# Surgical Video Image Restoration Method based on Deep Learning

**Jiang-Po Guo**[1], **Ke-Yu He**[1], **Ying Ma**[1], **Yuan-Yan Ye**[1], **Yue Zhang**[1]
[1]Department of Computer Science, Xiamen University, China

## Abstract

Medical electronic endoscope is one of the direct and effective medical devices for medical personnel to observe the internal pathological tissue of human body, which is known as ”the third eye of human beings”. In the process of endoscope image acquisition, illumination is artificially added through external equipment. Therefore, under the influence of lighting conditions, the initial image acquisition of endoscope will show the effect of bright light sufficient area and dim light insufficient area, which will lead to the degradation of endoscope image quality and affect the doctor’s diagnosis to a certain extent. Based on the deep neural network, Retinex theory, and the image enhancement model of bilateral grid, we realize the real-time recovery of surgical video under insufficient light. (1) The most advanced image enhancement model was applied to establish a new medical image data set through gamma correction and manual processing, and minor changes were made to the model to improve its performance in medical image problems. (2) A new error function is generated by referring to the linear combination of mean absolute error and structural similarity error, and the weighted least square filter is used to enhance the image contrast and improve the image quality. By comparing multiple models on the data set established by us, our model has achieved good recovery effect from both subjective analysis and objective evaluation.

## Introduction

Low-illuminance images are images taken under low light. Such images often have problems such as low brightness and poor contrast. Therefore, the research of low-light image enhancement has strong practical significance. This article mainly focuses on the two aspects of low-light image and medical image.

The photos taken by the camera may be underexposed due to low light and backlight. This kind of image cannot capture what the user wants, because underexposed areas can hardly see details, low contrast, and dim colors. However, low-light image enhancement is a challenging task, because underexposed areas are usually hard to detect, and the enhancement process is highly nonlinear and has subjective factors.

There are many methods to solve this problem. Early research mainly focused on contrast enhancement. This method is not effective in restoring image details and colors. Recent research uses data-driven methods to learn color, contrast, and adjust the brightness to produce more expressive results.

Different from low-light images, medical imaging refers to a non-invasive image of internal tissues of the human body or a certain part of the human body for medical or medical research. At present, minimally invasive surgery uses endoscopic equipment and technology. Compared with traditional surgery, minimally invasive surgery has a wider field of view. Because of this, it has strict requirements on the level of vision of endoscopic technology. Nowadays, the widely used 3D endoscope provides the color information of the endoscopic image by illumination, and the image acquisition process of the illumination endoscopy technology is artificially added by external equipment (Chen et al. 2016). Therefore, affected by the lighting conditions, the initial image acquired by the endoscope will appear bright in the middle and dark around, resulting in degradation of the quality of the image. Therefore, we need to restore the original image collected by the endoscope to enhance the quality and improve the recognizability of image details.

## Related Work

### Medical image quality enhancement

For thousands of years, due to the structure of the human body, we cannot directly see the inside of the human body. How to explore the inside of the human body has been a difficult problem. In 1804, Philip Bozzini, the German doctor who first proposed the idea of an endoscope, was hailed as the ”first inventor of an endoscope”, and in 1806 he produced a candle that uses a candle as a light source and mirrors the light. The reflex function is used to observe the equipment inside the bladder and rectum-”lighting device” (Cunningham and Peterson 2003). Over the next 100 years, scholars made improvements to light sources, lighting methods, and the invention of light bulbs, which led to a significant development in endoscopy technology. In 1983, the Welcn Allyn Company in the United States developed an electronic endoscope. The birth of an electronic endoscope was another historic breakthrough in the history of endoscope development. The quality of endoscopic medical images has been greatly improved. In 2006, Olympus released an endoscope system with integrated technology. In 2014, Stryker launched the world's first medical lens camera. The image capture speed has increased by 33% compared to the previous year, and the signal-to-noise ratio, brightness and clarity have all made improvements.

