A deep learning approach for detecting and correcting highlights in endoscopic images

Antonio Rodríguez-Sánchez[1], Daly Chea[1], George Azzopardi[2] and Sebastian Stabinger[1]
[1]Intelligent and Interactive Systems
Department of Computer Science
University of Innsbruck, Innsbruck, Austria
email: antonio.rodriguez-sanchez@uibk.ac.at
[2]Johann Bernoulli Institute for Mathematics and Computer Science
University of Groningen
Groningen, the Netherlands

***Abstract*— The image of an object changes dramatically depending on the lightning conditions surrounding that object. Shadows, reflections and highlights can make the object very difficult to be recognized for an automatic system. Additionally, images used in medical applications, such as endoscopic images and videos contain a large amount of such reflective components. This can pose an extra difficulty for experts to analyze such type of videos and images. It can then be useful to detect - and possibly correct - the locations where those highlights happen. In this work we designed a Convolutional Neural Network for that task. We trained such a network using a dataset that contains groundtruth highlights showing that those reflective elements can be learnt and thus located and extracted. We then used that trained network to localize and correct the highlights in endoscopic images from the El Salvador Atlas Gastrointestinal videos obtaining promising results.**

***Keywords*— Image processing theory, Image processing tools, Image processing applications, Template, Typesetting.**

I. Introduction

Specular and diffuse reflections present in images can be a nuissance for algorithms dealing with stereo matching, segmentation, tracking, object recogntion and other applications. The appearance of a surface can significantly vary in the presence of reflected lights. For those applications, reflections may cover surface details and appear as additional features that are not intrinsic to the object. Highlights can have more serious consequences in cases such as when present in medical images. They can pose a factor in the correct evaluation of an image or video, or at least make more difficult such evaluation from an expert. One such example is cervical cancer screening; i.e. to detect precancerous lesions during digital colposcopy. These images contain reflective components that generally appear as bright spots heavily saturated with white light. Another example is endoscopic examination, where pictures from the inside of the human body are displayed on a computer monitor. They often contain large areas with light reflections. Usually the physician can avoid these highlights by changing the perspective, turning the tip of the endoscope. However, this solution is not effective in case of a camera-in-pill examination because it is not possible to force the pill to move to a better position.

9978-1-5386-1842-4/17/$31.00

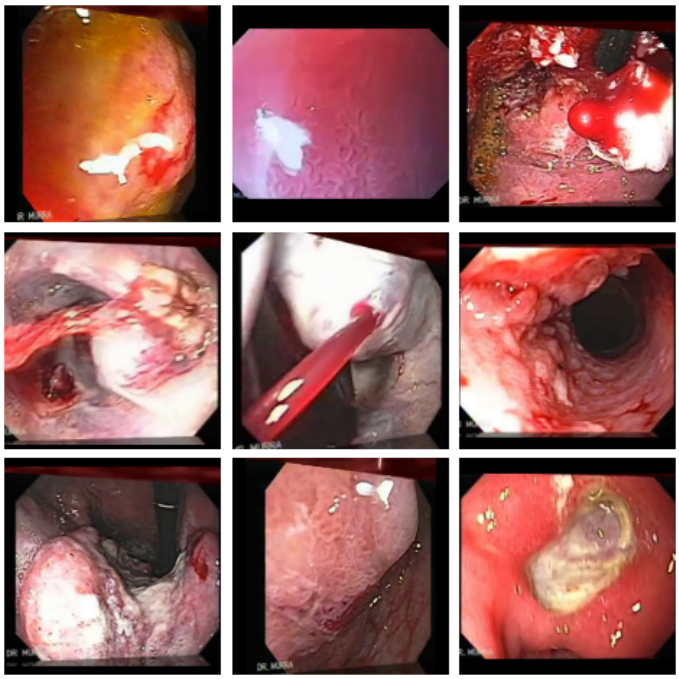

Fig. 1: Endoscopic images obtained from the El Salvador Atlas of Gastrointestinal videos.

