Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science International Medical Journal

Medicine Science 2021;10(3):1002-7

A generative adversarial network based super-resolution approach for capsule endoscopy images

Mehmet Turan

Bogazici University, Institute of Biomedical Engineering , Istanbul, Turkey

Received 25 June 2021; Accepted 20 August 2021
Available online 23.08.2021 with doi: 10.5455/medscience.2021.06.218

Abstract

Medical wireless capsule endoscopy is an effective method for diagnosis and evaluation of gastrointestinal diseases. However, due to energy and size limitations, it produces low-resolution images, which makes it difficult to detect and diagnose the abnormality and may even lead to an incorrect diagnosis. Recently, endoscopy methods with improved resolution have been shown to be more effective than conventional endoscopy approaches in disease detection and characterization, and it is expected that they will have the same success in the field of capsule endoscopy. In this study, a novel method based on deep learning techniques is proposed that can generate high-resolution counterparts of low-resolution endoscopic images. Conditional GANs and spatial attention blocks are combined to increase the resolution by 8x, 10x and 12x. Extensive qualitative and quantitative analyses show that the proposed method is more successful than the recent deep super-resolution technologies, such as Deep Back Projection Network (DBPN) and Residual Channel Attention Networks (RCAN).

Keywords: Capsule endoscopy, deep super resolution, spatial attention blocks, generative adversarial networks

Introduction

The development of Deep Learning (DL) based Super-Resolution (SR) algorithms has accelerated with the increase in the amount of available data and advances in graphics cards. Glasner [1] proposed an approach to achieve successful SR results by using inter-scale patch redundancy in the image. In Gu's study [2], it has been shown that focusing on the entire image instead of overlapping image patches leads to better SR image results. In [3], learning-based detail synthesis was combined with an edge-driven SR algorithm to reconstruct image details at higher frequencies. In another study [4], a nonlinear mapping of low resolution (LR) and high resolution (HR) images is learned using convolutional neural networks (CNN). Another research group [5] investigated the conversion of endoscopy images from LR to HR using an analysis approach based on a convex set. Medical capsule endoscopy is an effective medical device technology for examination, and diagnosis of gastrointestinal organ diseases. Although the minimally invasive methods of capsule endoscopy are evolving day by day, capsule endoscopy has several limitations which impede its wider clinical application, including insufficient precision of motion control, limited field of view, and low image resolution provided by capsule cameras [6]. On the other hand, with the use of deep learning algorithms, it has become possible to successfully and robustly apply methods such as depth estimation, polyp detection and characterization on endoscopic images, making the resolution quality of endoscopy images of significant importance, as the performance of SR can directly influence the success of these subsequent techniques [7,8].

Although it is possible to increase image resolution by adding hardware equipment such as lens and sensor in the camera used, research has been done in the field of computer vision to computationally convert LR images into HR images, as hardware-based solutions generally require additional space, energy consumption, or cost [9,10]. In that sense, recent studies have resulted in significant advances with the increase in image resolution enabling better anomaly detection, region segmentation, and 3D reconstruction [11,12]. In this study, a novel deep SR method for image analysis in capsule endoscopy is presented and summarised in Figure 1.

***Corresponding Author:** Mehmet Turan, Bogazici University, Institute of Biomedical Engineering , Istanbul, Turkey
E-mail: mehmet.turan@boun.edu.tr

