Deep Autoencoder Approach to Enhance Low-Light Images during Minimally Invasive Procedures

Caio Jordão de Lima Carvalho
Instituto Federal da Bahia
Salvador, Bahia, Brasil
caio.carvalho@ifba.edu.br

Antonio Carlos dos Santos Souza
Instituto Federal da Bahia
Salvador, Bahia, Brasil
antoniocarlos@ifba.edu.br

***Abstract*—The process of analyzing images from Minimally Invasive Procedures suffers from some problems related to the conditions of these images. Among these problems, it is possible to notice that the illumination issues are one of the most complicated to solve, as these surgeries are performed through small incisions in the human body. Therefore, the present project presents a model based on a deep neural approach that is able to improve low-light surgical images, making them more optimized for analysis processes performed by human surgeons or assistant robots.**

***Index Terms*—Minimally Invasive Procedures, Autoencoders, Image Enhancement, Low-Light Images, Laparoscopy.**

I. Introduction

Minimally Invasive Procedures (MIPs) have become notorious in the modern surgical medicine field because of some factors that can benefit the patient, specially when this category of surgery is compared to the conventional open surgery procedures. Among these beneficial factors, it is possible to focus on the decrease of the hospital stay and the reduction of pain and postoperative traumas [1]. Over the last three decades, these procedures have gained prominence in the medical field, including new techniques and using vanguard technologies to assist all the medical team. The use of robots for automation of some procedures and the remote surgery have been constantly influencing the development of these techniques in comparison with the other ones which are conventionally used.

In addition, robotic assistants are becoming more successful than human assistants in the execution of surgical tasks because of their improved precision when doing some routines. There are two specific categories of assistant robots that can help in MIPs. The first is a robot that is able to do some basic tasks, being specially focused on helping the surgeon, assisting with the application of suture, incisions and anesthesias [2]. The robotic intervention can be done in an automatic way, where the robot can execute all the tasks independently, or in a supervised way, where a surgeon can guide the robot. The other category of robot is only responsible for the control of the microcamera that is inserted inside of the patient's body, allowing the surgeon to follow the procedure in some external screen [3]. To accomplish this task, there are some tracking algorithms that are implemented into the robots' logic, which are mainly focused on following the medical instruments used in the procedure.

Today, robots focused on the microcamera control are being highly used when compared to the other surgical robot types. Computer Vision algorithms are used to help the locomotion of these robots. To follow the medical instruments that are operated by an human surgeon, these robots need to know the area that is occupied by these tools, mapping the entire location of them in real time [4]. This map is acquired using two main techniques. The first of them is known as object tracking, which is characterized mainly by the process of identification of the most relevant points of an object, such as contours or regions of interest (ROIs), applying image filters to outline the desired object and follow its trajectory in a certain environment. The second technique is known as image segmentation, which is characterized by the identification and mapping of image segment that identify an object. These segments are processed as pixel maps, building a polygon over the area of interest.

Among the current automatic algorithms for object tracking and image segmentation, the ones that stand out in the state-of-the-art are those focused in deep learning processes, which are based in the use of neural models with a large number of layers, providing an increased level of abstraction on multidimensional data [5]. One advantage of these models is that they have a considerable good performance when dealing with huge training datasets containing multivariational data, which optimizes their predictive power. However, there are some challenges when using these types of algorithms in the surgical field, being possible to notice that these methods need optimized images, without errors caused by external factors and containing a precise information, as they are acting in high risk processes that demand a certain level of quality. Errors caused by low-lighting, focus and movement blur can make their performance decrease and cause other errors that can be life threatening to the patients that are being operated. Although medical centers use high quality visual equipment, any error due to these factors related to the environment may disrupt the surgical routine. Therefore, the use of automatic image enhancement algorithms in order to optimize the quality of the analyzed frames appear as an aid to these tracking algorithms.

During the image preprocessing stage, enhancements algorithms are capable of applying optimization techniques on them to improve some details for a particular context [6]. In this process of improvement, the focus of the algorithm is to perform a cleaning of all possible imperfections present in a certain frame. Thus, it is possible to generate images with a better quality and the reduction of errors as the focal loss, noise and illumination failures. These changes can be done by using some concepts related to matrix convolution techniques, where the focus is to obtain an optimized filter capable of modifying the image for a certain state. Generative algorithms combined with adversarial learning approaches have been used to find the optimal filter, enhancing the image for a specific context [7]. According to that, the possibility of combining these algorithms to other object tracking routines is observed, improving the image quality through an enhancement method.

Thus, this project presents an artificial neural network model based on an autoencoder architecture to automatically generate filters that are capable to enhance images, removing errors in surgical images caused by lighting failures that can affect the performance of algorithms focused in medical instrument detection and the image analysis that is done by medical and surgical professionals. Two experiments will be executed with the developed model, including an evaluation of its performance and its comparison with a model based on a statistical approach, aiming to analyze the performance of the proposed image enhancement method and the quality of the improved images.

The main contributions of the proposed solution are:

1) Develop an artificial neural network model based on an autoencoder approach to enhance frames captured during minimally invasive procedures.
2) Build a dataset focused on the automatic low-light image enhancement of minimally invasive procedures with variational data composed by different levels of illumination on different surfaces on the interior of the human body.
3) Reduce lighting errors present in surgical images in order to optimize the execution of assistant robots, especially focusing on those responsible for controlling micro-cameras inserted into the patients' bodies, and the analysis of the frames during the surgeries that is done by human surgeons.
4) A comparison of the developed neural model with another model available in the state-of-the-art of image enhancement, specially focusing on models that are based on statistical approaches.

This document is organized as follows: Section II presents the theoretical background that was necessary for the study of this project, describing the state-of-the-art of the areas focused on minimally invasive procedures, deep learning and image enhancement. Section III describes some related works that contain approaches similar to this project. The proposed solution is presented in Section IV, including the description of its architecture and the challenges that were faced during its implementation. The set of tests performed and the results obtained are presented and cataloged in Section V. Lastly, Section VI will conclude this paper by listing the obstacles that were found during the project and describing some future works.

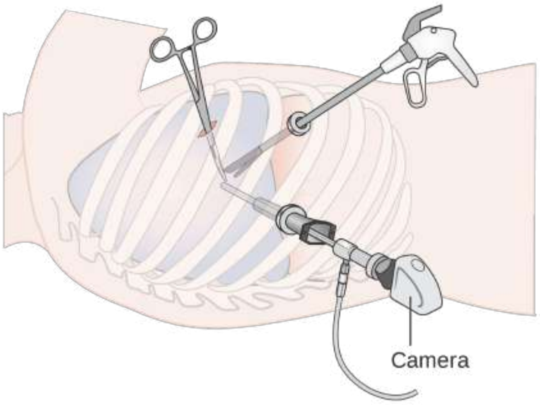


Fig. 1. Thoracoscopic procedure performed with micro-camera. [9]

II. Theoretical Background

A. Minimally Invasive Procedures

Compared with conventional open surgeries, or non-invasive surgeries, minimally invasive surgical procedures appear as a way to reduce some types of damage caused by these alternative surgical methods. One of the main reasons that is associated with this damage mitigation is the reduction of the size of the incision made in the patient's anatomy, which allows physicians to work with smaller openings, requiring only small fissures through which the laparoscopic instruments will be inserted [1]. Because of this advantage, it is possible to obtain a set of benefits, such as reduction of postoperative pains, bleeding and inflammation, including a more pleasant aesthetic result for the operated region.

The history of these methods dates back to the early twentieth century, where some devices were invented to perform procedures in the pelvic region and gastrointestinal system with the use of the cystoscope and the first laparoscopes [8]. After the evolution of these mechanisms, MIPs have started to become popularly studied in the state-of-the-art of surgical medicine in some medical centers on the United States and Europe, specially in the 1980s. Other medical techniques were developed after the inclusion of technologies that were capable to assist these procedures, easing the process through the years.