There is a large and interesting amount of work in highlight segmentation and removal in computer vision. Most of it deals with the separation between diffuse and specular reflectance. We can classify them depending on whether they use a polarization reflectance model [1], color spaces [2] or segmentation [3], diffusion [4] or a multiview (stereo, motion) approach [5], [6].

Concerning medical images, some algorithms deal just with specular highlights and thus apply intensity thresholding [7], [8], [9]. These thresholding methods use either a fixed range of intensity values or implement a method where these thresholds are adaptive in order to overcome the problem of having to deal with different thresholds. The main problem of these thresholding methods is the over/under-estimation of highlight areas. Another group of methods rely either on averaging

[10], [11] or on difussion [12], [13]. Averaging algorithms consist of selecting a specific neighborhood (e.g. L-shape) and computing the average or median intensity of that neighborhood. These methods rely on the assumption that the area underneath of the highlight is homogeneous, failing then in the case of large highlight areas (such as the ones in endoscopic videos). Diffusion methods rely upon what is called *digital inpainting*, which is image interpolation and suffer from being insufficiently efficient. More recently, hybrids that combine both methodologies have also been presented [14], [15].

We propose here a method where the type of highlight is not important. Our method does not distinguish between diffuse or specular highlights, as both can be problematic. It does not make any assumptions either, but it learns the attributes that correspond to highlights using a deep learning architecture similar to the one used for object segmentation. We use a dataset of natural images that contains groundtruth highlight labelling for diffuse and directed lighting to train the network. This trained network is then used for the analysis of highlight areas in endoscopic medical images, of the type shown in Fig. 1.

II. Methods

A. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a deep learning method first introduced by LeCun *et al.* in 1989 [16]. They gained renewed interest in 2012 when Krizhevsky *et al.* [17] won the ImageNet [18] competition, for image classification, by a large margin using a CNN. Since then, CNNs were adapted to tasks other than classification and have become one of the work horses for machine learning and computer vision. In this paper we use an adapted version of the SegNet [19] architecture for reflection segmentation as well as color correction.

B. SegNet

SegNet has become the standard architecture for image segmentation. It follows a general encoder–decoder architecture, somewhat similar to autoencoders. The encoder part consists of a VGG16 network [20] without the fully connected layers at the end that are usually used for classification. In VGG16 a sequence of convolutional and pooling layers is repeated until a sufficiently small resolution is reached. In the standard SegNet architecture, this pattern repeats five times. This highly reduced information is then decoded by repeated upsample and convolution layers until the output has the same resolution as the input image. Information about which elements were chosen during pooling are forwarded to the upsample layers to improve upsampling. A softmax layer is used to get class probabilities for each pixel position.

C. Highlight-Segnet

This Segnet architecture is trained on pairs of images and dense per–pixel–labels. For the task of reflection segmentation, the input is an image with reflections. The labels specify for each pixel whether it is part of a reflection or not. After training, new images can be presented, and the network will determine a reflection probability for each pixel. For color reconstruction, the input is the reflection part of an image and the per pixel labels are colors from a discrete set of possible colors which were selected by octree color quantization [21]. The reason behind the use of octree color quantization is that In the color correction task we have to consider that we would have a very large number of classes if we consider every color in RGB space ($2^{24} = 16,777,216$). We then need to reduce that large number of classes. Figure 2 shows the architecture of the proposed system.

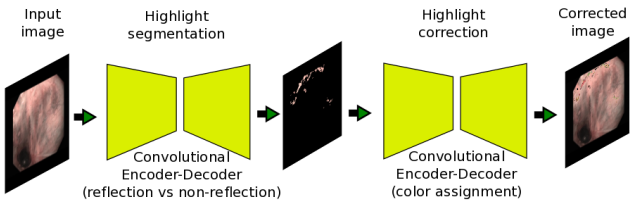

Fig. 2: Architecture of Highlight-Segnet

We used the implementation of SegNet provided with the Caffe [22] deep learning framework. The number of neurons in the last layer were adapted to correspond with our class numbers.

