Performance Evaluation of a Generative Adversarial Network for Deblurring Mobile-phone Cervical Images

Prasanth Ganesan, M.S.*†, Zhiyun Xue, Ph.D.*, Sanjana Singh*, Rodney Long, M.A.*,
Behnaz Ghoraani, Ph.D.†, Sameer Antani, Ph.D.*‡, *Senior Member, IEEE*

Abstract— **Visual examination forms an integral part of cervical cancer screening. With the recent rise in smartphone-based health technologies, capturing cervical images using a smartphone camera for telemedicine and automated screening is gaining popularity. However, such images are highly prone to image corruption, typically out-of-focus target or camera shake blur. In this paper, we applied a generative adversarial network (GAN) to deblur mobile-phone cervical (MC) images, and we evaluate the deblur quality using various measures. Our evaluation process is three-fold: first, we calculate the peak signal to noise ratio (PSNR) and the structural similarity (SSIM) of a test dataset with ground truth availability. Next, we calculate the perception based image quality evaluator (PIQE) score of a test dataset without ground truth availability. Finally, we classify a dataset of blurred and the corresponding deblurred images into normal/abnormal MC images. The resulting change in classification accuracy was our final assessment. Our evaluation experiments show that deblurring of MC images can potentially improve the accuracy of both manual and automated cancerous lesion screening.**

Keywords: Generative Adversarial Network, Cervical Image Deblurring, Uterine Cervix Cancer Classification

I. Introduction

Visual analysis of the uterine cervix is routinely used as a cervical cancer screening procedure. Images acquired during the gynecological exam using smartphones can be used for expert telemedicine assessment or for automated decision making. Although such mobile-phone cervical (MC) imaging technologies (e.g., EVA COLPO, MobileODT Inc., Tel Aviv, Israel) provide an easy-to-use alternative for traditional colposcopy cameras, they have the limitation of image contamination due to blur caused by poorly focused target or camera shake. Such image blur introduces distortion, making it harder for the clinicians to examine for lesion sites or it may result in erroneous automated screening. Hence, an automatic deblurring system would be valuable for reducing image contamination and thereby alleviating resulting ill-effects. In this paper, we evaluate the performance of a GAN-based deblurring method for restoration of focal-blurred MC images.

GANs have been used for a wide variety of applications since its introduction by Goodfellow *et al.* [1]. The Vanilla-GAN was primarily modeled to learn the joint distribution of the input data (training set), but later, the introduction of conditional-GAN (C-GAN) [2] enabled the model to learn specific I/O image correlations, thereby leading to the conception of various blur-restoration GAN models. Previously, GANs have been modeled for various image restoration tasks such as image deblurring with paired training data [3], [4], [5]; image deconvolution, including deblurring, decompression, and denoising (GAN-D) using VGGNet as the discriminator network [6]; video deblurring using 3D convolutional network embedded in a GAN (DBLRGAN) [7], motion deblurring using a densely connected GAN [8]; and image denoising by learning the noise distribution in the absence of clean data [9].

In this paper, we adopt an existing GAN model that was developed for motion deblurring of natural images (deblurGAN) [4] for performing the task of focal deblurring of MC images. We train the model (hereafter referred as MC-deblurGAN) on paired blur and matching sharp MC images and then extensively evaluate its performance using both metric-based and empirical (classifier-based) techniques. Previous works have been done in medical image denoising and deblurring applied to CT and PET images [10], [11], [12], however, to the best of our knowledge, our work is the first application of a deblurring GAN on colposcopy images. Our evaluation shows that the images output by MC-deblurGAN are of significantly higher quality than their blurred input counterpart. More importantly, our preliminary evaluation on images with biopsy-validation results shows that deblurring has the potential to improve the detection accuracy of cervical abnormality.

II. Method

A blurred image is basically a convolution of the sharp image with a blur kernel. This blur kernel differs based on the type and amount of blur. The primary challenge of the MC-deblurGAN is to directly learn the blur-reversal function without the kernel estimation step. A blurred image can be mathematically represented as follows:

$$I_b = I_s * k(X) + N \tag{1}$$

where, I_b is the blurred image, I_s is the sharp image, $k(X)$ is the kernel function over variable X, and N is additive noise.

