Multi-Scale Structural-aware Exposure Correction for Endoscopic Imaging

Ricardo Espinosa
respinosa@up.edu.mx

Axel Garcia-Vega
caaxgave@gmail.com

Gilberto Ochoa-Ruiz
gilberto.ochoa@tec.mx

Tecnologico de Monterrey, School of Engineering and Sciences, Mexico
Tecnologico, 64849 Monterrey, N.L.

Universidad Panamericana, Facultad de Ingeniería, Aguascalientes, 20290, México.

Abstract

Endoscopy is the most widely imaging technique used for the diagnosis of cancerous lesions in hollow organs. However, endoscopic images are often affected by illumination artefacts: image parts may be over- or underexposed according to the light source pose and the tissue orientation. These artifacts have a strong negative impact on the performance of computer vision or artificial intelligence based diagnosis tools. Although endoscopic image enhancement methods are greatly required, little effort has been devoted to over- and underexposure enhancement in real-time. This contribution presents an extension to the objective function of LMSPEC, a method originally introduced to enhance images from natural scenes. It is used here for the exposure correction in endoscopic imaging and the preservation of structural information. To the best of our knowledge, this contribution is the first one that addresses the enhancement of both types of endoscopic artefacts (under- and over-exposure) using deep learning methods. Tested on the Endo4IE dataset, the proposed implementation has yielded a significant improvement over LMSPEC reaching a SSIM increase of 4.40% and 4.21% for over- and underexposed images, respectively.

1. Introduction

Endoscopy plays a central role in minimally invasive surgery or for carrying out examinations in hollow organs, such as colon or stomach. In recent years, computer aided endoscopy has become an important area of research. In particular, Computer Vision (CV) has the potential of becoming an essential tool for assisting endoscopists in various tasks [1–3].

However, a major hurdle that most of these CV methods must face is related to the uncontrolled and highly changing illumination conditions in endoscopic scenes. Figure 1 shows two colonoscopic images in which strong illumination changes are visible.

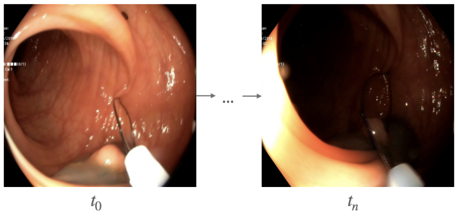

Figure 1. Strong illumination change example in almost consecutive frames of a colonoscopic image sequence

The results obtained by numerous methods in the EAD challenge have shown that image enhancement (IE) algorithms are of high interest for improving the robustness and generalization capabilities of endoscopic image preprocessing techniques. This contribution focuses on the exposure correction in white light endoscopy. It is noticeable that this issue has only been partially addressed in the IE field, as most methods (see [4]) were dedicated to the correction of either over- or underexposed images, but did not deal with both effects occurring concurrently.

García-Vega et al. proposed a paired "normal-exposed" image dataset [5,6] to assess the ability of machine learning (ML) based methods to correct the effects of non-optimal lighting conditions. However, the need for both accurate and real-time IE techniques highlighted the shortcomings of most current methods. Nonetheless, the LMSPEC deep learning (DL) method proposed by Afifi et al. [4] outperformed other models in terms of accuracy and inference time, whilst obtaining a satisfactory enhancement performance. However, in some images, LMPSEC introduced undesired textural and color artifacts, which could lead, for instance, to false diagnoses in endoscopy or errors in automated methods.

This contribution shows how to alleviate the loss of texture and color information during the exposure correction process by introducing a structural similarity-aware extension to the overall loss function of the LMSPEC pipeline. This modification allows to preserve fine-textured details. The results show that is possible to achieve this goal both for over- and underexposures, while maintaining a relatively low inference time.

The rest of the paper is organized as follows. Section 2 gives an overview of the works relating to endoscopic image enhancement. Section 3 discusses the dataset used to perform the experiments presented in this contribution, as well as the metrics used to evaluate the performance of the solution introduced in Sections 4 (DL-model) and 5 (DL-model tuning). Section 6 gives, through a set of ablation studies, quantitative and qualitative comparisons between the different configurations of the proposed DL-model. Finally, Section 7 globally discusses the results and outlines perspectives.