III. Experimental evaluation

We perform two types of analysis. The first one consists of segmenting areas of the images that contain highlights, while the second involves removing those highlights and replacing them with their corresponding values under ambient light illumination. Most works on highlight detection and removal report their results on sample images, here we include as well a quantitative analysis. For this purpose we use a dataset that includes the groundtruth highlight segmentation and the Euclidean distance to the correct color under ambient light conditions over a set of 200 images. Images are of size 150×150 pixels, and since the input to the training are the pixel values and their labels, the learning algorithm has then a total of 4,500,000 classification samples. The datasets used, training, preprocessing and experimental results are described below.

A. Datasets

We use two publicly available datasets for the evaluation of our methodology. The Purdue RVL SPEC-DB Color Image dataset [23][1] contains 300 real images with specular highlights under three different conditions, namely ambient, diffuse and directed. This dataset comes with a groundruth segmentation of the image reflection areas for 200 of those images. There

[1] https://engineering.purdue.edu/RVL/Database/specularity_database/index.html

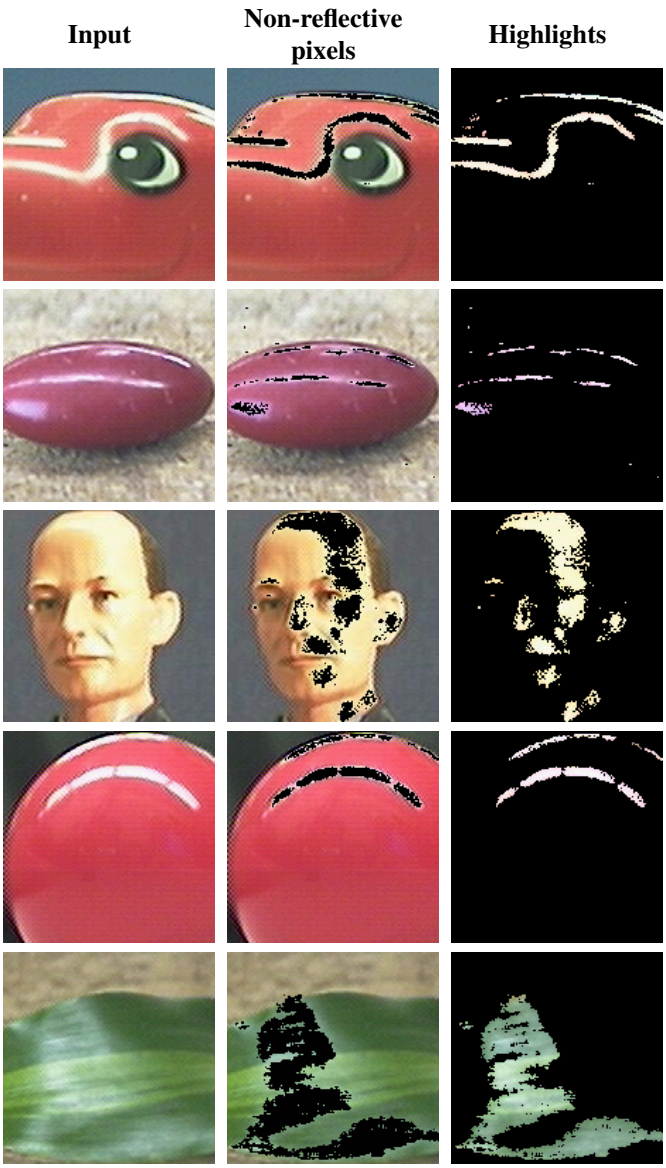

Fig. 3: Examples of highlight segmentation for the Purdue RVL SPEC-DB Color Image dataset

were no constraints on the type of materials used. This dataset is used for training the model and a first evaluation of its performance. Another suitable dataset is the one reported in [4]. this is, however, not publicly available and hence we could not use it.

The second dataset, El Salvador Atlas of Gastrointestinal Image dataset[2] contains 4446 video clips from which we extracted 340 endoscopic frames randomly from 15 different random videos. As there is no groundtruth images from these videos, we will use the images extracted from this dataset purely for qualitative evaluation purposes. Figure 1 shows some examples of these video frames.