Over the years, although the level of endoscopy technology has been rapidly improved, the lighting method has always been a difficult obstacle to the development of endoscopes. Therefore, some scientists have thought of starting with image processing to indirectly solve the problem caused by the lighting method image problem. At the same time, various medical companies are constantly developing image enhancement technologies. Including technologies such as Flexible Spectral Image Color Enhancement from Fujinon and Storz Professional Image Enhancement System from Karl Storz.

The basic principle of FICE technology (Togashi et al. 2009) is based on spectrum estimation technology. FICE technology uses electronic spectroscopy technology to decompose and simplify the different color number elements collected by the color image sensor, and provides an image processing mode that generates any combination of wavelengths, which significantly improves the contrast between the lesion and the surrounding tissue structure, and more effectively improves detection rate of lesions.

The SPIES system (Kamphuis et al. 2016) provides four image enhancement methods: Chroma, Spectra A, Spectra B, and Clara Chroma mode can sharpen the image and make the image clearer. Spectra A uses a dedicated color conversion algorithm to enhance the contrast between fine blood vessels and blood vessels on the membrane surface. Spectra B retains the blood vessels in the deep layer of the milk film on the basis of Spectra A, making the image more detailed. Clara mode can enhance the brightness of dark areas of the image.

### Low illumination image quality enhancement

Low-illuminance image enhancement is one of the popular research directions in the field of computer vision. It mainly deals with the problems of noise, low brightness, and low contrast in images with insufficient lighting and uneven illumination, so as to improve visual quality.

Low-light video enhancement technology is mainly to enhance each video frame in the video, so its core is still low-light image processing technology. Traditional low-light image enhancement algorithms mainly include the following:

- The image enhancement algorithm based on histogram equalization redistributes the pixel values of the image by stretching the dynamic range of the gray scale, so that the number of pixels in a certain gray scale range is almost the same.
- The image enhancement algorithm based on wavelet transform decomposes the image into the low-frequency sub-band of the image approximate signal and the high-frequency sub-band of the image detail signal. Non-linear image enhancement of low frequency subbands can enhance contrast and suppress background. Denoising processing on high frequency subbands can effectively reduce the influence of noise. However, these methods are very difficult to reconstruct high-frequency components and low-frequency components, and the amount of calculation is huge.
- The Retinex theory was proposed by Edwin.H.Land et al. in 1963. It is based on the idea that the color of the image of an object is influenced by the reflection properties of the object's surface and the environmental lighting. It believes that the essence of the color of the object's imaging is not the environmental lighting, but the reflection of the incident light on the target scene itself.

In recent years, with the rapid development of artificial intelligence technology, deep learning has achieved far better results than previous technologies in the recognition of text and speech, as well as image recognition and processing. At present, the use of deep learning for low-illumination image quality enhancement research has also made some progress.

In terms of low-light image enhancement, K.G. Lore et al. proposed a deep autoencoder method (Lore, Akintayo, and Sarkar 2017) for low-light image enhancement, which can extract image features, enhance image brightness, and remove image noise. Shen Liang et al. proved that the traditional MSR (Multi Scale Retinex) method can be regarded as a feedforward convolutional neural network with different Gaussian convolution kernels, and proposed a MSR-Net (Shen et al. 2017) that can directly learn the end-to-end mapping from dark images to bright images.

Inspired by the Retinex theory, Chen Wei et al. proposed RetinexNet (Wei et al. 2018). Its overall structure mainly includes two networks: DecomNet and RelightNet. DecomNet is used to decompose the picture into reflection components and illumination components, and RelightNet is used to correct the illumination components and reconstruct them with the reflection components to obtain the corrected image. At the same time, they put forward the idea of collecting data in real scenes for the first time, which not only enhances the brightness of the image, but does not destroy the image texture details and boundary information.