Among the main types of minimally invasive surgeries, it is possible to identify their use in cardiac, gastrointestinal, gynecological, vascular and thoracic procedures (as shown in Figure 1) [10], [11]. These different types of surgical procedures differ depending on the area where the surgery is performed, being categorized as endoscopic, laparoscopic, arthroscopic and retroperitoneoscopic procedures. In general, all of these procedures fit into the concept of endoscopy, which involves the use of a camera mounted on a platform or device that provides an internal view of the patient's body

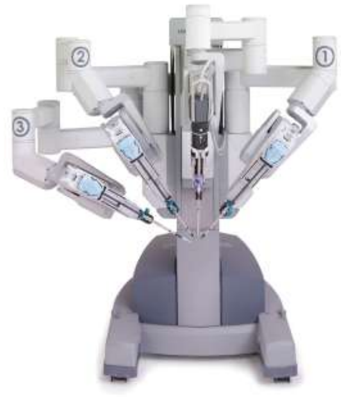


Fig. 2. *da Vinci*, surgery robot used in image-guided laparoscopic procedures. [14]

without the need for wide incisions. However, laparoscopy is mainly characterized as an endoscopic procedure performed in the abdomen region. The procedures of arthroscopy and retroperitoneoscopy are focused on surgeries in the regions of joints (e.g. knees and shoulders) and peritoneum (i.e. adrenal regions or kidneys), respectively.

The laparoscopic procedure is executed with the help of an instrument known as the laparoscope. This instrument is composed by a thin tube containing a camera with a support for lighting [12]. A small incision is done in the patient's body, which is used as a region for the insertion of the laparoscope, allowing the medical team to visualize the operated area. Carbon dioxide is used to inflate the area, increasing the distance between the internal organs and the patient's skin, which can assist the insertion of the instrument and provide a larger space for the movement and execution of the operation.

Despite the high range of applications of these procedures in surgeries, not all the patients are qualified to perform some categories of MIPs. There are some pre-operative conditions that can prevent the patient from becoming fitted for the surgery, such as high body mass index, anatomic incompatibility in the operated area or previous operations that led the patient to obtain abdominal adhesions (especially in the case of surgeries in digestive organs) [13]. However, it has been observed that these procedures have been convenient for an optimization in the necessary operative tasks on the patients who are able to perform these types of operations.

After the creation of telemanipulators in the 1990s, robots have began to be inserted into the surgical field as assistant during the operative procedures. In the beginning, they were used mainly for cholecystectomies, which are characterized as surgeries focused on gallbladder removal [15]. Designed to extend the skills of human surgeons, enhancing their tactile feedback, dexterity and coordination of the instruments, robots have allowed the evolution of minimally invasive procedures to a new level, where an increase in surgical accuracy and a reduction of possible failures caused by human beings is clear. Nowadays, robots like *da Vinci* (Figure 2) are examples used in MIPs in various hospitals and medical institutes around the world [16].

As tools that have enabled the evolution of surgical processes guided by images and videos, these robots have facilitated the execution of surgical telementoring and telemanipulation. In the case of *da Vinci*, a surgeon is able to operate it through a control terminal equipped with tactile controllers. These controllers are capable of sending motion signals to four robotic arms and a stereoscopic camera that are able to aid the surgery and enable the dissection of internal organs remotely. Being created in 2000 and focused on enabling long distance telemanipulated surgeries, *da Vinci* was developed in order to evolve conventional laparoscopic procedures, allowing the surgeon to perform the surgery with improved motion skills due to the sensors included in the robotic arms, and achieving some satisfactory results in thoracic, vascular (bypass procedures) and oncological procedures.

Other examples of assistant robots have appeared in the medicine field over the last years. Projects as the *AutoLap* device [17], showed in Figure 3, which are becoming relatively popular in the state-of-the-art of robotic surgeries, are focused on controlling only the video-camera that is used during the operation, not playing the role of interacting directly in the operative processes. Conventionally, these cameras are controlled by one or more human assistants. Therefore, automating video camera control facilitates this process, allowing only one human, the main surgeon, to perform the procedure. To accomplish this, the robot must initially locate the positions of the medical instruments used by the surgeon and, as they are moved, follow them to provide a better view of the area.

Laparoscopy, such as other types of image-guided surgeries, was one of the main categories of surgical procedures that improved the field of surgical technologies. Previously, only humans were able to perform the task of assisting the main surgeon with the control of the camera that provides the vision of the operated area. However, methods have been developed to make this process fully automated, as it is seen in the case of *AutoLap*. Initially using stereotactic interventions, tomography and sonic/optical localization, images were used as a way to enable the physician to be guided during the procedure in the patient's internal organs without the need for cuts or incisions with a large width. However, the demand for a larger number of images to perform real-time procedures led to the use of high performance cameras. The use of computer vision, augmented reality and 3D rendering techniques in these procedures allowed the evolution of this area to a new level.

In the case of *AutoLap*, several differences arise in comparison to conventional robot assistants that control cameras. These robots are usually oriented through voice and head/eye movements. However, *AutoLap* allows this entire process to be automated through image analysis. Recent studies show a satisfactory stabilization of the tool, as well as security improvements during its use, but there is a lack in some image optimization features that can compromise the performance during its execution [17]. Among the compromising characteristics, it is possible to indicate that the quality of the images analyzed have an important impact on the way that the robot moves, as illumination and motion noises can bring some performance issues. Because of this problem, the difficulty that this type of technology has to be inserted into the current market is increased, influencing directly on how image-oriented surgeries proceed and making it impossible to increase the use of robots such as the *AutoLap*. Algorithms or technologies focused on improving the quality of the analyzed images become necessary for the evolution of these robotic assistants.

B. Deep Learning

The classical artificial intelligence methods have been passing through some significant changes since the creation of machine learning processes oriented by neural models. Unlike the models that follow the statistical learning paradigm, such as Markov chains and bayesian inferences, the way how neural models are built is strongly influenced by the natural organization of the nervous system, having a group of neuron layers that can abstract the evaluated information to a recognizable and normalized pattern [5]. Being initially recognized as supervised learning techniques, these models have actually evolved and have included some other learning categories, such as the non-supervised learning processes, even covering some generative models, where it is possible to autonomously generate new samples of data from an initial dataset with encoding and decoding methods.

However, the conventional neural models, which are commonly known as artificial neural networks, used to have some issues caused by the expansion of the training datasets or the increase on the dimensions of the data, being unable to achieve some data abstraction levels, including the examples where the data required tensor manipulation, as in the case of images. To solve this problem, the deep neural networks were created. These networks are similar to the conventional ones, but they have different types of categories of layers, where each of them has a specific purpose during the process of abstraction that is executed by the network, and include a significant increase in the number of the layers [5].

Fig. 3. *AutoLap*, assistant robot that is capable to control the camera during laparoscopic procedures [17].

Created in the 1990s by the computer scientist Yann LeCun, who is also the main responsible for the creation of the first convolutional neural network (CNNs) model, the architecture *LeNet-5* was the first artificial neural network topology to present a deep learning process [18]. Being used for hand-written and printed digit recognition, this architecture used a group of convolutional, fully-connected and pooling layers, abstracting the images to feature maps until the analyzed data become an uni-dimensional vector to be processed by a softmax function to classify the digits that are present in the evaluated image.

In that time, these networks were mainly used in computer vision processes, having a good performance, but requiring a significant amount of computational power and huge datasets that were not easy to obtain. Because of this situation, the support vector machines (SVMs) were more used than CNNs at that time, specially in trivial classification problems. Nonetheless, since 2012 [19], the interest in these models were restored as the state-of-the-art has experienced an exponential improvement in its performance when used on large collections of data.