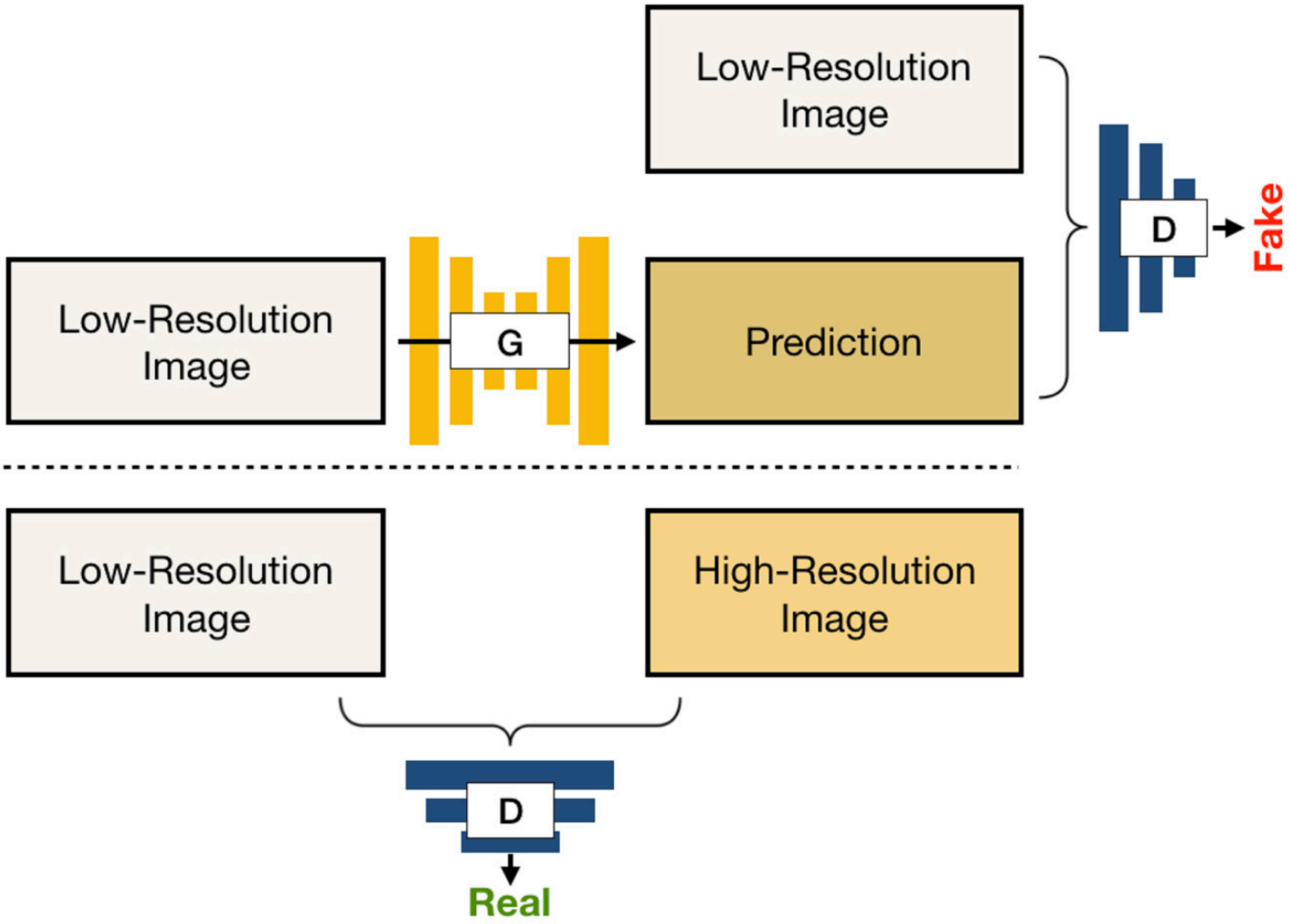

Figure 1. System architecture. A conditional GAN architecture is used to generate HR counterparts from LR input images. In contrast to the traditional unconditional GAN approach, low-resolution images are given as input to both generator and discriminator networks

Materials and Methods

Super-resolution (SR) is the process of generating high-resolution (HR) images from their low-resolution (LR) counterparts. In the following equations, G^{HR} and G^{LR} stand for high and low resolution images, whereas I and δ denote the reduction function and reduction coefficient, respectively. LR image is used as input for inverse degradation function [13] and the main purpose of this function is to use LR images as input to approximate target HR images as reference by the following procedure:

$$G^{LR}=I(G^{HR};\delta) \quad (1)$$

$$G^{SR}=F(G^{LR};\theta_F) \quad (2)$$

Herein, G^{SR} stands for generated HR images, F stands for the HR model, and θ_F stands for the coefficients of this model. The unidentified degradation procedure is affected by different factors including sensor noise, pixel noise, misfocusing, and compression artifacts that may complicate the SR process. Therefore, the degradation process is modeled as a combination of various distortion effects as follows:

$$I(G^{HR};\delta) = (G^{HR} \otimes \kappa) \downarrow r + n\varsigma, \{\kappa, r, \varsigma\} \subset \delta, \quad (3)$$