Wang Ruixing et al. proposed a new end-to-end network for enhancing underexposed photos (Wang et al. 2019). They introduced intermediate lighting in the network, and correlated the input with the expected enhancement result, thereby using the simple nature of the lighting in natural images to enhance the ability of network to learn complex photographic adjustments from modified images. Through these means, their network can restore clear details, sharp contrast and vivid colors in underexposed photos.

## Methodology

### Data augmentation

Experiments need to imput the impaired images of well-exposed surgical images and corresponding underexposed images. Image enhancement based on natural image generally includes underexposed images and expert modified images for training. Because of the small number of medical image data sets and the lack of corresponding images modified by experts, we propose an under-exposure image pro-

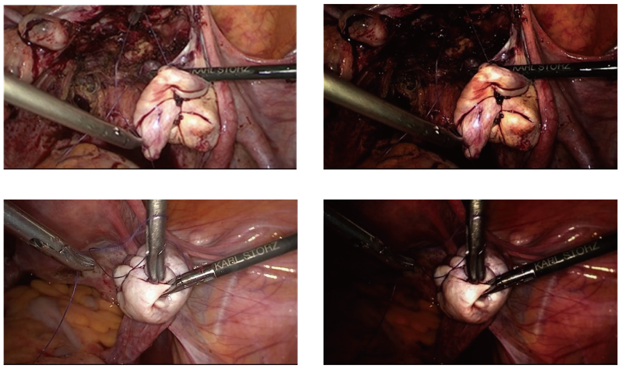

Figure 1: Normal light image and low light image after gamma transform processing

cessing method based on the characteristics of endoscope images to establish image pairs.

**Gamma Correction** The human eye's perception of brightness change is nonlinear and approximate to power function, and more sensitive to the details of the darker region.Gamma Correction was originally used to correct the storage and display of images, using the nonlinearity of human eye cognition color and brightness, so that the images can retain more details of the lower brightness regions and ignore the details of some higher brightness regions. In order to produce the image which is more consistent with the image obtained by endoscope with the characteristics of light in the middle and dark in the sides, Gamma Correction is used to correct different parameter values, The picture is nonlinearly adjusted to darker, while ensuring that the dimmed image is closer to the real endoscope observed by the doctor.

$$L_{out} = L_{in}^{\gamma} \qquad 1$$

## Network structure

First, the image is down-sampled to $256 \times 256$ resolution, and extracted the local and global features on low-resolution images, After feature fusion, the bilateral network is generated by point-by-point convolution, and then the color affine transformation matrix is obtained with the same resolution as the original input image by using the guidance map interpolation upsampling. For each pixel on the original input image, the affine transformation matrix is applied to obtain the predicted input brightness image *S*, and the modified image is obtained by combining the original input low light image.

**Low-level features** the input image is uniformly scaled to $256 \times 256$ resolution by bilinear interpolation, and then by

$$S_c^i[x,y] = \sigma\left(b_c^i + \Sigma_{x'}, y'_{i,z} w_{cc'}^i \left[x', y'\right] S_{c'}^{i-1}\left[sx + x', sy + y'\right]\right) \qquad 2$$

extracting low-level features while further reducing the resolution of the data space (filter $3 \times 3$, stride2), c and c' are the RGB channels for each convolutional layer. After $\mathrm{n_s}$ layer convolution calculation, the extracted feature map size is further reduced by $2^{n_s}$ times. $\mathrm{n_s}$ control the degree of down-sampling from the low-resolution input image to the bilateral grid, and also affect the complexity of the model. The bigger $\mathrm{n_s}$ is, the richer the extracted information.

**Local features and Global features** The fusion of local features and global features is very important for extracting sufficient semantic information. The extraction of local features can provide the spatial position information of the input image. On the basis of the extracted low-dimensional features, the convolution kernel of different sizes (stride=1) is used in each layer to extract the semantic information of different sizes. Ensure the spatial size of the output feature map unchanged, to accomplish local feature extraction. The global feature consists of x layer convolution layer and y layer full connection layer. Let the information contained in the feature map is integrated into a vector with equal number of channels and local feature channels, which facilitates the next feature fusion and provides the prior information of the whole input image.