The use of convolution in these models have provided a good advance on how artificial neural networks learn, as this is the main operation that is executed inside of the convolutional layers. Largely used by conventional image filters, these operations are capable of transforming a multidimensional data to a tensor known as a kernel and executing an operation based on linear algebra that is much faster than the matrix multiplication. While conventional neural networks use techniques based on vector and matrix multiplication to update their weights, convolutional networks use convolution in at least one of their inner layers [5].

The kernel can be understood as a low-dimension matrix (commonly represented by square matrices of order 2 or 3) that is used to scan the analyzed tensor, doing a mathematical operation on each of the item subsets of the original tensor and generate a feature map as output. This method is used by the CNNs as a way to abstract the input tensor through the layers in a more efficient way.

$$s[t] = (x * w)[t] \tag{1}$$

In the Equation (1), the function s in a time period t will generate a feature map based on the x values, which is the input tensor, and w, which is the kernel. There are different types of convolution and, depending on these types, the dimension of the input can be completely changed after this operation. In the case of a pooling or a up sampling process, a summarized statistic of the analyzed region is generated, resizing the input tensor to a smaller size during the pooling and to a bigger size during the up sampling. In this sense, each type of convolutional operation has its own purpose and the combination of these operations in a network will determine how the output will be after this process.

The usage of convolution on these layers allow the network to apply a group of relevant filters on the data abstraction

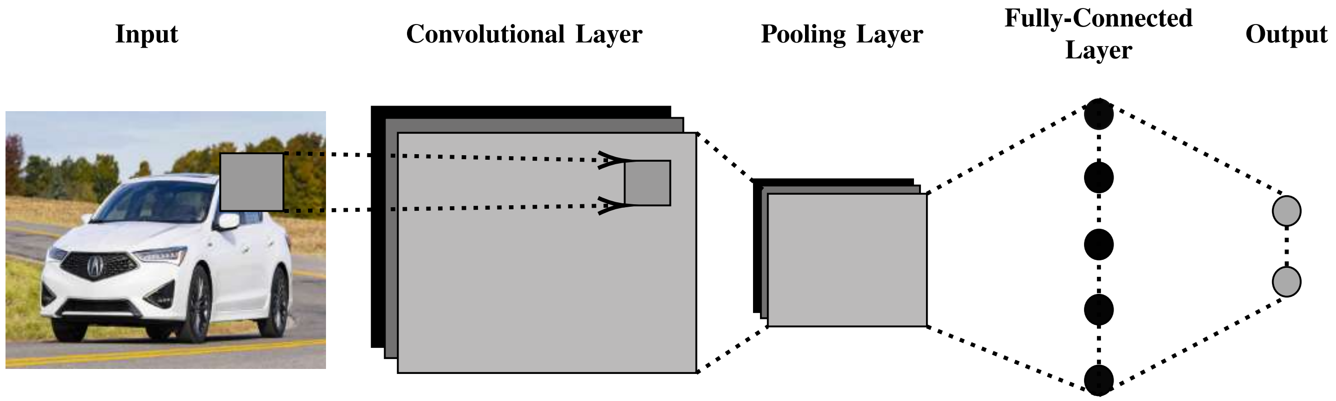


Fig. 4. Example of a convolutional neural network applied to an image classification problem. The input will be processed by a convolutional layer that will send its output to a pooling layer. During the pooling process, the feature maps will have a reduction on its dimensions. Then, a fully-connected layer will receive this data which will finally be transmitted to an output layer that will contain a classification function capable of categorizing the result of the network.

and, because of the decrease of the number of parameters caused by the kernel, there is an improved adaptation of the dataset and smaller chance of occurring errors caused by a process known as overfitting, which allows the model to obtain a good performance with training data, but an unacceptable performance with new data. This process can make the model to be skewed according to a data pattern and can make it stop predicting correctly on new information, such as in test or validation data.

In the field of object detection and automatic segmentation of images, purely convolutional networks had some performance gaps during the detection processes with images containing multiple instances of certain objects, as some errors used to occur with the focus on the regions of interest (ROIs) where the objects were located, and these networks did not have any type of hierarchy that was focused on locations or regions. To solve this problem, [20] proposed a new architectural pattern for convolutional networks where a selective search and grouping are executed for the identification of these regions. Known as region-based convolutional neural networks (R-CNNs), the models based on this new architecture become capable of generating a set of sub-segments and candidate regions, applying a greedy algorithm to combine closer regions and generate wider ones, using them to predict the expected locations.

This proposal has produced an improvement in the feature extraction for the generation of maps in the intermediary (or hidden) layers and, with its subsequent evolution through architectures such as *Fast R-CNN*, *Faster R-CNN* and *YOLO* [21], [22], [23], it has gained significant notoriety for the resolution of classification problems in the field of computer vision. Another architecture that is based on R-CNNs and that achieve a considerable performance is the Mask-RCNN [24], which is mainly applied to instance segmentation processes. This architecture locates candidate regions where the ROIs may be found and, using these regions, include a pixel mask over the object in the region with the highest confidence level.

Even though the deep neural networks focused on classification problems had a significant success and constant advances, there were some discussions about the inclusion of these models in unsupervised and reinforcement learning contexts. It was also discussed about the ability of these networks to work in an environment where the evaluated datasets were not labeled. With the application of these learning categories, it would be possible to evaluate similarities between data units and to recognize patterns present in them, creating groups and identifying relevant differences. To implement this idea into deep networks, two concepts of the classical machine learning area were combined to build new models that are able to learn without using the standard supervised paradigms.

With this type of learning, it is possible to evaluate similarities between data units and to recognize patterns present in them, performing groupings or identifying relevant differences. Two concepts of the machine learning area were united to the construction of new models able to learn without using the standard supervised paradigms, but following the line of algorithms of unsupervised learning and reinforcement. So, approaches were developed using generative models combined with an adversarial reinforcement process [5].

In the statistical field, the generative models can generate data without a specific reference (i.e. having almost no knowledge about the evaluated dataset) [5]. The main idea is that only basic parameters, such as constants and the dimensions of the data, are known and the rest of them is hidden, influencing the model to randomly generate these hidden parameters and assisting it to find a new sample present in that scope. Thus, the Equation (2) will act as a joint probability distribution to approximate randomically the parameters available in the evaluated set to create a new data unit. Bayesian networks, autoencoders and hidden Markov layers are examples of generative models used in the artificial intelligence area.

$$P(X|Y) \tag{2}$$

In a deep learning environment, generative models are highly applied in optimization problems or to learn to produce random examples using an initial dataset, as seen in the case of

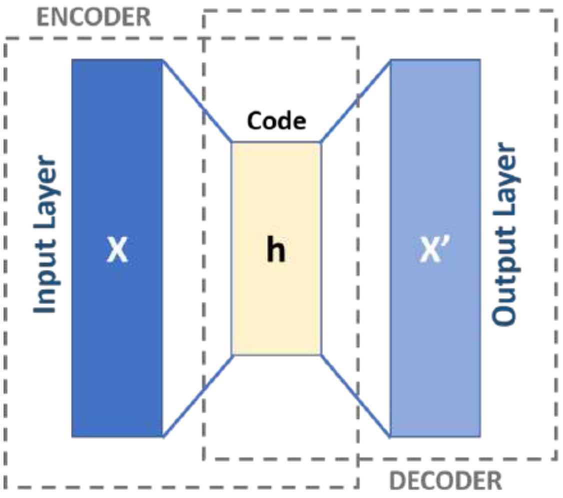


Fig. 5. Example of an autoencoder architecture.

generative-adversarial models (GANs) and autoencoders [5]. In problems related to image-to-image translation and image enhancement, generative models can have a significant performance because of their ability to find encodings between the original data and the resulted output. Following this problem, an autoencoder architecture is composed by two levels that are able to complete this task, which can compress the original data into a pattern and then uncompress it to an output, as shown in Figure 5.

