov, C. Wang, and F. A. Cheikh, "Generative smoke removal," arXiv preprint arXiv:1902.00311, 2019.

[4] C. Wang, C. Xu, C. Wang, and D. Tao, "Perceptual adversarial networks for image-to-image transformation," IEEE Transactions on Image Processing, vol. 27, no. 8, pp. 4066-4079, 2018.

[5] L. Chen, W. Tang, N. W. John, T. R. Wan, and J. J. Zhang, "De-smokeGCN: Generative cooperative networks for joint surgical smoke detection and removal," IEEE Transactions on Medical Imaging, vol. 39, no. 5, pp. 1615-1625, 2020.

[6] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in 2017 IEEE International Conference on Computer Vision (ICCV), Oct 2017, pp. 2242-2251.

[7] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks" in 2017 Conference and Workshop on Neural Information Processing Systems (NIPS), Dec 2017, pp. 700-708.

[8] D. Engin, A. Genç,and H. K. Ekenel, "Cycle-Dehaze: Enhanced cyclegan for single image dehazing," in 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Jun 2018, pp. 825-833.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in 2015 International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Oct 2015, pp. 234-241.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative Adversarial Nets," in 2014 Conference and Workshop on Neural Information Processing Systems (NIPS), Dec 2014, pp. 2672-2680.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul 2017, pp. 5967-5976.

[12] K. Simonyan, and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in 2015 International Conference on Learning Representations (ICLR), May 2015.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 2016, pp. 770-778.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 2009.

[15] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in 2018 International Conference on Learning Representations (ICLR), May 2018.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, 2004.

[17] N. Venkatanath, D. Praneeth, M. C. Bh, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in 2015 IEEE Twenty First National Conference on Communications (NCC), Feb 2015, pp. 1-6.

[18] A. Mittal, R. Soundararajan, and A. Bovik, "Making a completely blind image quality analyzer," IEEE Signal Processing Letters, vol. 20, no. 3, pp. 209-212, Mar 2013.

[19] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in 2017 International Conference on Neural Information Processing Systems (NIPS), Dec 2017, pp. 6629-6640.

[20] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," CoRR, abs/1701.07875, 2017.