**feature fusion** After splicing the local and global features along the depth dimension, the features are fused by point convolution to generate a new feature map, and the spatial size of the feature map is adjusted, and then activated by the ReLU function

$$F_c[x,y] = \sigma\left(b_c + \Sigma_{c'} w'_{cc}, G_{c'}^{n_G} + \sum_{c'} w_{cc'} L_{c'}^{n_L}[x,y]\right) \qquad 3$$

Outputting the faature matrix of $16 \times 16 \times 64$, and then the coefficients of the pixel value transformation affine function are predicted by a linear regression prediction model of $1 \times 1 \times 64$, and the expanded coefficient matrix of $16 \times 16 \times 96$ is obtained.

$$A_c[x,y] = b_c + \Sigma_{c'} F_{c'}[x,y] w_{cc} \qquad 4$$

**Bilateral grid** Compress the coefficient matrix along the third dimension,$A_{dc+z}[x,y] \leftrightarrow A_c[x,y,z]$, Bilateral grid depth d=8, bilateral grid size $16 \times 16 \times 8$, Each grid contains 12 parameters, that is, the parameters of the 3/4 affine color transformation matrix. From the change of feature map size, combined with the extraction process of low-level features, it can be seen that the stepwise convolution on the x,y dimension combines the information on the c and z dimensions, and enhances the communication between the data.

**Slicing** After obtaining the bilateral grid A, we need to do the upsampling operation to obtain the output result with the same resolution as the original input image. A low-resolution bilateral grid A and a gray image with the same resolution as the original input image guidance map (He, Sun, and Tang 2010) g are used as inputs. The slicing operation of the proposed guidance map based on Paris (Paris and Durand 2009) and so on obtains the new feature map $\tilde{A}$ with the same spatial domain resolution as the same spatial domain resolution. This operation is differentiable for both A and g, i.e. A and g can be adjusted by error backpropagation algorithm.

$$\tilde{A}_c[x,y] = \sum_{i,j,k} \tau\left(s_x x - i\right) \tau\left(s_y y - j\right) \tau(d \cdot g[x,y] - k) A_c[i,j,k] \qquad 5$$

$\tau(\cdot) = \max(1 - |\cdot|, 0)$ the linear interpolation kernel function,$s_x, s_y$is the sampling rate on the wide and high dimensions of the grid and the original resolution input image; The x,y is the spatial coordinate of a point on the g, which is determined by the gray value of the store on the g by the

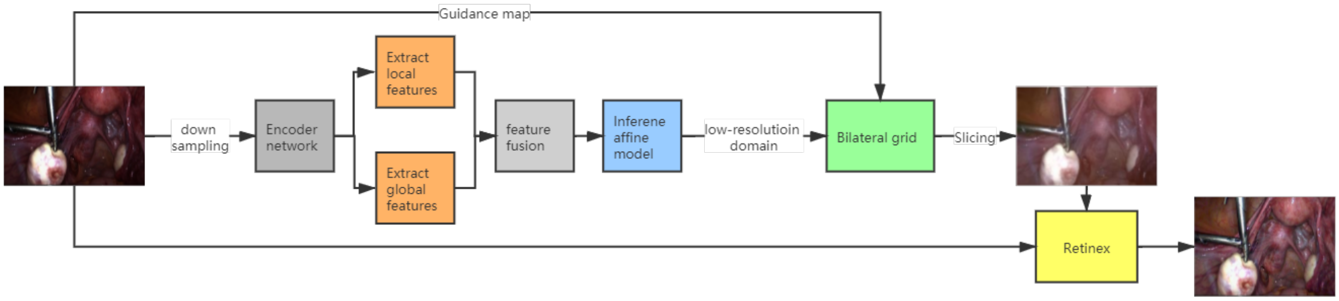

Figure 2: network structure

third dimension coordinate of the projection point of the low resolution bilateral grid, that is, the frequency domain coordinate.