One of the applications of models based on generative architectures is in the image enhancement field [25]. With proposals of variations in the initial architectures, other neural network topologies based on autoencoders and GANs have been developed to solve some other problems. Recent results showed that this type of architecture can be widely used as a way to optimize and automate the image pre-processing stage during the execution of computer vision algorithms. Thus, the applicability of these models in robotics is observed, specially in surgeries that are assisted by video-camera control robots, which need high quality on images during processes of orientation and localization.

C. Image Enhancement

During data analysis procedures, variances between the multiple data instances can represent differences that can influence the result of the algorithm. This problem is applied to the most different types of analysis, because, as the work may involve more data and real-time processing, the divergences can be increased without a proper control, especially when the context involves new data that was not analyzed before, as in the moment where the prediction based on test data is done. According to this situation, enhancement techniques used during the pre-processing stage are necessary to allow the algorithm to be prepared to these differences. In the image processing field, there are some techniques focused on the process of enhancement to optimize classification algorithms [6].

There are two categories of image enhancement methods: spatial approaches and frequency approaches [26]. The spatial approaches are focused on the group of pixels available in the analyzed frames. In this case, the enhancement occurs through the application of a enhancement function in the values available in a selected pixel region. Frequency approaches act in a similar way, but they are focused on the full image and work in a frequency domain (e.g. Fourier transform). Both of these methods work according to (3), where g is the result generated by the function T when it is applied in a selected pixel region (in the case of spatial methods) or in a complete image (in the case where frequency methods are used) known as f. Thus, it is observed that T will search for certain characteristics from f and change the ones that are incorrect to enhance the original data into a improved one.

$$g = T(f) \tag{3}$$

As the images may not always be following a certain pattern of optimization, especially when there are some interference from the environment that may cause some errors, as seen when low lighting or focus loss occur, a fixed enhancement process cannot be successful in all of the cases. Because of that, it is observed that solutions based on automatic enhancement algorithms become necessary, as these algorithms can learn to find the best optimization for the map of parameters used by the function T [26].

Some machine learning approaches are focused on using predictive methods to assist with image enhancement tasks. Thus, a model seeks to find a g function that can significantly enhance images to a specific standard. Generative models are one type of these examples, as in the case of autoencoders, which are capable of performing image-to-image translation tasks. Consequently, it is observed that these models can be used to optimize an image without using a fixed function, adapting the algorithm to images present in different state types and variations, as in the case of the problem that the present project aims to solve.

III. Related Works

In order to optimize medical instrument detection processes in minimally invasive procedures, some methods and approaches are aimed to perform automatic image enhancement using techniques derived from the state-of-the-art of artificial intelligence. Including both the endoscopic field and the laparoscopic area, these models are focused on solving problems that include interferences caused by smoke in images, specular reflections and different types of noises [27], [28], [29], [30], [31]. As can be observed in these studies, one of the challenges currently encountered is that image enhancement models based on classical AI methods make these algorithms specialists in only one type of enhancement technique, and it is difficult to adapt them to a context where it is necessary to carry out a set of processes that aim to optimize different characteristics of an image, specially when the problems are related to light and brightness.

In the case of the models proposed in [27], the focus is mainly on smoke and noise removal on videos of minimally invasive procedures. During the image enhancement process, these two types of errors can cause significant losses in textures and, depending on the concentration of the smoke or noise, may generate errors that are primarily associated with depth and distance factors between objects, which can interfere in the performance of object tracking algorithms that use these images as input. In [28], an unified model based on an algorithm of expectation-maximization with variational Bayesian inferences was developed to perform the enhancement of regions with smoke, reflections and noise. However, the proposed model is valid only for these problems, as it does not cover issues related to illumination issues.

In [29], a generative-adversarial model was developed to optimize the enhancement process of laparoscopic images containing smoke. Using an architecture of image-to-image transformation and applying a perceptual-oriented GAN model that is capable to learn the mapping between deformed input images and restored output images, combining it with a multi-scale structural similarity method, significant qualitative results were obtained, but a difficulty on the application of this model in real-time processing environments due to issues related to its processing time was noticed. However, with this experiment it was possible to notice that this model category can aid the processes of enhancement of surgical images as it considers important characteristics such as texture and depth.

In the field focused on mist and specular reflection removal, there are some approaches focused on image enhancement without the lost of colors on the enhanced regions [32], [31]. In such cases, the chromatic information tends to be modified by the interaction with lighting. During the enhancement, deconstruction processes were performed in these regions, which are capable of causing losses of these types of information, generating errors and failures. In [31], a statistical approach was used to remove fog signals, but there was some errors on images with a high quantity of fog and there was not a routine to verify if the processed image really contained errors or not, which could cause some unnecessary executions.

In some areas that are focused on problems that are different from the medical field, the combination of these automatic image enhancement techniques with algorithms focused on object detection is being tested, as seen in the experiments performed in [33], which have showed the need for a process capable of improving the analyzed frames. However, the currently applied models suffer from questions as the persistence of specific errors in the images, since these proposals usually seek to cover models that are specialist only in a type of enhancement. In this perspective, the present project seeks to obtain a model that is able to optimize more lighting error points in images in order to improve the performance of algorithms used to detect medical instruments during minimally invasive procedures and to improve the analysis of these imagens by surgeons and other medical professionals.

IV. Proposed Solution

A. IR-MIP

The IR-MIP (Illumination Restorer for Minimally Invasive Procedures) model, which will be presented in this section, is a neural approach based on autoencoders to improve the process of image enhancement with focus on the illumination of regions that are present in surgical images captured in the context of MIPs. The model LLNet, proposed in [34], was used as a baseline for the proposed solution. However, some adaptations of the original model were included to support color images, as their approach only supported black-and-white frames. Moreover, this model was trained with some specific datasets from MIPs, with the focus on the optimization of this approach for this context. A dataset for the training of illumination restoration algorithms based on MIPs was created with the use of surgical images provided by [35].

The use of LLNet as the base for the proposed solution was mainly because of the categories of images that were used. The collection of training and evaluation datasets include images present in PNG (Portable Network Graphics) format. The proposed network has obtained a significant result with this format of data. Some synthetic illumination noises were applied to show the efficiency of this algorithm over different levels of illumination and errors in variant regions. With the adjustment of some network hyperparameters in the hidden layers, the developed autoencoder was improved for different types of solid that were present in the human internal organs (e.g. different tones of flesh, noises caused by blood and internal liquids, bones and vascular regions).

Autoencoders have a relevant performance with this type of problem, because they are capable of abstracting a set of encodings for the data present in irregular formats, as in the case of noised images. Thus, the main task of IR-MIP in this context is to learn about how to adapt a collection of data that are available in a wrong pattern called W to a improved pattern called I. This transformation will be learned through the discovery of a function called E, which is an encoding capable to transform W into an optimized form similar to I.

The proposed model will be used in the pre-processing stage of object detection algorithms that are used by camera-control robots that operate as surgical assistants during MIPs. The camera-control pipeline involves the following stages: the first one will be described as the capture of an image by the camera; the second one involves the pre-processing stage, where some image enhancement algorithms are going to be used to improved to a specific quality pattern and our model will be inserted in a level of this stack; the third stage consists of the process of object detection, that will be executed by another algorithm; and the final stage will involve the camera orientation according to the detection of the medical instruments, as they are part of the ROI (region-of-interest) during the surgery.

The main objective of our solution involves the enhancement of the darker regions of a MIP image. Therefore, this solution can improve the execution of the object detection algorithms that are going to be used in the described pipeline, having a significant contribution as it has a better performance than some statistical models, as the one proposed in [36], that will be compared to our approach and which consists in an algorithm that estimates an illumination map on the image, generating a fixed filter that is based on a mathematical function. Some metrics are going to be used to measure the efficiency of our model and compare its performance and benchmarking, with focus on the evaluation of the level of enhancement of the proposed approach.