[2]http://www.gastrointestinalatlas.com/english/english.html

Table 1: Reflection and Non-reflection Segmentation Precision, Recall and F1-score for our model trained on the Purdue RVL SPEC-DB Color Image Dataset, and thresholding methods on the 40 test images

	Precision	**Recall**	**F1-score**
Highlight-Segnet	**0.80**	**0.81**	**0.80**
Threshold (75)	0.43	0.96	0.59
Threshold (80)	0.52	0.93	0.67
Threshold (85)	0.32	0.86	0.47
Threshold (90)	0.18	0.52	0.27

B. Training

We use two different networks, one for each task, namely segmenting reflective pixels (two classes) and correcting those pixel values (75 classes as described below in section III-D). We train each model using 150000 iterations, in batches of 4, with a decay rate of 0.9, and with a base learning rate of 0.001. For every 100 iterations, we test the learned network on the validation data. The weights in all layers are initialized with random values from 0.1 to 0.5.

C. Segmenting highlight areas

We train the network described in section II-C for classifying two types of pixels: reflective and non-reflective. Since the Purdue dataset contains the groundtruth reflective areas, we train the network using this dataset. Most works in the field of highlight detection do not perform a quantitative analysis. But, using this dataset will allow us to perform a quantitative analysis of our model. Training was done with 160 images randomly selected, and the evaluation set consisted of the remaining 40 images. We will evaluate the model using the classical Precision, Recall and $F1$ score. They are given by combinations of the true positives T_p, false positives F_p and false negatives F_n,

$$P = \frac{T_p}{T_p + F_p}, \; R = \frac{T_p}{T_p + F_n}, \; F1 = \frac{2PR}{P + R} \qquad (1)$$

where P is the precision and R is the recall. The perfect case would have a recall value of 1 for any precision. The $F1$ measure summarizes both as their harmonic mean.

Figure 3 shows examples of highlight segmentation for the Purdue dataset. A quantitative analysis is provided in table 1. In order to confirm that the network was not just learning high intensity pixel values, we include four different threshold values (above 75%, 80%, 85% and 90% brightness) – following the works based on thresholding (see section I) – for comparison.

The trained network using the Purdue dataset images and groundtruth is used for segmenting highlights in the 340 endoscopic frames extracted from the El Salvador Atlas Gastrointestinal videos. Results are shown in Fig. 3. We cannot report on a quantitative analysis in this case as there is no groundtruth for this dataset. Nevertheless, this evaluation shows the cross-dataset validity of our approach. Even though