2. State of the art

In the past, IE methods based on different approaches have been proposed, such as methods using histogram equalization [7] or Retinex theory-based models [8]. More recently, frameworks based on ML models have emerged. In these approaches, the IE mapping is learned instead of computed, and DL models have excelled at this task since its inception. Although these DL-methods have shown great promise, they still suffer from various shortcomings. First and foremost, most of the methods in the existing literature can enhance only over- or underexposed images, but cannot simultaneously perform both tasks with efficiency. Thus, [9], [8] and [10] are examples of methods that can be successfully applied in underexposed images, taking advantage of DL networks and the Retinex assumption. Nonetheless, some recent methods have been proposed to address the enhancement of both over- and underexposed images with inference times which are near to real-time.

In a prospective study, Garcia-Vega et al. [5] compared various image enhancement methods, using a recent dataset containing synthetically generated over- and underexposed endoscopic frames which are paired with their non-corrupted ground truth image. The authors assessed the capabilities of different IE methods to enhance the quality of endoscopic images, while maintaining a high degree of fidelity. To do so, the texture quality was quantified in terms of peak signal to noise ratio (PSNR) and structural similarity index measure (SSIM), as well as subjectively graded by human evaluator. The inference times of each model was also measured. In this study, LMSPEC demonstrated an astounding performance for both types of exposure artifacts, while attaining an almost real-time performance. However, the IE-model introduced, on the one hand, some artifacts that removed high frequency content (texture details) from the enhanced images, and led, on the other hand, on other undesired color artifacts.

3. Materials

3.1. Dataset

The dataset used in this contribution is a combination of three different existing datasets (EAD [11], EDD [12] and HyperKVisir [13]) using the procedure described in [5]. The authors used image-to-image translation to take unmodified endoscopic frames and generated frames with over- and underexposure artefacts. For implementing this task, they used CycleGAN architecture [14] since the main issue to tackle was the lack of paired data (CycleGAN is an efficient method for working with unpaired data).

This dataset is composed of three different types of images: i) 2,216 unmodified (acquired) endoscopic frames (without exposure errors) that act as ground truth data, ii) 1,231 synthetically overexposed frames, and iii) 985 synthetically underexposed frames. Every ground truth image in sub-dataset 1 is associated with its either over- or underexposed synthetic version, belonging to sub-datasets 2 and 3, respectively. Both paired sub-datasets (sub-datasets 1+2 and sub-dataset 1+3) were split as follows: 70% , 27% and 3% of the images were used for the train, test, and validation steps, respectively.

3.2. Metrics

Two standard full-reference metrics were used to compare different IE methods: PSNR and SSIM. Both metrics can evaluate globally over the image the quality of the obtained results.

4. Proposed DL Method

Given a poorly exposed input image **I** acquired under white light, the proposed method (depicted in Figure 2) aims to predict an output image **Y** being a version of **I** with no exposure errors. As in LMSPEC, the color and detail errors of **I** are sequentially corrected. Basically, a multi-resolution representation of **I** is given by a pyramidal Laplacian decomposition derived from a pyramidal Gaussian decomposition of ground truth **T**.

The heart of the method is based on the original implementation of LMSPEC, which randomly extracts n small patches $I'_1, ..., I'_n$ from **I** and decomposes each patch into two components: i) a four-level Gaussian Pyramid (GP) and then ii) a four-level Laplacian Pyramid (LP). This LP can be seen as a set of frames with different frequency levels, $LP = \{l_1, l_2, l_3, l_4\}$, where l_1 and l_4 contain the high- and low frequency components, respectively. This LP decomposition is carried out to feed four U-Net-like sub-nets in a cascade configuration with sub-images with different levels of detail. Each sub-net is used to extract relevant features from the image and to carry out a reconstruction of each l_i input in reverse order, as shown in the LMSPEC block in Fig. 2.