B. Architecture

The IR-MIP architecture uses the proposed architecture for the LLNet model as baseline [34]. This architecture consists of an autoencoder capable of performing the task of image-to-image translation on images that have low light. These images with brightness errors can be improved automatically, as models developed based on this architecture can generate an encoding that can identify which points of that image should be improved. The problem that LLNet seeks to address is with its focus on generating natural lighting in outdoor images, and this model only supports black and white images. Therefore, one of the focuses of IR-MIP was to transform this baseline model so that it could support color images, as the problem involves identifying regions present within human internal organs, and being able to act with significant accuracy. within that surgical field.

The architecture is based on a sparse denoising, or SDA, autoencoder. An autoencoder by itself will always try to extract a set of unlabelled data features to generate an encoding capable of transforming an W pattern into another pattern called I. The set of inner layers of an sparse autoencoder is larger than its input and output layers. As a consequence, the autoencoder can abstract more features from the input data with this greater number of intermediate layers. In combination with denoising algorithms, these autoencoders can generate noise in the input data. These noises are used to improve the encoding process, so this procedure does not simply create a copy of input to output, but also learns how to extract an even larger set of features from that data.

The architecture originally proposed for LLNet has an input layer capable of receiving misleading images, a set of intermediate layers capable of encoding and decoding with the focus on denoising focused on lighting and contrast adjustment. Finally, enhanced images are returned through the output layer. These enhanced images have the same dimensions as the original images. In addition, during the training and prediction process, some procedures are used to optimize both the input images and the images generated by the inference process.

A set of functions used in the training process are aimed at identifying the levels of image corruption and adjusting the learning rate. These functions act as tuning agents to optimize model learning against the analyzed dataset. After this process, a pre-training stage is performed, where these functions estimate the value of the learning rate to be used. This process consists of about 30 initial epochs and it is performed along with the use of these adjustment functions. A maximum value of 100000 training epochs/iterations has been adjusted for cases where the loss value is very variable. However, it is possible to observe training cases where a smaller number of epochs were required, since the loss value remained stable after a good number of iterations.

Unlike the originally proposed LLNet, which only works with one color pattern (i.e. black and white), the proposed model is able to analyze images available in the RGB standard. This process is performed primarily in the inference phase of the model, where the image is processed and restored. To restore, you must divide the image tensor into three arrays, each representing a color pattern, and reconstruct each of these patterns individually. This process is also performed by LLNet, but its focus is to reconstruct the black color pattern. With this, it was possible that the developed study could obtain the improvement of the color images.

C. Challenges

A number of challenges are associated with the problem that this project seeks to solve. Some projects in the object detection area have already pointed out that errors caused by lighting can significantly impair the execution of algorithms in this field. Therefore, it is observed that the problem faced has significant relevance and that the use of image enhancement algorithms may be useful for its solution. Listed below are some of the key challenges faced during the implementation of this project. These challenges are related to both the field of action and the construction of the proposed model.

Because the proposed model is based on neural methods, the process of building a dataset or data collection becomes relatively important within the project. Therefore, there is a need to build a data set capable of assisting the proposed model to converge correctly in the face of surgical images of minimally invasive procedures. This dataset should allow the model not to perform the overfitting process, converging only to the training set. In addition, it is noticeable that the use of data augmentation within the evaluated database can assist with cases where data coming from a real context has certain disparities.

Another challenge associated with this process is the correct recovery of coloration of the regions present within the enhanced image. Since images taken during MIPs are internal regions of the human body and have a set of fluids such as blood, the proposed algorithm must be able to differentiate these regions, as this directly impacts the evaluation of the image. The proposed model has support for color images, however an evaluation was performed to the point of optimizing it, identifying the different shades of red or pink and white that are present in the blood, organs, veins and other internal regions. This also involves issues related to possible cases of over saturation or too much brightness in images.

Another issue involves retrieving occluded parts and identifying different levels of placement of objects in the image. As an example, a medical instrument can positioned in a layer higher than the operated region, generating an occlusion.

This leveling is very important as it directly influences the identification of how the region is being illuminated and how the 3D environment layers are represented in a two-dimensional image, this also includes the identification of edges present in the enhanced regions.

In addition, movement issues also play a significant role in this problem, as motion blur caused by the movement of medical instruments may occlude certain operated regions. This occlusion may cause the lighting enhancement algorithm to understand that this noise is an extra layer and to perform the region enhancement differently, causing inconsistencies or generating artifacts that do not exist in the actual image. However, the proposed model should focus on identifying these corner cases and correctly improving the regions of interest.

Finally, the last most relevant problem is associated with the incidence of shadows in the image. These shadows act similarly to the problem related to occluded regions and the problem related to colorization, but this issue is more associated with the possibility of allowing the developed algorithm to identify lower layer regions that have darker tones. However, the identification of shadows also becomes important for the evaluation of the analyzed images, allowing the algorithm to recognize different shades present in a region of interest.

V. Evaluation

The process of evaluation of the IR-MIP model have consisted in a group of stages that were focused on obtaining the best parameters and the tuning of the proposed model. Firstly, a research was done with the intent to gather an image set of the MIP area that were significantly variant with the avoidance of possible cases of overfitting on our model. The collected data have passed through a process of augmentation that were composed mainly by the application of different levels of synthetic noises, rotation and changes in the axes of the ROIs. Images of laparoscopic procedures were collected from different sources to build the used dataset. Some of these images were taken directly from surgical videos through the use of some tools capable of splitting these captured videos into various frames, using free software applications like FFmpeg and ImageMagick to automate this process. A further description of the used datasets will be shown at Subsection V-A.

The training of IR-MIP was realized with a maximum number of epochs set as 100000. However, the model could balance its loss values on some lower epochs over most of the training processes. The evaluation of this quantity of epochs were done sequentially, focusing mainly on a number that were good enough that could allow the network to converge properly without causing errors due to some increase on its loss function or because of the overfitting or underfitting. In comparison to the original LLNet model, our approach has reached a slightly lower number of epochs as it is not as generalist as the base model and it is focused on a specific scope of problem. The validation were realized with around 25% of the images available in the datasets.

TABLE I
Google Colab specification.

CPU	Intel(R) Xeon(R) CPU @ 2.20GHz
GPU	Nvidia Tesla K80
RAM	16 GB
Operating System	Ubuntu
Hard Drive	320 GB

TABLE II
Octave environment specification.

CPU	Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz
GPU	Nvidia 940M
RAM	8 GB
Operating System	Ubuntu
Hard Drive	500 GB

The following metrics to evaluate enhanced images were used: Peak Signal to Noise Ratio (PSNR), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Structural Similarity Index (SSIM) [37], [38], [39]. These metrics are highly used in the field focused on the procedures of image enhancement to compute the differences between the original images and the enhanced ones. The use of these evaluation techniques in the present project will be more described in the Subsection V-B.

The main experiment of this project has consisted in the comparison of our model to another one available in the image enhancement state-of-the-art, which is known as LIME (or Low-Light Image Enhancement via Illumination Map Estimation) [36]. This method consists in the improvement of the quality of RGB images by an estimation of the light present on each pixel and the construction of an illumination map over the image. This approach is built over some statistical methods and aims to help the execution of object detection techniques with the image enhancement support on the pre-processing stages. An adaptation of the algorithm (https://github.com/cjlcarvalho/lime-octave) was developed to run it on the Octave platform, which is a free software, while the original implementation is written with Matlab, which is proprietary. The experiments that are described in Subsection V-C show the comparison of this method with IR-MIP during MIPs.

The tests for the IR-MIP were done in the Google Colab platform (https://colab.research.google.com), using GPU acceleration support. The following hardware specification described in Table I was used. As Google Colab has only native support for Python scripts, the LIME implementation was run in a different environment, which contains the hardware specification described in Table II. Both of them were evaluated with the same test data and our results show that IR-MIP has reached a better performance as it is more specialized in the MIP context than the statistical method used by LIME.