The MC-deblurGAN essentially predicts I_s from I_b by estimating a residual correction factor (I_r, $I_r = I_s - I_b$). Although attempts have been made by some researchers to model this prediction using unpaired blur and sharp images in the training set [13], it is relatively easier to achieve loss saturation and higher accuracy if we train using a ground truth (GT) pair. Hence we apply the deblurGAN model with paired training [4] for deblurring MC images.

*National Library of Medicine, National Institutes of Health, Bethesda, MD
†Dept. of Electrical Engg., Florida Atlantic University, Boca Raton, FL
‡Corresponding author, Email: sameer.antani@nih.gov

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

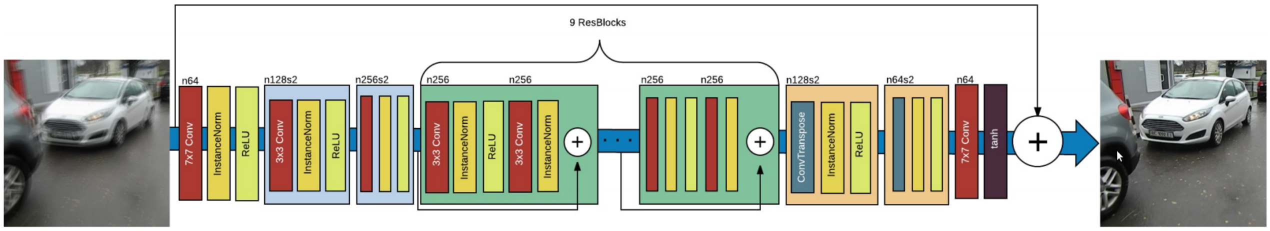

Fig. 1. Generator architecture of DeblurGAN – Model adopted for MC-deblurGAN. Image source: [4]

The following sections give an overview of the existing deblurGAN model and then discuss the MC-deblurGAN dataset and training details.

A. DeblurGAN Model Overview

A GAN consists of a generator and a discriminator network, which are interconnected to play a min-max game until the generator learns the training set distribution. In case of the deblurGAN model [4], the network is a C-GAN, so the goal of the generator is to learn only the blur-sharp distribution in the training set, while the discriminator is trained to classify blur and sharp images.

As shown in Figure 1 (figure adopted from the deblurGAN manuscript [4]), the generator is a CNN consisting of two strided convolutions (stride=1/2), nine residual blocks with dropouts after each conv-1 layer, and then two transposed convolution blocks. The discriminator is a Wasserstein GAN (WGAN) with LeakyReLU and gradient penalty. The loss function is a combination of adversarial loss and content loss. The adversarial loss is a WGAN loss with a gradient penalty, that provides a higher stability over the least squares loss. The content loss is a differential L2 loss, also known as perceptual loss, that avoids blurry artifacts which could otherwise be introduced with traditional mean absolute error (MAE) or mean squared error (MSE) loss functions. More details on the networks and other parameters can be found in the deblurGAN manuscript [4].

B. Dataset and Training Details of MC-deblurGAN

The dataset was provided by MobileODT Inc., Tel Aviv, Israel, under special agreement with the National Institutes of Health, Maryland, USA. The cervical images were collected using a mobile cervical screening device called EVA Colpo. The images were visually filtered for undesirable characteristics, such as iodine-stained images and images with special color filters. The resulting filtered dataset consisted of 2400 sharp and 1209 blurred labeled MC images (see Figure 2).

To train the model, paired blur and matching sharp MC images are required. The sharp images form the ground-truth (GT). However, as described above, our dataset consists of unpaired blur and sharp images. So, we prepared our training set by manually introducing blur to the sharp MC images. First, we split the 2400 sharp images into 2048 for training and the rest for testing. Note that we already have 1209 blurred MC images which is a different test set (see various test sets shown in Figure 2. Second, we apply randomized blur convolution filters to the 2048 training images and 352 test images. The following section describes more about these blur filters.