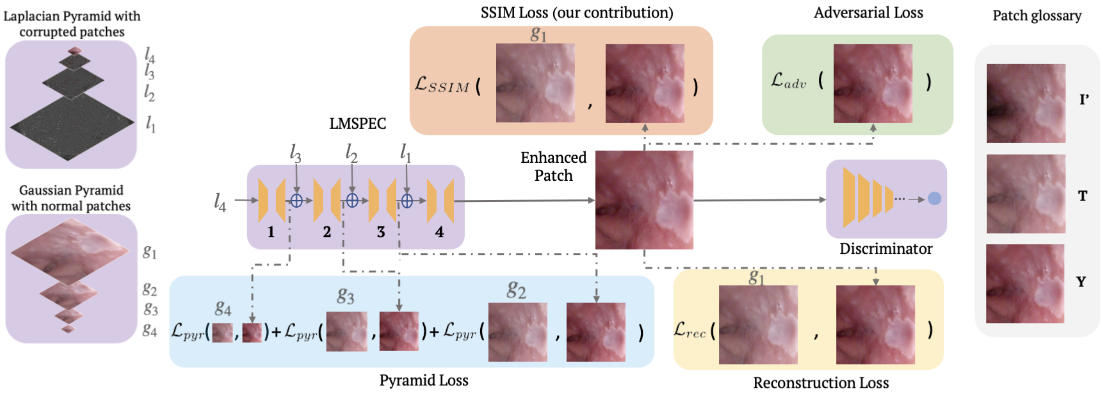
Laplacian Pyramid with corrupted patches
l_4
l_3
l_2
l_1
Gaussian Pyramid with normal patches
g_1
g_2
g_3
g_4
SSIM Loss (our contribution)
$\mathcal{L}_{SSIM}$ (g_1 ,)
Adversarial Loss
$\mathcal{L}_{adv}$ ( )
Patch glossary
I'
T
Y
LMSPEC
l_4
l_3
l_2
l_1
1
2
3
4
Enhanced Patch
Discriminator
$\mathcal{L}_{pyr}$(,)+$\mathcal{L}_{pyr}$(,)+$\mathcal{L}_{pyr}$(,)
g_4
g_3
g_2
Pyramid Loss
$\mathcal{L}_{rec}$(g_1 ,)
Reconstruction Loss

Figure 2. DL-mode. On the left: Laplacian pyramid decomposition over patches **I'** with exposure artefacts and Gaussian pyramid decomposition over ground truth patches **T**. On the right: $\mathcal{L}_{pyr}$ is computed with the up-sampled output from sub-networks **1**,**2** and **3**, whereas $\mathcal{L}_{rec}$, $\mathcal{L}_{SSIM}$ and $\mathcal{L}_{adv}$ are computed with the final up-sampled output **Y** from the sub-network **4**. In addition, when the discriminator network is enabled, it is simultaneously trained with the final output and its respective ground truth.

Processing input image **I** in this manner permits to independently deal with each sub-net output and compute Pyramid Loss $\mathcal{L}_{pyr}$. This loss is the weighted sum of the four $\mathcal{L}_1$ losses, one for each last three LP level predictions. Thus, in order to compute $\mathcal{L}_{pyr}$, the target for each level is given by the Gaussian pyramid GP of the patch extracted from ground truth T. For $GP = \{g_1, g_2, g_3, g_4\}$, $\mathcal{L}_{pyr}$ is computed as follows:

$$\mathcal{L}_{pyr} = \sum_{i=2}^{4} 2^{i-2} \mathcal{L}_1(g_i, \hat{l}_i) \tag{1}$$

The value of $\mathcal{L}_{rec}$ is also based on the $\mathcal{L}_1$ loss, which measures the pixel-wise error between the prediction and the ground truth patches T_j as shown in (2), where j is the j-th patch extracted from the frame.

$$\mathcal{L}_{rec} = \mathcal{L}_1(T_j, Y_j) \tag{2}$$

The last sub-net makes the final prediction Y_j, which is used to compute three of the losses: i) a Reconstruction Loss $\mathcal{L}_{rec}$ ii) an Adversarial Loss $\mathcal{L}_{adv}$ and iii) a Structural Similarity Loss $\mathcal{L}_{SSIM}$. In [15], Shao et al. combined the $\mathcal{L}_1$ and $\mathcal{L}_{SSIM}$ losses to improve the enhancement results in comparison to a single use of the $\mathcal{L}_1$ loss. As discussed above, both the $\mathcal{L}_{pyr}$ and $\mathcal{L}_{rec}$ losses are based on $\mathcal{L}_1$ and thus, by adding the $\mathcal{L}_{SSIM}$ loss (see Eq. 3) to the overall training objective, enforces the model to learns from the pixel distribution in the ground truth patch, thus leading to a model with more consistent outputs without increasing the inference time of the method. The attractive characteristic of SSIM loss is the fact that it has been proved to be successful when dealing with complex illumination changes [16]. This fact enabled the proposed approach to improve the results of the original LMSPEC implementation.

$$\mathcal{L}_{SSIM} = (1 - SSIM(T_j, Y_j))/2 \tag{3}$$

For preserving realism, LMSPEC integrates a discriminator, which takes **Y** as input and returns a scalar score that indicates how realistic the image looks like. This block is trained along the main network and is used for computing an adversarial loss $\mathcal{L}_{adv}$ shown in (4).