## Loss function

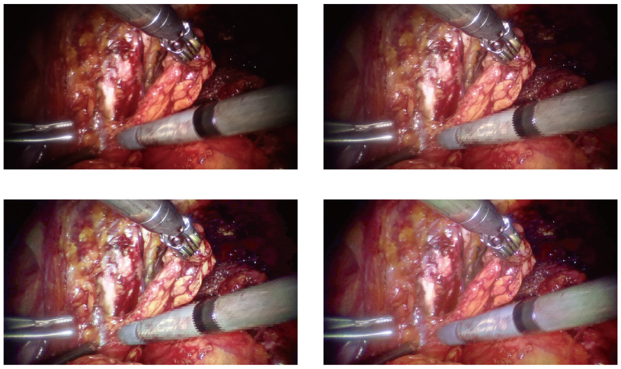

Figure 3: The result of image enhancement of the endoscopic image. The upper left corner is the original image under the endoscope. The loss function of the upper right image is $\mathcal{L}_m$, the loss function of the lower left image is $\mathcal{L}_m$+2$L_s$, and the loss function of the lower right image is $\mathcal{L}_{MAE}$+$\mathcal{L}_{SSIM}$.

In order to measure errors in many aspects, a linear combination of multiple loss functions is used as the loss function, and different weights are set according to experience. It is expressed as

$$\mathcal{L} = w_m \mathcal{L}_m + w_t \mathcal{L}_t + w_s \mathcal{L}_s \tag{6}$$

**Mean square error** We use the mean square error as the loss function. As shown in Equation 7, where N is the number of pixels in the image, I is ground true, S is the brightness map of the input picture predicted by the model, I is the input underexposed picture, and the pixel values of the picture are all normalized to [0,1]. The restricted conditions ensure that the enhanced image will not be larger than the boundary value 1, nor will it be darker than the original image. The mean square error fully takes into account the difference in the number of pixel points of the picture, but it is easy to overfit the model and cannot allow the model to learn some deep features.

$$\mathcal{L}_m = \frac{1}{N} \sum_i \left\| S_i^{-1} * I_i - \check{I}_i \right\|^2 \tag{7}$$

$$s.t. \quad (I_i)_c \leq (S_i)_c \leq 1, \quad \forall channelc \tag{8}$$

**Smoothing loss** Zccv Harbman et al. (Farbman et al. 2008) proposed a weighted least square to enhance the edge information of the picture. Smoothing loss can enhance the generalization ability of the model and at the same time enhance the contrast of the output image. In order to make the output image u retain the edge information of the input image g as much as possible, the original problem is regarded as a filter model that minimizes the value of the following formula:

$$\Sigma_p \left( \left( u_p - g_p \right)^2 + \lambda \left( a_{x,p(g)} \left( \frac{\partial u}{\partial x} \right)^2 + a_{y,p(g)} \left( \frac{\partial u}{\partial y} \right)^2 \right) \right) \tag{9}$$

Among them, $(u_p - g_p)$ is used to measure the similarity between u and g, and $\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}$ is the gradient of image u at points x and y is used to measure the degree of smoothness. Therefore, the loss function to measure the smoothness of a point on the image is:

$$L_s = \sum_c \sum_p w_{x,p(I)} \left( \frac{\partial I}{\partial x} \right)^2 + w_{y,p(I)} \left( \frac{\partial I}{\partial y} \right)^2 \tag{10}$$

**Structural loss** Zhou Wang et al. (Wang et al. 2004) proposed a structural similarity image quality enhancement index, considering the similarity between two pictures from the three aspects of brightness L, contrast C, and structure S. SSIM is defined as

$$SSIM(x,y) = \left[ L(x,y)^{\alpha} \cdot C(x,y)^{\beta} \cdot S(x,y)^{\gamma} \right] \tag{11}$$