$G^{HR} \otimes \kappa$ refers to the convolution operation of a blur kernel and a high-resolution image, r is the undersampling operation, ↓r is the down-scaling coefficient, nς is the additive white Gaussian noise (AWGN) with the standard deviation ς. The main goal of the SR method is specified in the following function:

$$\hat{\theta}_F = argmin(L^{SR}(G^{SR}, G^{HR}) + \lambda \phi_F(\theta_F) \quad (4)$$

where $L^{SR}(G^{SR},G^{HR})$ refers to the loss function evaluated to minimize the difference between the reference HR image and generated SR image, which are denoted by G^{HR} and G^{SR}, respectively. θ_F shows the coefficients of the models, ϕ_F (θ_F) denotes the regularized term and λ refers to the variance compensation term. The architectural model, shown in Figure 2, is basically made up of four components, namely the spatial attention block (SAB), the hybrid loss function as well as the generator and discriminator networks. In this study, the attention U-Net generator network is defined as U and the G^{SR} is constructed with weight and bias parameters $\theta_U=\{W_{1:L};b_{1:L}\}$. In addition, a differential U model is created to distinguish the HR and SR images, which are the reference image and generated high resolution image, respectively. Thus, the hybrid loss function L_{obj} makes it possible to obtain SR endoscopic images with the desired

properties by optimizing U iteratively. The objective function is:

$$\theta\hat{}_U = arg\frac{1}{N}\sum_{n=1}^{N} L^{SR}\left(U_{\theta_U}, (G_n^{LR}), G_n^{HR}\right) \quad (5)$$

where the total number of training sets is N, G_n^{LR} and G_n^{HR}indicate the input LR image and the corresponding HR image, respectively. The adversarial min-max optimization objective function is used during training of generator and discriminator [14].

Following the first convolutional layer, the structure of the generator network contains a SAB, which is integrated into the U-Net architecture [15]. This attention module is specifically designed with the intuition to focus on clinically more relevant regions. The SAB module accomplishes its task by assigning higher scores to regions that are of utmost importance according to the relevance of the capsule endoscopy domain. In this process, the given input is encoded by successive convolutional layers until the characteristic features of the image are compressed into a latent vector, which is decoded afterwards. In addition to the characteristic features, the low-level information common to both the input images and their counterparts, the HR images, passes through the layers of the U-Net to preserve the positional information of different tissue pixels, which is important for diagnosis. To establish this information flow accurately and efficiently, the standard architecture of the U-Net is preferred, where n is the layer number and each layer i is connected to the n-i layer via skip connections. Moreover, the spatial-attention unit is capable of focusing on tissue regions with disease-related abnormalities due to the presence of greater gradient changes. To reduce the risk of overfitting the model and speed-up convergence, normalization is applied to each mini-batch [16]. The overall architecture of the encoder and decoder is depicted in Figure 2.