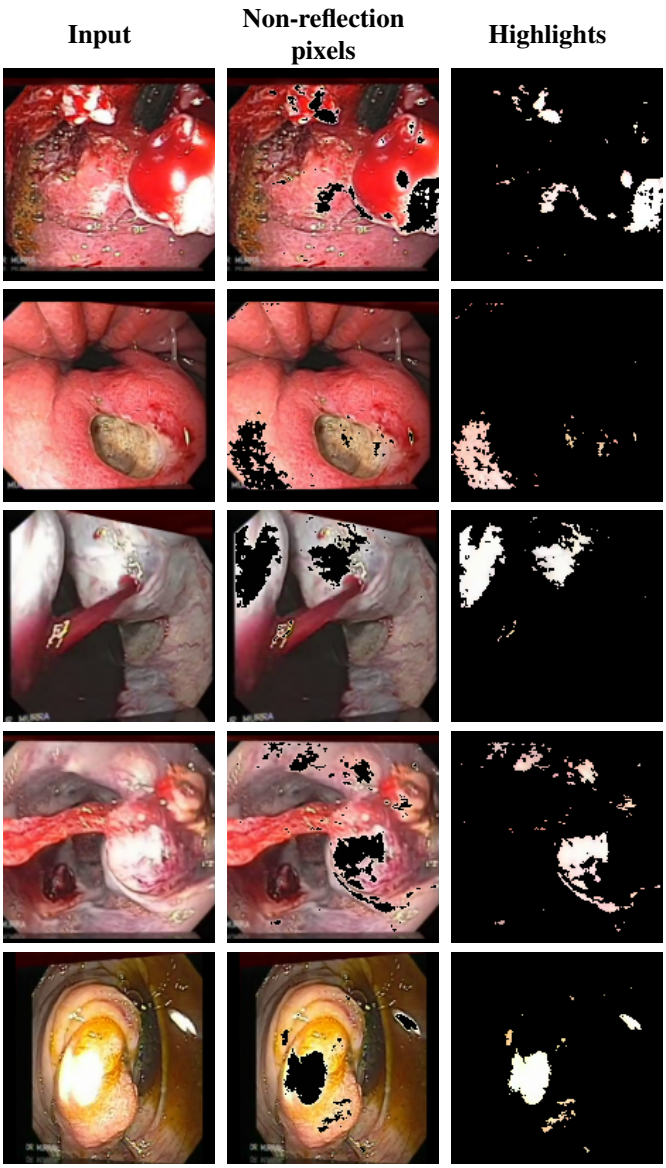

Fig. 4: Highlight segmentation for El Salvador Atlas of Gastrointestinal Image Dataset

the training consisted of natural images, we can see that the trained network can be applied to another very different type of images, from medical endoscopy in this case. Figure 5 shows a comparison of the results from the network to the two best performing threshold values (75 and 80). Processing an endoscopy image would take 0.030 seconds, which can be considered real-time. No preprocessing was applied on the images from both datasets.

D. Highlight correction

For the trained network in the color correction task we use an octree color quantization [21] to reduce the number of classes to 75. The reason we chose 75 is that after trying smaller and larger number of classes, we obtained the best results with 75 (using the Purdue RVL SPEC-DB dataset as

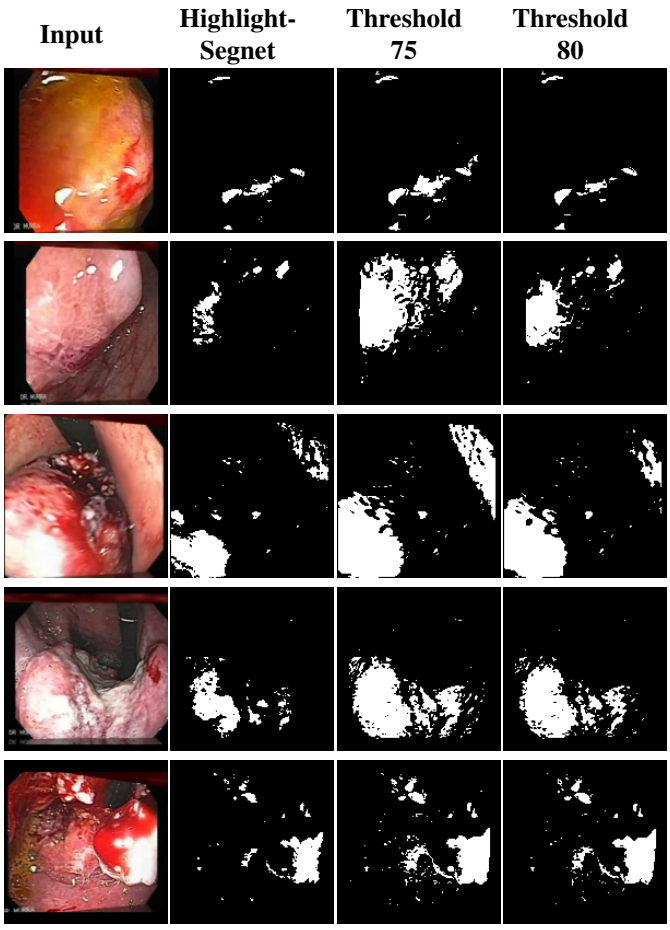

Fig. 5: Highlight segmentation for El Salvador Atlas of Gastrointestinal Image Dataset

our reference). We then map the pixels from the training images to our set of quantized color map by assigning them to the respective color class obtained from the network used from the experiments in section III-C. For this purpose we use the Euclidean distance value from the pixel to each one of the 75 classes and assign it to the closest one. The input to the network is the segmented highlight region. We only used the Purdue highlight dataset for this task since it was the one that provided images with highlights and their respective non-highlight counterparts (ambient light). Figure 6 shows some results of highlight correction.