A. Datasets

Nowadays, some challenges are related to the process of collecting a sufficient amount of data for image enhancement

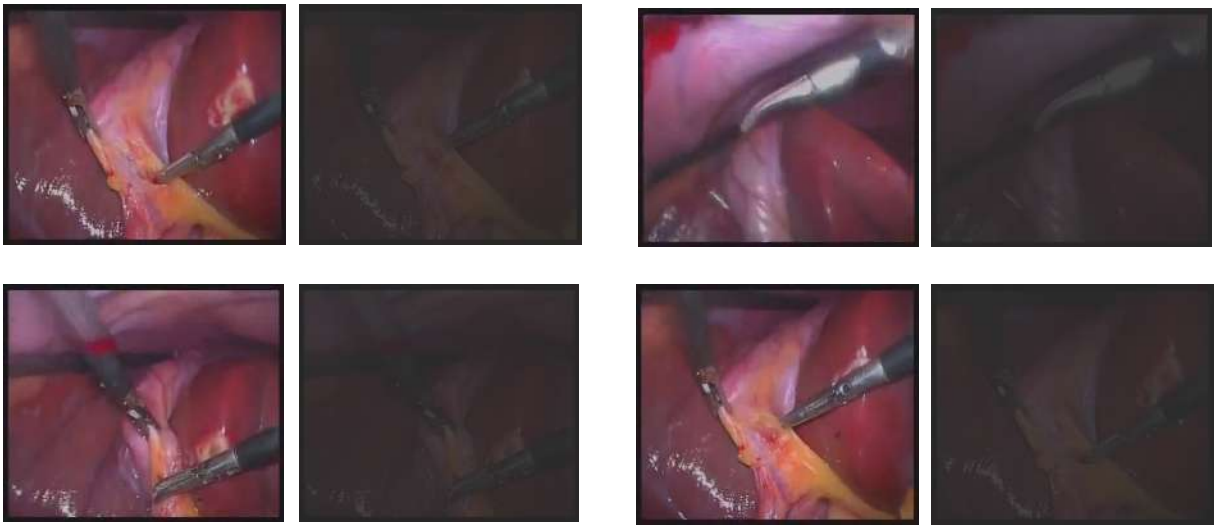


Fig. 6. Images from the processed dataset. Each example contains a sample of the original image and its version containing the synthetic lighting errors.

processes. One of these challenges is that most of the available data that are in public data sources are available in an optimized version (i.e. without any possible environmental errors caused by light, movement or other factors). Therefore, one solution for this situation is the process of generating synthetic errors in good quality datasets. As most of the minimally invasive procedures images are available for other types of computer vision problems, such as object detection and image segmentation, one of the contributions of this project was the creation of a image enhancement dataset with focus on illumination issues.

The dataset that was used in this project was built by the insertion of lighting errors into different image datasets that were originally used to solve other computer vision problems in the MIP field. The first of these datasets was a laparoscopy image sequence that was created by Sznitman et al [35]. This dataset was initially used to track laparoscopic instrument and their scales during the surgeries. The second dataset was composed by images from laparoscopic colorectal surgeries that were initially used to segment rigid medical instruments. The third dataset was focused on robotic surgeries and displays 2D poses of needle driver instruments.

Some scripts were developed to generate the synthetic errors in the images. These scripts used some external image manipulation tools and applications such as FFmpeg and ImageMagick to change the original images from the datasets and insert illumination, brightness and contrast issues into them. The focus of this process was to convert these images into analyzable samples according to the context of the problem that was faces in this project. Figure 6 shows some examples from the generated dataset, including a comparison between the original images and their outputs after the image manipulation.

However, another challenge has appeared during the implementation and tests that were done in the developed model because the images are not always with the same level of lighting. Thus, to simulate the illumination issues as they occur in a real-world context, a data augmentation was done into the generated datasets, where different levels of contrast and lighting issues were inserted into the dataset, increasing its size significantly. After this process, our model could train with a light enhancement dataset composed by surgical images, where each of them has an enhanced counterpart that allowed the autoencoder approach to generate the translation between a image in unacceptable conditions and an adequate one.

B. Metrics

This subsection shows a list containing the description for each metric used to evaluate the model and its importance on the performance benchmarking of our algorithm.

1) Mean Squared Error (MSE) [38]: Used to compute the differences between a generated image and the original one. This metric will calculate the cumulative sum of the squared errors between two images, outputting a value that is always greater than zero and that represents the distance of the resulted image from the original. Therefore, a high MSE value will represent the quantity of errors present in the output.

$$MSE(O,G) = \frac{1}{MN}\sum_{y=1}^{M}\sum_{x=1}^{N}(O(x,y) - G(x,y))^2 \qquad (4)$$

The Equation (4) describes the MSE formula, where M and N are the dimensions of the compared images, while O and G represent the original image and the enhanced image respectively.

2) Peak Signal to Noise Ratio (PSNR) [38]: PSNR is a metric used in the image enhancement field to calculate the quality of the enhanced images after the process executed by a transforming algorithm. This method will compare a signal (i.e. the original image) to a transforming noise, which represents the changes introduced by the used process. The relation between this metric and the MSE is the opposite, as a high value of PSNR represents a high quality on the resulted image.

$$PSNR(O,G) = 10 * \log_{10}(\frac{MP^2}{MSE}) \quad (5)$$

In this project, the PSNR is calculated by the usage of the formula described in Equation (5), which compares two images (i.e. the original and the enhanced), where MP is the maximum pixel value available on these images.

3) Root Mean Squared Error (RMSE) [38]: Similar to the MSE method, the RMSE will calculate the standard deviation of the prediction errors. This method shows how distant are the errors when compared to a desired output, representing this distance as a correlation between the original image and the result. Being a measure of the accuracy for image enhancement techniques, this metric will be proportional to the value of the MSE, but having a relevant use when the focus is to detect significant discrepancies on the result data, as in the case of outliers. This metric is illustrated by the Equation (6).

$$RMSE(O,G) = \sqrt{MSE(O,G)} \quad (6)$$

4) Structural Similarity Index (SSIM) [39]: SSIM is a metric to quantify the degradation caused by a transforming model to an image, being specially focused on its losses. This method has a good performance when the evaluated images are available in compressed formats (e.g. JPEG and PNG), as in the case of the images used in this project. While PSNR is focused on the noise present in the result, this method will compare the visible structures of the image, calculating the errors according to two images of a fixed size $N \times N$.

$$SSIM(O,G) = \frac{(2\mu_O\mu_G + C_1) + (2\sigma_{OG} + C_2)}{(\mu_O^2 + \mu_G^2 + C_1)(\sigma_O^2 + \sigma_G^2 + C_2)} \quad (7)$$

The Equation describes the SSIM metric formula, where μ_O and μ_G represent the mean of O and G respectively, while σ_O^2 and σ_G^2 represent the variance of O and G respectively. The C_1 and C_2 variables are two variables used to stabilize the division.

C. Experiments

The experiments that were done in this project were divided into two different stages. The first stage has involved the evaluation of the proposed method and the execution of LIME, including the estimation of the values of the metrics according to random samples from the analyzed data. Then, the second stage involved a comparison of the results of these two methods.

1) IR-MIP and LIME executions: During the training process of IR-MIP, some adjustments were done on its architecture to obtain better results according to the used image enhancement metrics. To accomplish these outcomes, tuning procedures and changes in the training dataset were done with the focus on the reduction of cases of generalization and overfitting. As a result, the proposed model could converge to the obtained results, making the images generated in the output have a high level of similarity according to the original images from the test dataset that had not been present in the training dataset.

TABLE III
METRIC RESULTS.

Metric	IR-MIP	LIME
MSE	29965.31	128260.81
PSNR	27.78	27.50
RMSE	10.40	10.75
SSIM	0.19	0.07

The LIME tests were performed based on the same data used during the IR-MIP tests. In addition, the metrics used have showed that the divergence in the the pattern generated by LIME is significant and it is possible to observe that a change in the color pattern is considerable, especially when the improved region presents different shades of blood, fluids and internal organ stains. The optimization algorithm BM3D was used in combination with LIME to balance its outputs.