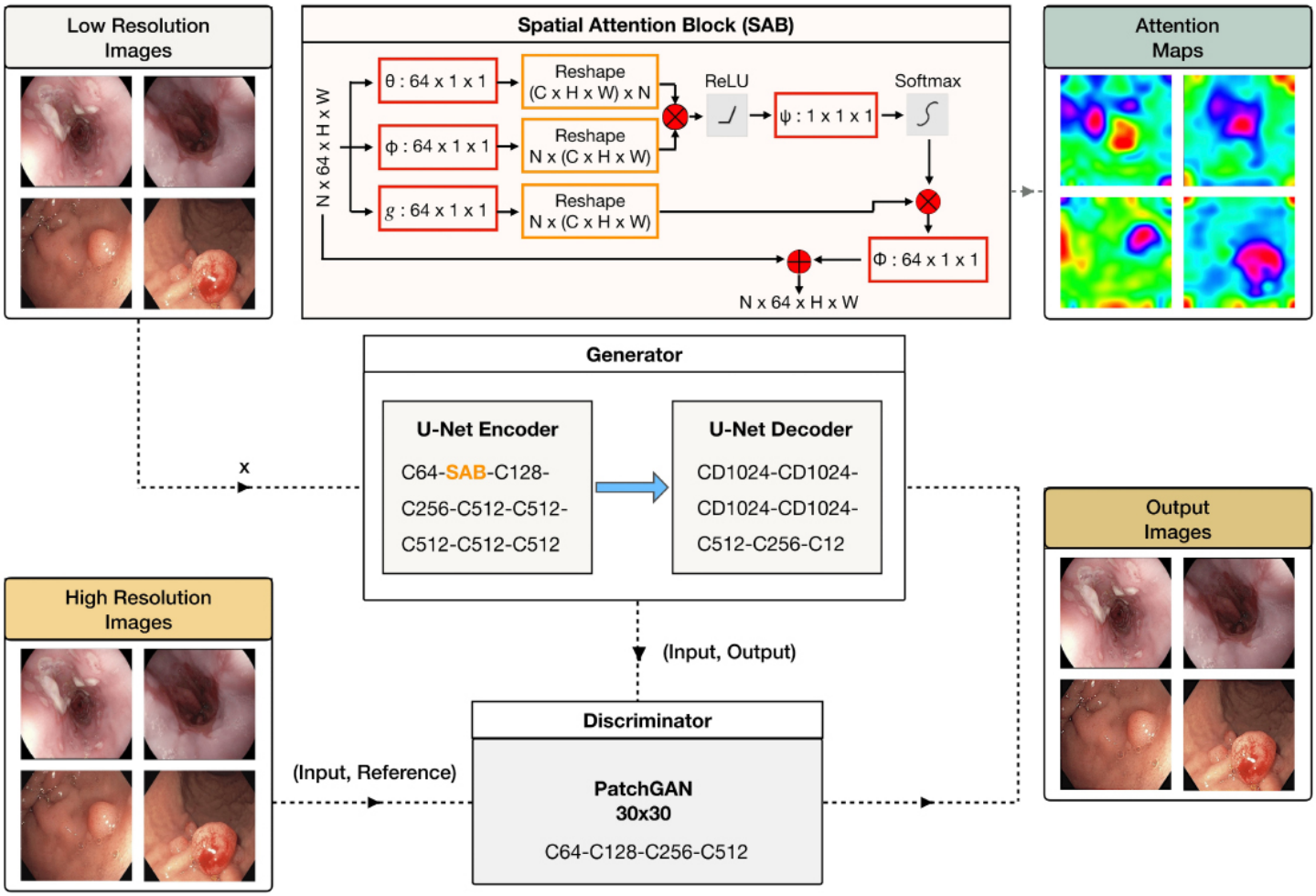

Figure 2. Overall network architecture. A low-resolution image (LR) is used as input to a U-Net based generator with an embedded spatial attention block that produces a corresponding high-resolution image (HR). The discriminator compares the low-resolution image with its high-resolution counterpart and attempts to determine if the generated image is fake or real. In the generator, the input tensor is downsampled and upsampled to the factors of two through the encoder and the decoder layers, respectively. As a discriminator, a convolutional PatchGAN model is employed to penalize the structure based on the patch size of 30x30. The outputs of the spatial attention module are represented as attention heatmaps using the Grad-CAM tool

Finally, the generated 512 feature maps are given to two fully connected layers followed by a sigmoid function to classify the super-resolution images (SR) and determine a probability value. Unlike standard GANs, a discriminator architecture based on PatchGAN is designed to control patch-level structures rather than the whole image. The designed discriminator network determines whether each NxN patch in the image is fake or real. We run the discriminator network over each patch and average all the results

to get the final result. In the patch-based design, reduced parameter sizes also enable a fast evaluation of the discriminator network. Furthermore, the patch discriminator represents the images as a Markov random field assuming that the model separates the pixels by larger than the patch size which is chosen to be 30x30 [17]. After the convolution process, the sigmoid function is applied in the last layer of the discriminator to obtain a one-dimensional output.

Learning objectives and hybrid loss function

In the proposed method, a hybrid loss function is designed specifically for super-resolution applications in capsule endoscopy by combining content loss, pixel loss, texture loss and adversarial loss to optimally train the network. With deeper insight, the pixel loss function (L_{pixel}) measures the deviation between two images at the pixel level. We employ a pixel L1 loss version, known as Charbonnier loss [18] to enforce that the generated SR image (G^{SR}) is close to the HR image (G^{HR}) on a pixel basis, which is mathematically the reference image. Since the PSNR metric is strongly associated with the pixel-difference of the compared images, such that minimizing the pixel-difference maximizes the value of the PSNR, pixel loss is frequently utilized in SR and related disciplines. However, the disadvantage of pixel loss is that the high frequency information and perceptual image quality cannot be preserved for very smooth textures [19]. To overcome this problem and evaluate the semantic variation between the SR and HR images, the content loss function ($L_{content}$) is preferred [20]. For this purpose, the content loss utilizes activation maps acquired from pre-trained networks for feature extraction and transfers HR image features to SR regions to ensure consistency in the content of the LR and SR images. Since unavailable or mislabeled information generated by SR can lead to misdiagnosis, checking consistency in content is critical. In order to generate more realistic and visually acceptable SR images, texture loss ($L_{texture}$) is integrated into the hybrid loss function by modeling the texture of the image as associations within various types of feature maps of a layer and mathematically representing it as the Gram matrix $Gr^{(l)} \epsilon R^{clxcl}$ [21]. Moreover, unlike standard traditional GANs that use the cross entropy based GAN loss, the least square error based adversarial loss (L_{adv}) is chosen to achieve faster convergence and greater stability during training, as shown in [22]. Based on these, pixel, content, texture and adversarial loss functions are summed by assigning different weights, whereupon the hybrid loss function (L_{obj}) is created and formulated as follows:

$$L_{obj}=\alpha L_{adv}+(1-\alpha)(1-\beta)(1-\gamma).L_{pixel}+\gamma.L_{content}+\beta.L_{texture} \quad (6)$$