In order to perform a quantitative evaluation, we used two measures, the RGB Euclidean distance and the ΔE_{ab} color difference which was established by the International Commission of Evaluation (CIE). The RGB average Euclidean distance ($AvgED$) is computed as follows:

$$AvgED = \frac{\sum_{i=1}^{N} \sqrt{(R - R_o)^2 + (G - G_o)^2 + (B - B_o)^2}}{N \times W \times H}, \quad (2)$$

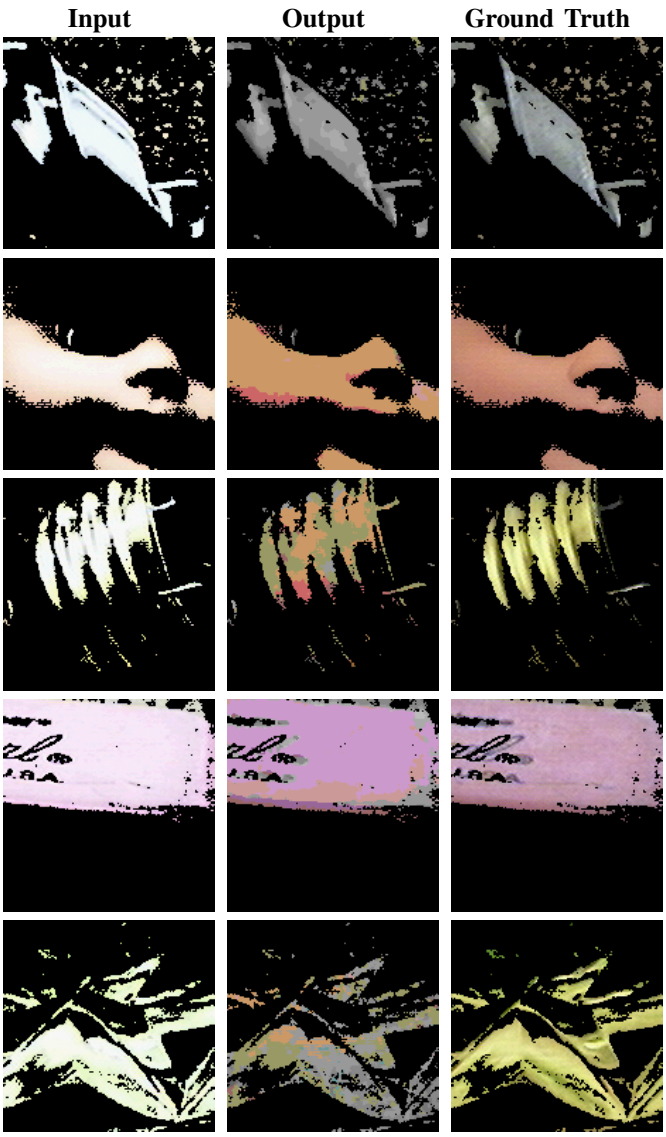

Fig. 6: Colors correction examples of Purdue RVL SPC-DB dataset. Input images are the segmented highlight regions from the highlight segmentation network.

where N is the number of images, W is the image's width, and H is the image's height. R, G, and B are the three color values of a pixel in the output image, while R_o, G_o and B_o are three color values of the corresponging pixel from the groundtruth image (before quantization).

We also evaluated the Lab space ΔE_{ab} difference since RGB color space is not perceptually uniform and the distance in Lab color space is more consistent with the human perception of colors. The ΔE_{ab} color difference which was established by the International Commission of Evaluation (CIE). The Average ΔE_{ab} is computed as follows:

Table 2: Reflection Removal: Average Euclidean distance (RGB) and Average ΔE_{ab} difference (CIE L*a*b* color space) results from our trained model, Interpolation and Median filters.