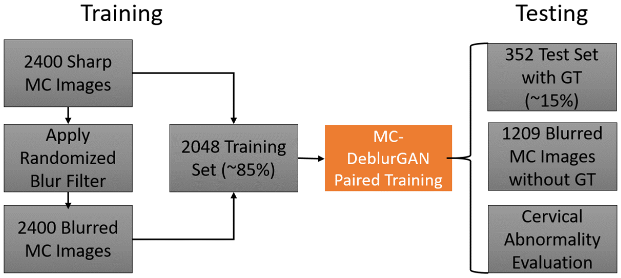
Training
Testing
2400 Sharp MC Images
Apply Randomized Blur Filter
2400 Blurred MC Images
2048 Training Set (~85%)
MC-DeblurGAN Paired Training
352 Test Set with GT (~15%)
1209 Blurred MC Images without GT
Cervical Abnormality Evaluation

Fig. 2. Dataset, training and evaluation Details – Training set is first prepared by introduction of manual blur and the MC-deblurGAN is trained with the blur-sharp pairs. The evaluation is performed on three different test sets for efficient assessment of the model performance.

Introduction of Focal Blur: We used Gaussian and Box (Average) blur kernels in a randomized manner to introduce blur in the sharp images. We not only randomized the blur type, but also the kernel size which defines the amount of blur created. The size was randomized between 3 and 6 pixel units in both x and y directions. Such randomizations to the training set helps to avoid overfitting of the GAN to a particular blur function, and hence offers more generalization.

The network was trained with a batch size of 8 on a high-performance machine with NVIDIA GTX 1080Ti GPU and 48GB RAM. The loss saturation occurred after 300 epochs (final learning rate=0), which took about 28 hours to achieve.

III. Performance Evaluation of MC-deblurGAN

The performance of the trained model was evaluated using various techniques. Since our manually blurred test set consists of GT sharp images, but the naturally collected MC blurred images do not have GT images, conventional evaluation metrics, such as SSIM and PSNR are not applicable for the latter. Hence, we designed a three step evaluation process, which provides a detailed view into the model performance, thus exposing the pros and cons of MC-deblurGAN. Table I gives a summary of all the evaluation results of the presented study.

A. Evaluation Using Images With Ground Truth

First, we evaluated the trained model on the test set of 352 images (see Section II-B) with manually introduced blur.

Since we have the GT for these images, we determined the PSNR and SSIM metrics for this test set.

PSNR defines the amount of clean pixels present in the image with reference to the GT image. Higher PSNR implies higher image quality. SSIM defines the similarity in structure between the target image and the GT. SSIM is determined pixel-by-pixel, this results in the SSIM map (last column in Figure 3). An SSIM value for an image is calculated as the average of all the pixel-wise SSIM values. Higher SSIM implies that the target image is very close to GT.

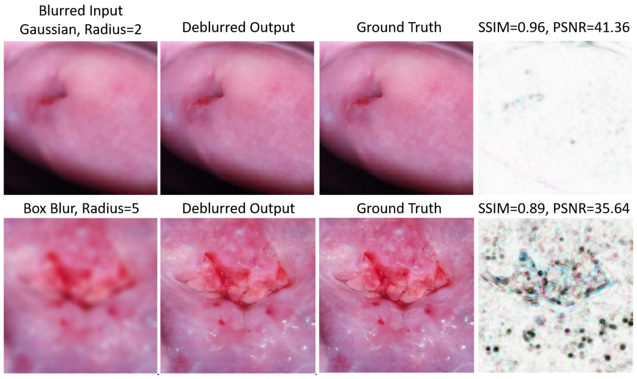

Fig. 3. Evaluation using SSIM and PSNR – The PSNR and SSIM are found to be high (top row) for images with less intensity changes and small blur radius, and vice-versa for low values.

For 352 test images, the value of PSNR was found to be 36.9 ± 2.76 and that of SSIM was found to be 0.92 ± 0.04. Some examples of images with high and low PSNR and SSIM is shown in Figure 3. Values of $SSIM > 0.9$ and $PSNR > 30$ are generally considered to be of a good quality image. Our values indicate that the MC-deblurGAN is indeed helpful in recovering the images with reasonable quality. However, further evaluation (performed in the upcoming sections) is required to fully explore the usefulness of the model in the detection of abnormal images.