The parameters α, β, γ can be used to modulate the effects of the loss functions L_{adv}, $L_{texture}$ and $L_{content}$ on L_{obj}, respectively.

Dataset and evaluation metrics

The analysis of the proposed method was performed using the Kvasir dataset, which consists of approximately 80,000 endoscopy images of gastrointestinal organs such as the small intestine, stomach, esophagus, and duodenum taken with various endoscopy devices [23]. A subset of 4,000 images at 1280 x 1024 resolution, including esophagitis, polyps, and ulcerative colitis classes, was selected for the training and testing phase. For quantitative evaluation, two metrics were used. 1) Peak signal-to-noise ratio (PSNR) [24] measures the quality of the generated image by evaluating the pixel-based similarity between the reference image and the SR output image. It is expressed as a logarithmic function of the ratio between the maximum possible pixel value and the mean square error (MSE) between G^{SR} and G^{HR}. The MSE is measured by averaging the squares of the pixel-wise distances of the estimated, and reference image. An increase in the error indicates a lower similarity between two images. 2) The structural similarity index (SSIM) [25] assesses the similarity between two images by considering the luminance, contrast and structural combination of pixels. It attempts to measure the quality perceived by Human Visual System (HVS), in values ranging from 0 to 1, and gives better indication compared to PSNR even for high frequency information such as textures and edges. The proposed technique was compared to DBPN [26] and RCAN [27] methods using these metrics.

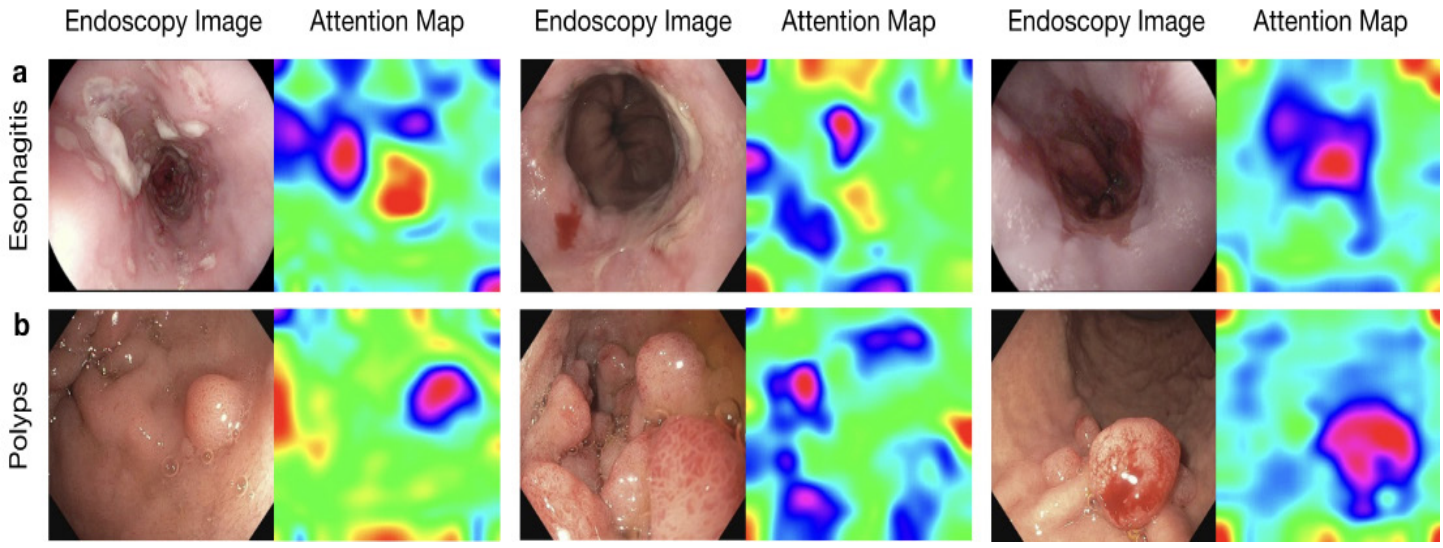

Figure 3. Endoscopy images and the corresponding attention maps. Each pair consists of a low-resolution input image and its corresponding output as attention maps derived from the spatial attention module. To visualize SAB activations, the Grad-CAM approach was used. It can be noticed that the attention unit encourages the model to pay attention to clinically more distinctive areas that are topologically and texturally distinct from surrounding regions