	$AvgED$ **(RGB)**	$Avg\Delta E_{ab}$ **(Lab)**
Highlight-Segnet	**38.79**	**17.36**
Interpolation Filter	163.54889	41.887936
Median Filter	217.11523	60.94023

$$Avg\Delta E_{ab} = \frac{\sum_{i=1}^{N} \sqrt{(L - L_o)^2 + (a - a_o)^2 + (b - b_o)^2}}{N \times W \times H}, \tag{3}$$

where N is the number of images, W is the image's width, and H is the image's height. L, a and b are the lighness, green-red and blue-yellow color opponnents of a pixel from the output network images, while L_o, a_o and b_o are the Lab values of a pixel from the groundtruth images (before quantization).

Table 2 shows both the average Euclidean distance in RGB and the average ΔE_{ab} difference in Lab color space.On a typical scale, the ΔE_{ab} value will range from 0 to 100. Our Average ΔE_{ab} is 17.36 which indicates that the output colors from our network are quite close to the groundtruth and much better than for example using an interpolation or median filters as used in a number of previous works (see I).

A few qualitative samples for Image highlight correction for endoscopic medical images is shown in Fig. 7, in this case we can see that the results are far from optimal and we definitely need to obtain better results for these type of images, either by increasing training size or using a different method other than the octree quantization. Processing an endoscopy image would take 0.024 seconds, which can be considered real-time. If we add this value to the segmentation timing (0.030+0.024), segmenting and correcting an endoscopic image takes less than 0.05 seconds.

IV. Conclusions

We have presented two convolutional encoder-decoder approach to segment and correct highlights from images. Reflection elements are learned by the network from a dataset that contains ambient and specular highlights. This learned network shows highlight segmentation and correction capabilities when faced with medical endoscopic images in real time, a type of images which the network was not confronted to during training. Even though results are quite promising, there is still much work to do. The color-corrected endoscopic images are far from perfect. For future work we would like to explore a modified version of the correcting network where the pixel values are obtained through regression instead of classification, thus avoiding the octree quantization process. Additionally we would need more training images. We also plan to explore the performance of our system with other types of medical images.

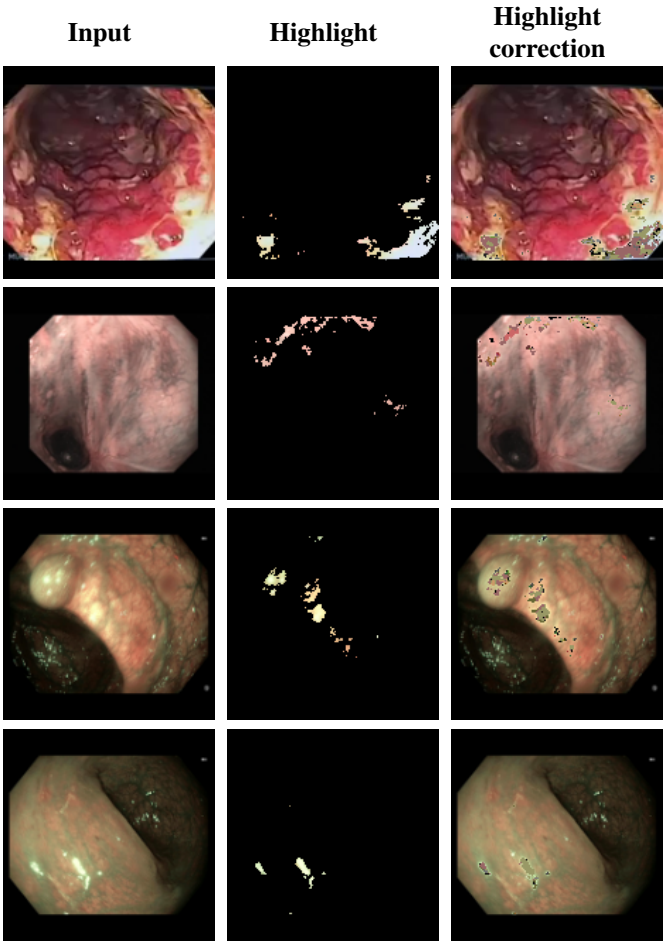

Fig. 7: Highlight correction examples for El Salvador Atlas of Gastrointestinal Image Dataset.

Nevertheless we can conclude that deep learning algorithms can be of great use to analyze highlights and medical images.