From the PSNR and SSIM values, it can be observed that the PSNR has a correlation with intensity variations present in the image, and the SSIM is correlated with the amount of blur present in the image. In other words, higher intensity variations (multiple sharp edges) is associated with lower PSNR, and higher blur kernel size is associated with lower SSIM. This is also evident intuitively, since higher blur kernel would destroy the edges in the image thus modifying the overall structure. This structure is difficult to recover by deblurring, resulting in low SSIM. Similarly, higher intensity variations would lead to more erroneous pixels while deblurring, thus reducing the PSNR.

B. Evaluation Using Images Without Ground Truth

For our test set without GT ($N = 1209$, see Figure 2), we must evaluate the model based on only the output images of MC-deblurGAN. For this purpose, we applied a no-reference image quality metric called PIQE – Perception-based Image Quality Evaluator [14]. PIQE is an unsupervised technique, which provides an image quality score between 0 and 100, 0 implying sharp and 100, blurred. The method works by estimating a block-wise distortion – starting from a Mean Subtracted Contrast Normalization, and then calculating a mask on the distorted artifact blocks present in the image. Then the PIQE score is calculated as the mean of the scores in these distorted blocks. According to the Mathworks® documentation on the PIQE score, the quality of the image can be assessed using the range of PIQE score defined as: "Excellent" if 0-20, "Good" if 21-35, "Fair" if 36-50, "Poor" if 51-80, and "Bad" if 81-100.

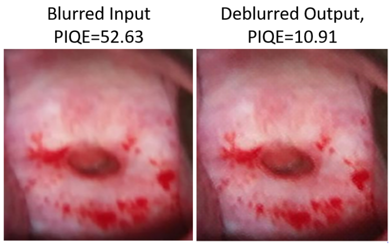

Fig. 4. Evaluation on test set without GT – The deblurred output is visually assessed to be of improved quality. The PIQE score is also improved from 52.63 ("Poor") to 10.91 ("Excellent")

We applied the PIQE metric on both the blurred input images and the deblurred output images of the MC-deblurGAN. The PIQE score for the blurred images was 63.9 ± 22.08 and that for the deblurred images was 9.57 ± 4.2, demonstrating a significant increase of 6-fold in the image quality. This indicates that the images which were "Poor" before deblurring fall under the category of "Excellent" after the deblurring was performed. An example of a blurred and deblurred image along with the associated PIQE scores is shown in Figure 4. This evaluation shows that the model has generalized well, because, even after training the model on images with manual blur functions, the image quality improves on the naturally blurred MC image dataset. However, on visually assessing the deblurred images (eg. in Figure 4), the improvement in image quality is not apparent, so this limitation must be addressed in future studies possibly by collecting a natural blur-sharp pair of MC images for the training set.

C. Evaluation Using Images labeled with Cervical Abnormalities

Apart from the test set with and without GT discussed above, we evaluated the model on another test set with blurred (captured naturally) MC images with the following manual binary labels – normal (healthy cervix, CIN1) or abnormal (CIN2+, cancer). The goal of this evaluation is to verify the MC-deblurGAN's effectiveness for abnormality detection. For this purpose, we first applied these abnormality-labeled MC images through a deep-network classifier and stored the output labels of the classifier. Second, we deblurred these images using the MC-deblurGAN and then applied the output deblurred images to the same classifier and stored the new labels. We finally compared the output classifier-labels of the blurred versus that of the deblurred images.

TABLE I

Performance Evaluation Result Summary – The table shows that the MC-deblurGAN produces higher quality images. Δ refers to the difference in score before and after automatic deblurring.

Dataset	Evaluation Method	Score	Conclusion
Manually Blurred Testset with Ground Truth	Avg. PSNR	36.9	>30, good pixel intensity recovery
	Avg. SSIM	0.92	>0.9, good structure recovery
Naturally Blurred Testset without Ground Truth	Δ Avg. PIQE	-54.33	3 steps improvement in PIQE range
	Avg. PIQE of Deblurred images	9.57	Falls in "Excellent" PIQE range
Naturally Blurred Abnormality-Labelled Testset	Δ Avg. PIQE	-63.3	3 steps improvement in PIQE range
	Avg. PIQE of Deblurred images	13.1	Falls in "Excellent" PIQE range
	ΔClassifier accuracy	+21.4%	Improved from 3 to 0 False Positives

To perform this evaluation, we first manually cleaned up our abnormality-labeled test set by filtering out based on the same criteria as done for the training set of MC-deblurGAN. Due to its retrospective nature, it was difficult to find a blurred ”Abnormal” labeled image. In order to circumvent this situation, without loss of generality, we retained $N = 14$ images of all ”Normal” label as our test set.