In this contribution, the loss function for optimizing the discriminator:

$$\mathcal{L}_{adv} = -3hwn \log(S(D(Y_j))), \tag{4}$$

where n is the number of pyramid levels (4 in this paper) and $S(D(Y_j))$ is the sigmoid function applied to the D value of the final prediction or the generated image. The complete loss function is then computed as follows:

$$\mathcal{L} = \alpha \mathcal{L}_{pyr} + \beta \mathcal{L}_{rec} + \gamma \mathcal{L}_{SSIM} + \delta \mathcal{L}_{adv}, \tag{5}$$

where, α, β, γ and δ are regularization weights. Figure 2 shows how each single loss was computed through out the entire pipeline.

5. Experimental Setup and Model tuning

The DL-model parameter tuning was carried out as follows. First, an ablation study was carried out to determine the appropriate values of the regularization parameters (α, β, δ and γ). Then, we fine-tuned the best of these configurations. Furthermore, a three-fold training stage was performed with a single underexposed (UE) dataset, a single overexposed (OE) dataset, and a combined over-underexposed (C) dataset (as in [4]). The best model from the ablation study was given by following parameter configuration: $\alpha = \beta = \delta = 0.25$ and $\gamma = 1.0$. This setting gives a strong importance to the SSIM term, which allows to preserve texture details. Moreover, this model (*Baseline*) and LMSPEC were initially trained with original hyper-parameters as shown in upper part of Table 1. Since the

Table 1. Hyper-parameter configurations. Phase 1 (128 pixels patches) in white rows, phase 2 (256 pixels patches) in gray.

Method	Training Set	Epochs	DSE	BS	lr_G	lr_D
LMSPEC [4] LMSPEC+	UE, OE, C	40	-	32	10^{-4}	10^{-5}
		30	15	8	10^{-4}	10^{-5}
Best models*	UE	50	-	32	10^{-4}	10^{-5}
		40	20	8	10^{-4}	10^{-5}
	OE	40	-	64	2×10^{-4}	2×10^{-5}
		30	15	32	2×10^{-4}	2×10^{-5}
	C	50	-	32	10^{-4}	10^{-5}
		40	20	8	10^{-4}	10^{-5}

*Fine-tuned models for each sub-dataset. DSE: discriminator starting epoch.
BS: batch size. lr: learning rate.

input type used in this contribution is different from the one in the original implementation, the hyper-parameters were tuned to maximize the performance on each training sub-dataset (UE, OE and C). The best hyper-parameters after training with each sub-dataset yielded three fine-tuned separated models, as seen in the lower part of Table 1. It is worth noticing that each training was done in two phases as follows: first the trained used 128 pixel square patches, then the weights were transferred as initialization of the second training phase with 256 pixel patches. For this second training phase, the discriminator was enabled at certain discriminator starting epoch (DSE) specified in configurations in Table 1, thus turning the network into a GAN-like architecture.

6. Results and Discussion

6.1. Quantitative Results

Table 2 summarizes the results for the inference phase for each exposure type, i.e., the UE and OE models were tested in over- and underexposure patch sets respectively, whereas the model C was tested over both (separated) test sets.

The results of the proposed model are compared with those of the baseline LMSPEC model. Table 2 shows that the proposed method outperforms LMSPEC best model (either for separated sets or combined) outperforms LMSPEC in terms of SSIM by 4.40% and 4.21% for over- and underexposed images, respectively. Therefore, also note that best performance of our proposed method was given by training our proposed model plus fine-tuning with separated datasets.

6.2. Qualitative Results

Figure 3 shows a qualitative comparison for a couple of frames from the Endo4IE [6] dataset. From the zoomed areas in the third and fourth columns (images enhanced by LMSPEC and the proposed method, respectively) it can be observed that the proposed method is able to produce a much more reliable prediction in comparison to the ground truth (first column), both for over- and underexposed frames (second column). However, a slight change in hue is introduced by both methods.

Table 2. Quantitative results on th Endo4IE dataset [5]. White rows: independent-data training. Light gray: combined-data training as in [4]. Highest criterion values are in bold.

	Overexposure		Underexposure	
Method	PSNR↑	SSIM↑	PSNR↑	SSIM↑
LMSPEC [4]	21.846	0.744	**24.204**	0.757
	22.286	0.772	23.064	0.760
Baseline	22.633	0.799	23.720	0.783
	22.442	0.795	22.877	0.786
Baseline*	**23.139**	**0.806**	24.201	**0.792**
	22.704	0.801	23.229	0.786

*Proposed model + fine-tuned model.