REFERENCES

[1] S. K. Nayar, X.-S. Fang, and T. Boult, "Separation of reflection components using color and polarization," *International Journal of Computer Vision*, vol. 21, no. 3, pp. 163–186, 1997.

[2] S. P. Mallick, T. E. Zickler, D. J. Kriegman, and P. N. Belhumeur, "Beyond lambert: Reconstructing specular surfaces using color," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. Ieee, 2005, pp. 619–626.

[3] R. Bajcsy, S. W. Lee, and A. Leonardis, "Detection of diffuse and specular interface reflections and inter-reflections by color image segmentation," *International Journal of Computer Vision*, vol. 17, no. 3, pp. 241–272, 1996.

[4] R. T. Tan and K. Ikeuchi, "Separating reflection components of textured surfaces using a single image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 2, pp. 178–193, 2005.

[5] Y. Sato and K. Ikeuchi, "Temporal-color space analysis of reflection," *JOSA A*, vol. 11, no. 11, pp. 2990–3002, 1994.

[6] S. Lin, Y. Li, S. B. Kang, X. Tong, and H.-Y. Shum, "Diffuse-specular separation and depth recovery from image sequences," in *European conference on computer vision*. Springer, 2002, pp. 210–224.

[7] G. Zimmerman-Moreno and H. Greenspan, "Automatic detection of specular reflections in uterine cervix images," *Medical Imaging*, vol. 6144, no. 1, pp. 61 446E–61 446E, 2006.

[8] M. Arnold, A. Ghosh, S. Ameling, and G. Lacey, "Automatic segmentation and inpainting of specular highlights for endoscopic imaging," *EURASIP Journal on Image and Video Processing*, vol. 2010, no. 1, p. 814319, 2010.

[9] R. Richa, P. Poignet, and C. Liu, "Three-dimensional motion tracking for beating heart surgery using a thin-plate spline deformable model," *The International Journal of Robotics Research*, vol. 29, no. 2-3, pp. 218–230, 2010.

[10] H. Greenspan, S. Gordon, G. Zimmerman, S. Lotenberg, J. Jeronimo, S. Antani, and R. Long, "Automatic detection of anatomical landmarks in uterine cervix images," *IEEE Transactions on Medical Imaging*, vol. 28, no. 3, pp. 454–468, 2009.

[11] A. Al-Surmi, R. Wirza, M. Z. Dimon, R. Mahmod, and F. Khalid, "Three dimensional reconstruction of human heart surface from single image-view under different illumination conditions," *American Journal of Applied Sciences*, vol. 10, no. 7, p. 669, 2013.

[12] C.-A. Saint-Pierre, J. Boisvert, G. Grimard, and F. Cheriet, "Detection and correction of specular reflections for automatic surgical tool segmentation in thoracoscopic images," *Machine Vision and Applications*, vol. 22, no. 1, pp. 171–180, 2011.

[13] G. Karapetyan and H. Sarukhanyan, "Automatic detection and concealment of specular reflections for endoscopic images," in *Computer Science and Information Technologies (CSIT), 2013*. IEEE, 2013, pp. 1–8.

[14] A. I. Avilés Rivero, P. Sobrevilla Frisón, and A. Casals Gelpi, "An approach for physiological motion compensation in robotic-assisted cardiac surgery," *Experimental & Clinical Cardiology*, vol. 20, no. 22, pp. 6713–6724, 2014.

[15] S. M. Alsaleh, A. I. Aviles, P. Sobrevilla, A. Casals, and J. K. Hahn, "Automatic and robust single-camera specular highlight removal in cardiac images," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015, pp. 675–678.

[16] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, pp. 1–42, April 2015.

[19] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] M. Gervautz and W. Purgathofer, "A simple method for color quantization: Octree quantization," *New trends in computer graphics*, pp. 219–231, 1988.

[22] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[23] J. B. Park and A. C. Kak, "A truncated least squares approach to the detection of specular highlights in color images," in *Robotics and Automation, 2003. Proceedings. ICRA'03. IEEE International Conference on*, vol. 1. IEEE, 2003, pp. 1397–1403.