For classifying normal images from abnormal images, we had trained a model based on an object detection network (Faster RCNN) using a set of biopsy validated MC images.

We first applied all 14 normal images to the classifier. Second, we deblurred the 14 images using our MC-delurGAN, and then applied these deblurred images to the same classfier. The classification accuracy for the blurred images was 78.6% (11/14 was classified correctly), whereas the classification accuracy for the deblurred was 100% – the images, which were previously misclassified as ”Abnormal” (3 false positives) were now correctly classified as ”Normal” (0 false positives). This increase (+21.4%) in accuracy provides a proof-of-concept that the MC-deblurGAN could assist in improving abnormality classification.

To further assess the quality of the deblurred ”Normal” images, we applied the PIQE metric (discussed in the previous section) on all 14 images before and after deblurring. The PIQE score before deblurring was found to be 76.4 ± 20.1 and the score after deblurring was 13.1 ± 5.9. This clearly shows an improvement in quality from ”Poor” category to ”Excellent” category.

IV. Conclusion

We adopted an existing GAN model for the task of deblurring cervical colposcopy images captured using a mobile-phone. We proposed a three-step evaluation technique - evaluation on test set with GT images using PSNR and SSIM, on test set without GT images using PIQE, and finally, on test set with labeled abnormalities using a deep-net-based abnormality classifier. Our evaluations demonstrate that the image quality significantly increases after deblurring, and that the model generalizes well over various blur functions. In addition, a study on a small set of images offers a proof-of-concept that deblurring of MC images could improve the accuracy of automated cervical abnormality screening. The results can be potentially improved in the future by collecting a paired MC image dataset and training on these natural blur-sharp pairs.

V. Acknowledgement

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC). We also thank MobileODT for providing the data used in this study.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[2] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[3] Thekke Madam Nimisha, Akash Kumar Singh, and Ambasamudram N Rajagopalan. Blur-invariant deep learning for blind-deblurring. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4762–4770. IEEE, 2017.

[4] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8183–8192, 2018.

[5] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3883–3891, 2017.

[6] Ha Yeon Lee, Jin Myung Kwak, Byunghyun Ban, Seon Jin Na, Se Ra Lee, and Heung-Kyu Lee. Gan-d: Generative adversarial networks for image deconvolution. In *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 132–137. IEEE, 2017.

[7] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Wei Liu, and Hongdong Li. Adversarial spatio-temporal learning for video deblurring. *IEEE Transactions on Image Processing*, 28(1):291–301, 2019.

[8] Sainandan Ramakrishnan, Shubham Pachori, Aalok Gangopadhyay, and Shanmuganathan Raman. Deep generative filter for motion deblurring. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2993–3000, 2017.

[9] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*, 2018.

[10] Jelmer M Wolterink, Tim Leiner, Max A Viergever, and Ivana Išgum. Generative adversarial networks for noise reduction in low-dose ct. *IEEE transactions on medical imaging*, 36(12):2536–2545, 2017.

[11] Yan Wang, Biting Yu, Lei Wang, Chen Zu, David S Lalush, Weili Lin, Xi Wu, Jiliu Zhou, Dinggang Shen, and Luping Zhou. 3d conditional generative adversarial networks for high-quality pet image estimation at low dose. *NeuroImage*, 174:550–562, 2018.

[12] Karim Armanious, Chenming Yang, Marc Fischer, Thomas Küstner, Konstantin Nikolaou, Sergios Gatidis, and Bin Yang. Medgan: Medical image translation using gans. *arXiv preprint arXiv:1806.06397*, 2018.

[13] Quan Yuan, Junxia Li, Lingwei Zhang, Zhefu Wu, and Guangyu Liu. Blind motion deblurring with cycle generative adversarial networks. *arXiv preprint arXiv:1901.01641*, 2019.

[14] N Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. Blind image quality evaluation using perception based features. In *Twenty First National Conference on Communications (NCC)*, pages 1–6. IEEE, 2015.