Results

The performance evaluations of the compared methods, as well as the efficiency of the spatial attention block (SAB) in the context of the super-resolution (SR) domain were analysed. The dataset was partitioned into a training set, a validation set and a test set, containing 3000, 500 and 500 images, respectively. Then, the models were set up with a total of 10^5 iterations and a learning rate of 10^{-4}, and trained for 200 epochs using Pytorch library with Adam optimization. Finally, proposed models were trained and tested on NVIDIA® GTX 3090 graphics cards.

Ablation study for SAB

The spatial-attention block is placed after the first convolutional layer of the generator to allow the network focusing on more clinically significant or relevant regions. Such an attentional mechanism attempts to utilize dependencies between contextual information and functional channels outside of local image regions, retaining low-level as well as high-level information from the LR input space. The efficacy of the SAB module is illustrated in Figure 3a-b using examples of polyps and esophagitis images from the Kvasir dataset with the corresponding attention heatmaps generated from the SAB layer results and visualized using the GradCAM approach. As can be seen, the attention module focuses more carefully on areas that are texturally and topologically different from their surroundings and places more emphasis on the SR quality of these areas.

Performance evaluation in clinical anomalies

Since maintaining clinically abnormal areas and enhancing perceptual quality are crucial in terms of medical relevance, the performance of the SR algorithm on tiny artifacts and abnormalities is very important. Therefore, the ability of the compared methods to preserve clinically significant small regions without degradation in the SR process was analyzed on three different classes of image sets: esophagitis, ulcerative colitis and polyps. In Figure 4, the reference images (HR) and the corresponding super-resolved results (SR) are shown for each method, including areas with pathological findings. As can be seen, DBPN and RCAN produce blurring and oversmoothing artifacts whereas the proposed method retains texture and pathological details more faithfully to the reference HR image. Overall, the proposed method was shown to be superior at representing sharp edges and small details in comparison to other methods. Moreover, the compared methods were trained and tested for large scaling factors (8× 10× and 12×) and mean values as well as standard deviations of the evaluation metrics obtained from each approach are shown in Table 1. In all cases, the proposed method yields higher mean values for PSNR and SSIM, which supports the qualitative assessments and approves the superiority of the proposed approach. As a further investigation, a statistical significance evaluation was implemented on 8x upscaling results to ensure that the proposed method produces more successful super-resolved outputs than DBPN and RCAN with respect to the PSNR and SSIM metrics. The p-values calculated from the proposed method and DBPN are 4.12e−1 and 3.42e−25 while p-values from the proposed method and RCAN are 3.48e−26 and 1.87e−85 for the PSNR and SSIM metrics, respectively. Since all p-values are less than or equal to 0.05, the results in both metrics disprove the null hypothesis as well as statistically demonstrate the success of the proposed method over the other two methods.