When the loss function is SSIM and mean absolute error, the performance of the model is better than using one of them alone. Therefore, the structure loss function is defined as shown in Equation 14, $\mathcal{L}_{MAE}$ is the average absolute error between the two images.

$$\mathcal{L}_{SSIM}(I) = 1 - SSIM(I, \check{I}) \tag{12}$$

$$\mathcal{L}_{MAE} = \frac{1}{N} \sum_i \left| S_i^{-1} * I_i - \check{I}_i \right| \tag{13}$$

$$\mathcal{L}_t = \alpha \mathcal{L}_{SSIM} + (1 - \alpha) \mathcal{L}_{MAE} \tag{14}$$

The human eye's perception of brightness change is nonlinear and approximate to power function, and is more sensitive to the details of the darker region. Gamma Correction was originally used to correct the storage and display of images, using the nonlinearity of human eye cognition color and brightness, so that the images can retain more details of the lower brightness regions and ignore the details of some higher brightness regions. In order to produce the image which is more consistent with the image obtained by endoscope with the characteristics of middle bright and dark, Gamma Correction is used to correct different parameter values. The picture is nonlinearly adjusted to darker, while ensuring that the dimmed image is closer to the real endoscope observed by the doctor.

## Experiment

### Evaluation index

In the image enhancement problem, PSNR(Peak Signal to Noise Ratio) and SSIM(Structural Similarity) are generally used to objectively evaluate the performance of the model. PSNR is based on the error between corresponding pixels. The larger the value, the smaller the distortion. However, it does not take into account that the visual characteristics of the human eye may be inconsistent with human subjective perception. The larger the value of SSIM, the more similar the two images in brightness, contrast, and structure. Because the endoscopic image has the characteristics of bright areas with sufficient light and dim areas with insufficient light. After image enhancement, in the subjective visual perception of the human eye, the overall brightness of the medical image is improved, the texture details are more obvious, and the contrast is greater.

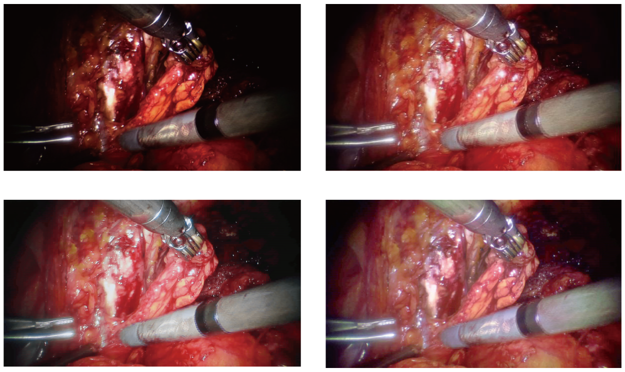

Figure 4: The results of four different models processing the medical images of the test set. The upper left corner is the result of FEQE, the lower left corner is the result of DPED, the upper right corner is the result of RetinexNet, and the lower right corner is the result of our model.

### Comparison

The images enhanced by our model is compared with the images obtained by the other three image enhancement algorithms FEQE (Vu et al. 2018), RetinexNet (Wei et al. 2018) and DPED (Ignatov et al. 2017). FEQE first downsamples low-quality images, samples the features into a low-resolution space, and then inputs the features into a series of N residual blocks for instance normalization and ReLU activation, and finally up-samples the output image. The entire model of RetinexNet can be divided into three parts: decomposition model, adjustment model and reconstruction model. The decomposition model decomposes an image into illumination components and reflection components; the adjustment model suppresses the noise of the reflection component of the low-light image, and enhances the illumination component of the low-light image; the reconstruction model restores the processed reflection component and illumination component to a normal illumination image. DPED is divided into three steps. First, use the generation and decomposition of two networks to learn multiple loss functions. Second, use the combined perception error function to combine multiple losses to improve the quality of the picture. Finally, use the effective method of calibrating the image to make the picture perform Self-learning to further improve the quality of the picture. The results obtained are shown in Figure 4. Calculate the average PSNR value and average SSIM value of the test set images respectively, and get Table 1. It can be seen that compared with other models, the expressiveness of our model has been significantly improved.