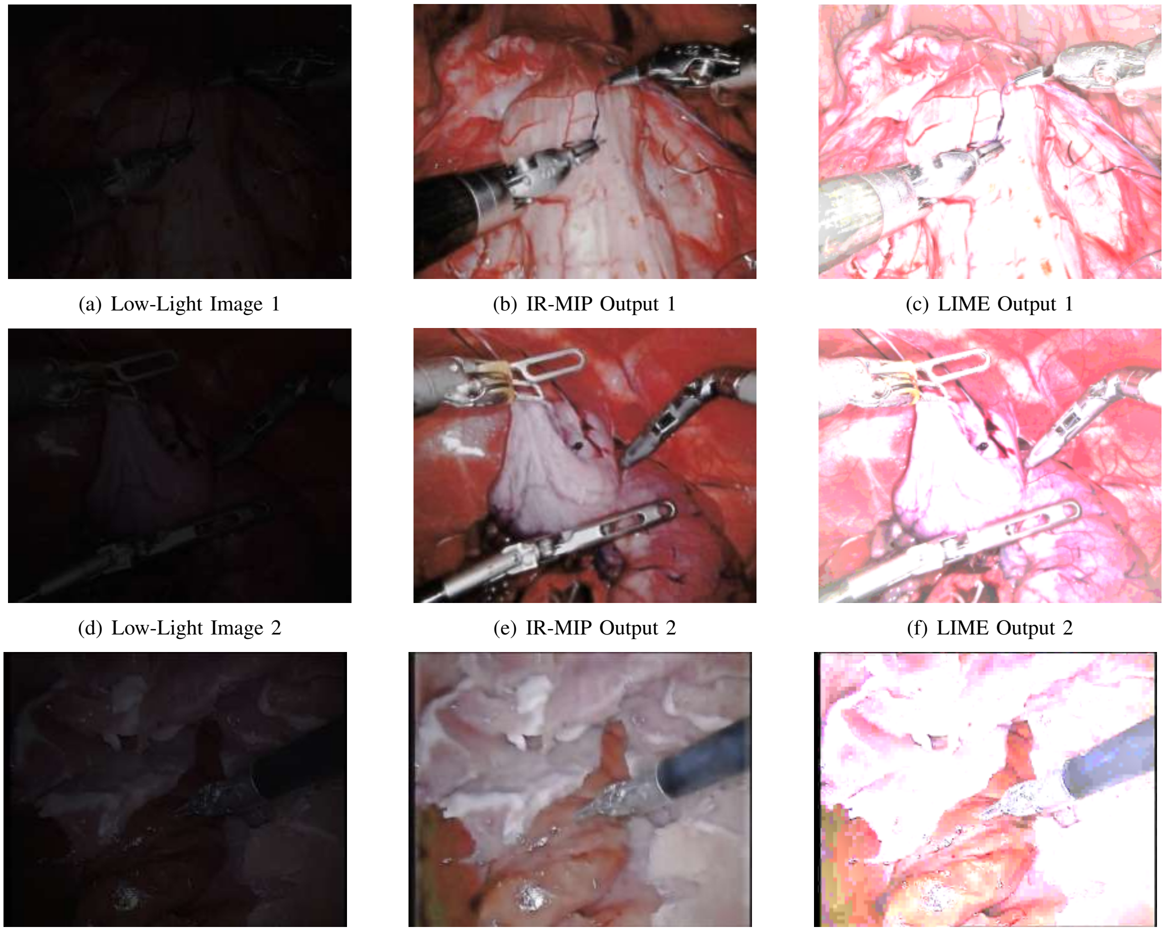
2) Metric evaluation: According to the metric evaluation process, the IR-MIP model has obtained a favorable result in comparison to the results that were obtained by the LIME model. Among its main advantages, it is possible to cite the fact that the proposed model was able to estimate the colors present in the image pixels with a high level of accuracy and this approach could also identify the variances between the various regions present in an image, including altered color tones according to occluded regions and with higher shadow levels. Table III compares the results of these two models according to the metrics, and these results were estimated based on the execution of these models in random dataset samples, showing how much LIME has generated divergences in the data that could be interpreted as inconsistencies in the images.

D. Discussion and Obtained Results

This subsection seeks to analyze and evaluate all the results obtained during the three experiments that were performed, including a reflection about the pros and cons of using the proposed model for the problem that was studied. The contributions of this project will also be debated in order to assess their impact on the state-of-the-art image enhancement area with focus on lighting problems.

As seen from the results obtained with the evaluation metrics presented in Table III, IR-MIP is a model that currently solves the low light problem in MIP images. The images in Figures 7(b), 7(e) and 7(h) show some examples of IR-MIP output, where it was possible to make the deteriorating lighting images could be significantly improved. Compared to the LIME model, a significant difference could be noted as LIME have failed to improve the light on the deteriorated image at some more intense levels, as shown in Figure 7(c).

An important challenge that was considered was related to the execution time of this model. As an approach based on neu-



(a) Low-Light Image 1

(b) IR-MIP Output 1

(c) LIME Output 1

(d) Low-Light Image 2

(e) IR-MIP Output 2

(f) LIME Output 2

(g) Low-Light Image 3

(h) IR-MIP Output 3

(i) LIME Output 3

Fig. 7. Different images before and after the process of light enhancement by IR-MIP and LIME methods.

ral methods, IR-MIP can have considerable higher execution time in a real environment than statistical methods. However, it is observed in the tests performed that its execution time was significantly compared with the other method analyzed, whose approach was based on a statistical analysis. Nevertheless, it is noticeable that an adjusted version of IR-MIP can be developed in the future with the focus of further optimizing this processing time, since MIPs are risk scenarios and the response time of the algorithms used should always be as short as possible.

The construction of the dataset used for the algorithm evaluation is a relevant contribution to the literature of the area. The built dataset has a low noise occurrence and its normalization is high. Regarding the dispersion present among the data, it was observed that this factor contributed to the execution of the proposed algorithm, since it could evaluate different types of images present in the dataset. The data augmentation process also have contributed to the fact that the IR-MIP could have a good hit rate at different light levels.

VI. Conclusion

Currently, one of the problems associated with the process of analyzing surgical images taken during MIPs is related to the lighting present in these images. Some automatic approaches based on machine learning are being developed to solve problems in this category. However, it was observed the possibility of using a model based on artificial intelligence so that these images could be preprocessed by the algorithm in order to improve them for a posteriori analysis.

The IR-MIP model, built on a deep neural approach and having an autocoder based architecture capable of handling lighting enhancement image-to-image translation processes, was able to solve the problem by presenting favorable performance in improving low-light surgical images. Through the construction of a training and testing datasets, it was possible to improve the model for the execution of this task, where it was efficient according to a set of experiments performed, including the evaluation of metrics proposed in the literature of the image enhancement area.

From the obtained results, it is possible to affirm that IR-MIP can be an efficient approach to solve this kind of problem, considering that the model was tested and compared with a state-of-the-art machine learning approach that aims to solve the same kind of problem. Therefore, neural methods based on deep learning can successfully perform the task of image enhancement in a minimally invasive surgery context.

A. Obstacles

Some obstacles had appeared during the process of development of the present project. Better results can be achieved after the resolution of these obstacles and the present model will be able to run in a real environment after that. Some of these obstacles are listed below:

- The process of training with real images could make the proposed algorithm able to be used in a real scenario. However, it was not possible to obtain public datasets with real images of surgeries where some lighting defects were present. As a result, the task of building a dataset with images containing synthetic errors became necessary.
- Deep learning algorithms tend to require significant computational power. Thus, it is noticeable that the use of platforms capable of processing these types of models becomes necessary in a real scenario.