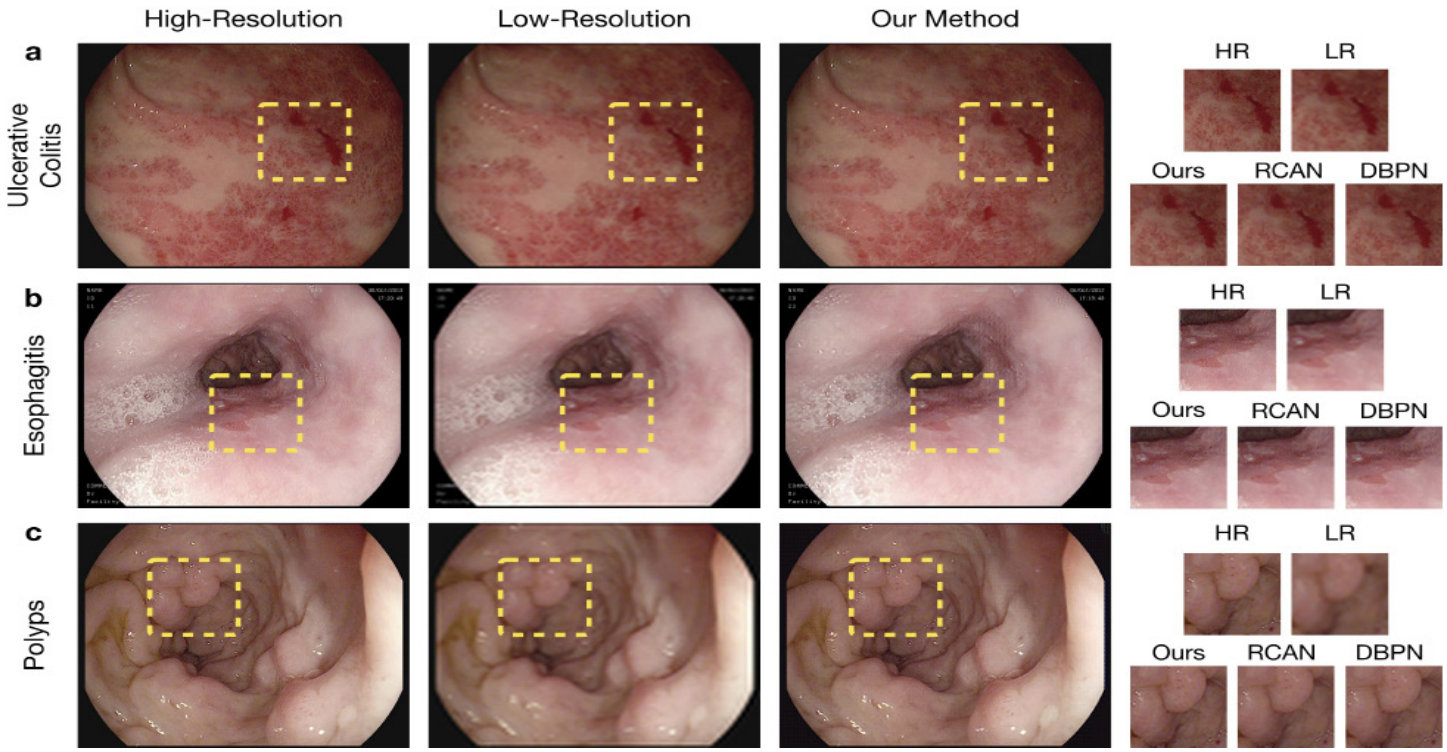

Figure 4. Qualitative evaluation of super-resolution results. Outputs on an 8x upscale factor of each method are demonstrated and cropped in yellow boxes to indicate their performances for three disease classes in the dataset. In general, our method produces super-resolution images with sharper edges, finer texture features, and lower blur compared to DBPN and RCAN. Moreover, our approach successfully maintains and amplifies clinically significant regions more faithfully to the reference high-resolution images, which is of utmost importance for a more reliable and accurate diagnosis

Table 1. Super-resolution efficiency of the proposed method is compared with other methods in terms of PSNR and SSIM. As the quantitative results show, our approach exhibits pixel-wise similarity and structural composition accuracy, with the highest PSNR and SSIM scores at all scales. With the texture loss function, our method is able to capture texture information while with content loss the model generates perceptually similar high-resolution images, leading to more convincing results

Metric	Scale Factor	Our Method	DBPN	RCAN
PSNR↑	8x	34.28±0.48	31.52±0.74	31.46±0.26
	10x	32.57±0.46	30.24±0.69	30.16±0.46
	12x	31.26±0.79	29.78±0.28	28.98±0.27
SSIM↑	8x	0.86±0.12	0.82±0.19	0.80±0.33
	10x	0.83±0.19	0.79±0.16	0.76±0.14
	12x	0.81±0.23	0.76±0.32	0.73±0.12

Discussion

In this study, an attention based conditional GAN network was specially developed for the super-resolution process in capsule endoscopy images. To preserve the clarity of images up to 10-12x scaling factors without adding blur, a hybrid loss function was created by combining pixel, content, texture and adversarial loss functions. An extensive qualitative analysis including investigation of performance in clinical anomalies on super-resolved images as well as visualisation and interpretation of attention heatmaps acquired from the spatial-attention module has proven the success of the proposed method. Its superiority over DBPN and RCAN methods has also been demonstrated by evaluations using PSNR and SSIM metrics. Future studies will involve analyzing the efficacy and usefulness of the proposed model in different applications of capsule endoscopy such as disease detection, segmentation, and classification.