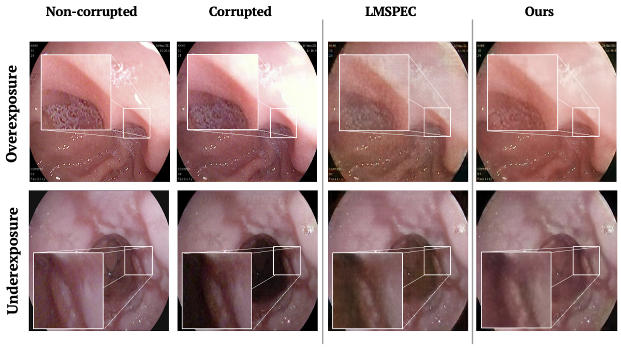

Figure 3. Visual assessment of the exposure correction and structure preservation. The structural enhancement is perceptible in the zoomed areas. The complete images include less artifacts.

7. Conclusions and Future Work

It was shown that the proposed extension of LMSPEC, in the form of an extra loss term for preserving texture details in exposure corrected images has yielded satisfactory results in the Endo4IE dataset: the experiments show a boost in terms of quantitative metrics, and a qualitative assessment has shown that the method produces more realistic images. However, some improvements are still possible: i) although the model makes use of 7 million parameters, we have been able to attain only a 8 FPS throughput (high inference time), and ii) the model sometimes produces images with a slight shift in hue. The last issue can probably be addressed by enforcing color preservation via an additional loss, while the former issue requires improvements in the model design.

References

[1] J. C. Ángeles Cerón, G. O. Ruiz, L. Chang, and S. Ali, "Real-time instance segmentation of surgical instruments using attention and multi-scale feature fusion," *Medical Image Analysis*, vol. 81, p. 102569, 2022.

[2] O. Zenteno, D.-H. Trinh, S. Treuillet, Y. Lucas, T. Bazin, D. Lamarque, and C. Daul, "Optical biopsy mapping on endoscopic image mosaics with a marker-free probe," *Computers in Biology and Medicine*, vol. 143, p. 105234, 2022.

[3] A. Martínez, D.-H. Trinh, J. El-Beze, J. Hubert, P. Eschwege, V. Estrade, L. Aguilar, C. Daul, and G. Ochoa, "Towards an automated classification method for ureteroscopic kidney stone images using ensemble learning," in *42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2020, pp. 1936–1939.

[4] M. Afifi, K. G. Derpanis, B. Ommer, and M. S. Brown, "Learning multi-scale photo exposure correction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9157–9167.

[5] A. Garcia-Vega, R. Espinosa, G. Ochoa-Ruiz, T. Bazin, L. E. Falcon-Morales, D. Lamarque, and C. Daul, "A novel hybrid endoscopic dataset for evaluating machine learning-based photometric image enhancement models," *arXiv preprint arXiv:2207.02396*, 2022.

[6] G. Garcia-Vega, Axel; Ochoa and R. Espinosa, "Endoscopic real-synthetic over- and underexposed frames for image enhancement," 2022. [Online]. Available: https://data.mendeley.com/datasets/3j3tmghw33/1

[7] C. Wang, J. Peng, and Z. Ye, "Flattest histogram specification with accurate brightness preservation," *IET Image Processing*, vol. 2, no. 5, pp. 249–262, 2008.

[8] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," *CoRR*.

[9] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1780–1789.

[10] Y. Wang, R. Wan, W. Yang, H. Li, L.-P. Chau, and A. Kot, "Low-light image enhancement with normalizing flow," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2604–2612.

[11] Ali, F. S., Zhou, B. Braden, and et al., "An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy." *Scientific Reports*, vol. 10, p. 2748, 2020.

[12] S. Ali, F. Zhou, A. Bailey, B. Braden, J. E. East, X. Lu, and J. Rittscher, "A deep learning framework for quality assessment and restoration in video endoscopy," *Medical Image Analysis*, vol. 68, p. 101900, 2021.

[13] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, D. Johansen, C. Griwodz, H. K. Stensland, E. Garcia-Ceja, P. T. Schmidt, H. L. Hammer, M. A. Riegler, P. Halvorsen, and T. de Lange, "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," vol. 7, 2020.

[14] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[15] S. Shao, Z. Pei, W. Chen, W. Zhu, X. Wu, D. Sun, and B. Zhang, "Self-supervised monocular depth and ego-motion estimation in endoscopy: appearance flow to the rescue," *Medical image analysis*, vol. 77, p. 102338, 2022.

[16] J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised scale-consistent depth and ego-motion learning from monocular video," *Advances in neural information processing systems*, vol. 32, 2019.