B. Future Works

During the development of the project, some points of improvement that could be considered as future works were observed. Moreover, some details related to the implementation of the proposed model appear as project extension ideas, allowing the evolution of the present model. Among these points, the following can be listed:

- Include some Attention-based intermediate layers so that regions of interest can be identified for the lighting enhancement process (i.e. just focus on lighting enhancement where it really is needed). This work can cause a reduction in the processing time and make the algorithm even more expert in the process of recognizing regions that really need improvement.
- Combine the proposed algorithm with other image enhancement models to create an image preprocessing stack capable of handling multiple problem categories. The focus of this future implementation would be to remove smoke, fog, and motion issues, along with the improved lighting seen in the present model.
- Allow the model to be able to evaluate images available in RAW format as the present solution deals only with PNG or JPEG images. The RAW standard has a higher dimensionality than these other standards, including all pure image data without being processed by any compression algorithm.
- Use reinforcement learning by inserting a discriminator into the training process. The discriminator can identify points where the proposed algorithm can improve and enable it to adjust its learning correctly quickly.
- Perform the test of the model along with various types of algorithms aimed at pattern recognition in surgical images, such as segmentation algorithms and image object detection.

References

[1] A. W. Mariani and P. M. Pêgo-Fernandes, "Minimally invasive surgery: a concept already incorporated," *São Paulo Medical Journal*, vol. 131, no. 2, pp. 69–70, 2013.

[2] R. H. Taylor, J. Funda, B. Eldridge, S. Gomory, K. Gruben, D. LaRose, M. Talamini, L. Kavoussi, and J. Anderson, "A telerobotic assistant for laparoscopic surgery," *IEEE Engineering in Medicine and Biology Magazine*, vol. 14, no. 3, pp. 279–288, 1995.

[3] K. Omote, H. Feussner, A. Ungeheuer, K. Arbter, G.-Q. Wei, J. R. Siewert, and G. Hirzinger, "Self-guided robotic camera control for laparoscopic surgery compared with human camera control," *The American journal of surgery*, vol. 177, no. 4, pp. 321–324, 1999.

[4] M. K. Chmarra, C. Grimbergen, and J. Dankelman, "Systems for tracking minimally invasive surgical instruments," *Minimally Invasive Therapy & Allied Technologies*, vol. 16, no. 6, pp. 328–340, 2007.

[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[6] R. Maini and H. Aggarwal, "A comprehensive review of image enhancement techniques," *arXiv preprint arXiv:1003.4053*, 2010.

[7] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6306–6314.

[8] C. B. Morgenthal, W. O. Richards, B. J. Dunkin, K. A. Forde, G. Vitale, E. Lin, S. F. E. Committee *et al.*, "The role of the surgeon in the evolution of flexible endoscopy," *Surgical endoscopy*, vol. 21, no. 6, pp. 838–853, 2007.

[9] Cancer Research U.K. (2014) Diagram showing video assisted thoracoscopy (vats). [Online]. Available: http://www.cancerresearchuk.org/prod_consump/groups/cr_common/@cah/@gen/documents/image/cr_116299.jpg

[10] G. Dieberg, N. A. Smart, and N. King, "Minimally invasive cardiac surgery: a systematic review and meta-analysis," *International journal of cardiology*, vol. 223, pp. 554–560, 2016.

[11] X. Cheng, M. W. Onaitis, T. A. D'amico, and H. Chen, "Minimally invasive thoracic surgery 3.0: lessons learned from the history of lung cancer surgery," *Annals of surgery*, vol. 267, no. 1, pp. 37–38, 2018.

[12] P. A. Finlay and M. Ornstein, "Controlling the movement of a surgical laparoscope," *IEEE Engineering in Medicine and Biology Magazine*, vol. 14, no. 3, pp. 289–291, 1995.

[13] D. O. Kavanagh, P. Fitzpatrick, E. Myers, R. Kennelly, S. J. Skehan, R. G. Gibney, A. D. Hill, D. Evoy, and E. W. McDermott, "A predictive model of suitability for minimally invasive parathyroid surgery in the treatment of primary hyperthyroidism," *World journal of surgery*, vol. 36, no. 5, pp. 1175–1181, 2012.

[14] Birmingham Bowel Clinic. (2014) Da vinci robot. [Online]. Available: http://www.birminghambowelclinic.co.uk/images/uploaded/20140507140033-robot--main-image-DaVinciRobot.jpg

[15] A. R. Lanfranco, A. E. Castellanos, J. P. Desai, and W. C. Meyers, "Robotic surgery: a current perspective," *Annals of surgery*, vol. 239, no. 1, p. 14, 2004.

[16] J. C. Byrn, S. Schluender, C. M. Divino, J. Conrad, B. Gurland, E. Shlasko, and A. Szold, "Three-dimensional imaging improves surgical performance for both novice and experienced operators using the da vinci robot system," *The American Journal of Surgery*, vol. 193, no. 4, pp. 519–522, 2007.

[17] P. J. Wijsman, I. A. Broeders, H. J. Brenkman, A. Szold, A. Forgione, H. W. Schreuder, E. C. Consten, W. A. Draaisma, P. M. Verheijen, J. P. Ruurda *et al.*, "First experience with the autolap$^{\text{TM}}$ system: an image-based robotic camera steering device," *Surgical endoscopy*, vol. 32, no. 5, pp. 2560–2566, 2018.

[18] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[19] 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.

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[21] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[22] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, 2015, pp. 91–99.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[24] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[25] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu, “Automatic photo adjustment using deep neural networks,” *ACM Transactions on Graphics (TOG)*, vol. 35, no. 2, p. 11, 2016.

[26] S. Bedi and R. Khandelwal, “Various image enhancement techniques-a critical review,” *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 3, 2013.

[27] S. Bolkar, C. Wang, F. A. Cheikh, and S. Yildirim, “Deep smoke removal from minimally invasive surgery videos,” in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 3403–3407.

[28] A. Baid, A. Kotwal, R. Bhalodia, S. Merchant, and S. P. Awate, “Joint desmoking, specularity removal, and denoising of laparoscopy images via graphical models and bayesian inference,” in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 2017, pp. 732–736.

[29] O. Sidorov, C. Wang, and F. A. Cheikh, “Generative smoke removal,” *arXiv preprint arXiv:1902.00311*, 2019.

[30] C. Wang, C. Xu, C. Wang, and D. Tao, “Perceptual adversarial networks for image-to-image transformation,” *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4066–4079, 2018.

[31] X. Luo, A. J. McLeod, S. E. Pautler, C. M. Schlachta, and T. M. Peters, “Vision-based surgical field defogging,” *IEEE transactions on medical imaging*, vol. 36, no. 10, pp. 2021–2030, 2017.

[32] F. Queiroz and T. I. Ren, “Endoscopy image restoration: A study of the kernel estimation from specular highlights,” *Digital Signal Processing*, vol. 88, pp. 53–65, 2019.

[33] H. Kuang, L. Chen, F. Gu, J. Chen, L. Chan, and H. Yan, “Combining region-of-interest extraction and image enhancement for nighttime vehicle detection,” *IEEE Intelligent systems*, vol. 31, no. 3, pp. 57–65, 2016.

[34] K. G. Lore, A. Akintayo, and S. Sarkar, “Llnet: A deep autoencoder approach to natural low-light image enhancement,” *Pattern Recognition*, vol. 61, pp. 650–662, 2017.

[35] R. Sznitman, K. Ali, R. Richa, R. H. Taylor, G. D. Hager, and P. Fua, “Data-driven visual tracking in retinal microsurgery,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2012, pp. 568–575.

[36] K. G. Bhavani and T. D. Rao, “Lime: Low-light image enhancement via illumination map estimation,” 2018.

[37] C. E. Shannon, “A mathematical theory of communication,” *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[38] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, and M. Carli, “Modified image visual quality metrics for contrast change and mean shift accounting,” in *2011 11th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM)*. IEEE, 2011, pp. 305–311.

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.