Conflict of interests
The authors declare that they have no competing interests.

Financial Disclosure
This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) with grant 2232 - The International Fellowship for Outstanding Researchers.

Ethical approval
Ethics committee approval was not obtained for this study. As no human and experimental studies were conducted in this study.

References

1. Daniel G, Shai B, Michal I. Super-resolution from a single image. IEEE International Conference on Computer Vision. 2009 12:349–56.
2. Shuhang G, Wangmeng Z, Qi X, et al. Convolutional sparse coding for image super-resolution. ICCV 2015 1:1823–31.
3. Yu-Wing T, Wai-Shun T, Chi-Keung T. Perceptually-inspired and edge-directed color image super-resolution CVPR'06. 2006 2:1948–55.
4. Chao D, Chen C, Xiaoou T. Accelerating the super-resolution convolutional neural network. LNCS. 2016 9906:391–07.
5. Häfner M, M Liedlgruber, A Uhl. POCS-based super-resolution for HD endoscopy video frames. CBMS. 2013 185–90.
6. El-Matary W. Wireless capsule endoscopy: Indications, limitations, and future challenges. J Pediatr Gastroenterol Nutr. 2008;46:4–12.
7. Hirasawa T, Kazuharu A, Tetsuya T, et al. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. Gastric Cancer. 2018 21 653–60.
8. Lee J, Young J, Yoon W, et al. Spotting malignancies from gastric endoscopic images using deep learning. Surg. Endosc. 2019 33:3790–97.
9. Turan M, Almalioglu Y, Araujo H, et al. Deep EndoVO: A recurrent convolutional neural network (RCNN) based visual odometry approach for endoscopic capsule robots. Neurocomputing. 2018;275:1861–70.
10. Isaac JS, Kulkarni R. Super resolution techniques for medical image processing. ICTSD 2015.
11. YLi, B Sixou, F Peyrin. Review of the deep learning methods for medical images super resolution problem. IRBM. 2020;42:120–33.
12. Yasin A, Kutsev B , Abdulkadir G, et al. EndoL2H: Deep Super-Resolution for Capsule Endoscopy. IEEE Transactions on Medical Imaging. 2020;39:4297-309.
13. Wang Z, Chen J, Hoi S. Deep learning for image super-resolution: A survey. IEEE TPAMI. 2019;2:1-7.
14. Isola P, Zhu J, Zhou T, et al. Image-to-image translation with conditional adversarial networks. CVPR. 2017;5967–76.
15. Zhou Z, Rahman S, Tajbakhsh N, et al. Unet++: A nested u-net architecture for medical image segmentation. Lect. Notes Comput Sci. 2018;11045:3–11.
16. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. ICML. 2015;1:448–56.
17. Li C, Wand M. Precomputed real-time texture synthesis with markovian generative adversarial networks. Lect. Notes Comput Sci. 2016;9907:702–16.
18. Lai W, Bin H, Ahuja N, et al. Fast and Accurate image super-resolution with deep laplacian pyramid networks. IEEE Trans Pattern Anal Mach Intell 2019;41:2599–13.
19. Mertens K, Baets B, Verbeke L, et al. A sub-pixel mapping algorithm based on sub-pixel/pixel spatial attraction models. Int J Remote Sens. 2006;27:3293–10.
20. Zhang W, Liu Y, Dong C, et al. RankSRGAN: Generative adversarial networks with ranker for image super-resolution. Proc. IEEE Int Conf Comput Vis. 2019;3096–05.
21. Sch B, Hirsch M. EnhanceNet: single image super-resolution through automated texture synthesis. ICCV. 2019;4491–00.
22. Xintao W, Ke Y, Shixiang W, et al. ESRGAN: Enhanced super-resolution generative adversarial networks. ECCV. 2019;11133:63–79.
23. Pogorelov K, Kristin R, Carsten G, et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. ACM Multimedia System. 2017;164–69.
24. Tong T, G Li, X Liu, et al. Image super-resolution using dense skip connections. ICCV. 2017;4809–17.
25. Wang Z, Simoncelli E, Bovik A. Multi-scale structural similarity for image quality assessment. Signals, Systems and Computers. 2003;2:1398–02.
26. Haris M, Shakhnarovich G, Ukita N. Deep Back-projection networks for super-resolution. CVPR. 2018;2:1664–73.
27. Zhang Y, Li K, Wang L, et al. Image super-resolution using very deep residual channel attention networks. ECCV. 2018;2:294-10.