| | PSNR | SSIM |
|---|---|---|
| Our model | 24.5800 | 0.8862 |
| DPED | 23.4647 | 0.8851 |
| RetinexNet | 20.3577 | 0.5260 |
| FEQE | 12.5424 | 0.3725 |

Table 1. The average PSNR value and average SSIM value of the results obtained by the four different models

## Conclusion

Based on the deep neural network, Retinex theory, and the image enhancement model of bilateral grid, we realize the real-time recovery of surgical video under insufficient light. (1) The most advanced image enhancement model was applied to establish a new medical image data set through gamma correction and manual processing, and minor changes were made to the model to improve its performance in medical image problems. (2) A new error function is generated by referring to the linear combination of mean absolute error and structural similarity error, and the weighted least square filter is used to enhance the image contrast and improve the image quality. By comparing multiple models on the data set established by us, our model has achieved good recovery effect from both subjective analysis and objective evaluation.

# References

Chen, J.; Adams, A.; Wadhwa, N.; and Hasinoff, S. W. 2016. Bilateral guided upsampling. *Acm Transactions on Graphics* 35(6): 203.

Cunningham, L. L.; and Peterson, G. P. 2003. Historical development of endoscopy. *Atlas of the Oral and Maxillofacial Surgery Clinics of North America* 11(2): 109–127.

Farbman, Z.; Fattal, R.; Lischinski, D.; and Szeliski, R. 2008. Edge-preserving decompositions for multi-scale tone and detail manipulation. *ACM Transactions on Graphics (TOG)* 27(3): 1–10.

He, K.; Sun, J.; and Tang, X. 2010. Guided image filtering. In *European conference on computer vision*, 1–14. Springer.

Ignatov, A.; Kobyshev, N.; Timofte, R.; Vanhoey, K.; and Van Gool, L. 2017. Dslr-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 3277–3285.

Kamphuis, G.; de Bruin, D.; Fallert, J.; Gultekin, M.; De Reijke, T.; Laguna Pes, M.; and de la Rosette, J. 2016. Storz professional image enhancement system: a new technique to improve endoscopic bladder imaging. *Journal of Cancer Science & Therapy* 8(3): 71–77.

Lore, K. G.; Akintayo, A.; and Sarkar, S. 2017. LLNet: A Deep Autoencoder Approach to Natural Low-light Image Enhancement. *Pattern Recognition* 61: 650–662.

Paris, S.; and Durand, F. 2009. A Fast Approximation of the Bilateral Filter Using a Signal Processing Approach. *International Journal of Computer Vision* 81(1): 24–52.

Shen, L.; Yue, Z.; Feng, F.; Chen, Q.; Liu, S.; and Ma, J. 2017. MSR-net:Low-light Image Enhancement Using Deep Convolutional Network .

Togashi, K.; Osawa, H.; Koinuma, K.; Hayashi, Y.; Miyata, T.; Sunada, K.; Nokubi, M.; Horie, H.; and Yamamoto, H. 2009. A comparison of conventional endoscopy, chromoendoscopy, and the optimal-band imaging system for the differentiation of neoplastic and non-neoplastic colonic polyps. *Gastrointestinal endoscopy* 69(3): 734–741.

Vu, T.; Van Nguyen, C.; Pham, T. X.; Luu, T. M.; and Yoo, C. D. 2018. Fast and efficient image quality enhancement via desubpixel convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 0–0.

Wang, R.; Zhang, Q.; Fu, C. W.; Shen, X.; and Jia, J. 2019. Underexposed Photo Enhancement Using Deep Illumination Estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13(4): 600–612.

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560* .