Revolutionizing Computer Vision: An Unsupervised Training Methodology For Semantic Segmentation and Localization Models

Sina Saadati[1]*

Abstract

Artificial intelligence and machine learning methodologies are pivotal in image processing applications, enabling computers to learn patterns from masked images and utilize the extracted knowledge in subsequent tasks. However, image masking and annotation remain challenging and time-consuming for human experts. For the first time, this paper proposes an unsupervised deep learning method for training image semantic segmentation models, inspired by human learning psychology. First, the Self-Learning methodology is introduced, a semi-supervised approach for segmenting numerous medical images from laparoscopic or robotic surgeries videos, requiring only two manually annotated images. In the Self-Learning method, the neural network models autonomously mask additional images with high accuracy, producing fully masked images that can also train other neural networks. Next, an accurate object detection algorithm is proposed to automate the masking process using physical and geometric concepts. By integrating these methods, an unsupervised methodology for training and developing semantic segmentation neural networks is presented. Additionally, two ensemble neural network frameworks are developed to evaluate the proposed method, that can semantically segment the uterus and ovary during laparoscopic and robotic surgeries. The ensemble frameworks are developed by providing two laparoscopic videos to the computer without any masked or annotated images. Consequently, the computer learned the information and could meticulously segment the uterus and ovary in the laparoscopic videos with high accuracy, which shows that the proposed approach is reliable. The proposed methodology can be used for developing accurate, reliable, and interpretable deep learning models for various tasks and applications, both medical and non-medical, as well as for segmentation or localization tasks.

Keywords: Unsupervised Semantic Segmentation, Fully Automatic Learning, Unsupervised Localization Training, Automatic Machine Learning, Computer and Robotic vision, Revolutionizing Computer Learning.

1. Introduction

Machine learning is a critical field within computer science, artificial intelligence, and data science. It involves methods in which computers learn patterns from data, a process known as training, and utilizing the extracted knowledge to analyze new data in the future. The data in this context can include images, voice messages, texts, videos, and more. As a result, machine learning methods enable the development of intelligent applications for various purposes such as data classification, data generation, semantic

[1] Department of Computer Engineering, Amirkabir University of Technology(Tehran Polytechnic), Tehran, Iran.
sina.saadati@aut.ac.ir

segmentation, and localization [1-7]. These applications play a significant role in social and industrial processes, education, and especially in medical procedures [8-14].

Artificial Neural Networks (ANNs), a subfield of machine learning models, are inspired by the structure of the brain in humans and animals. In ANNs, neurons are mathematically simulated in software, leading to the development of neural networks. These models are computationally powerful, as they can detect the meaning of texts, generate data, classify images, generate artificial images, and enhance the quality of the movies, among other tasks. Consequently, their presence is evident in real-life applications such as autonomous driving, automated manufacturing processes, and even medical diagnosis and treatment [15-17].

Despite the pioneering role of neural network systems, the process of training a neural network model remains challenging in many areas. Specifically, in image processing and semantic segmentation, neural network models require a vast amount of annotated data, where pixels of regions of interest are meticulously distinguished from other pixels by human experts. Hence, the task of labeling and annotating the data needed for training machine learning and neural network models is highly challenging, time-consuming, and demands significant concentration from human experts [5,9,12,18].

This paper draws inspiration from the principles of human learning psychology and neurology to propose an unsupervised methodology for developing and training artificial neural network models for semantic image segmentation. This approach aims to fully automate the computer training process. As a result, the need for image masking and annotation from humans will be eliminated. With this methodology, all that is required for a computer to accurately detect and segment a component in an image is a video showcasing that element, along with a single instance of pointing to that component in only one frame of the video. The proposed method enables computers to autonomously learn patterns from the given video and subsequently detect and segment the same component in other videos. This strategy is inspired by the human learning process, where a teacher, professor, or parent introduces an object to the learner by pointing to that component. The learner then focuses on the object and observes it from different directions and angles independently, without further assistance from the trainer. This process is clearly evident in infants as they explore their environment and objects by touching, holding, and observing them during the first years of life, commonly referred to as the practicing year [19].

The proposed approach consists of two methodologies. The first method, termed 'Self-Learning,' is a semi-supervised approach for the semantic segmentation of images. This method requires only two masked images for the computer to accurately detect and segment components within a large set of images. The second method involves a physics- and geometry-informed algorithm inspired by K-nearest neighbors (KNN), which automates the image masking process. In this method, the user needs only to touch or click the component of interest in a single frame without meticulously segmenting or annotating the object. Since the physical and geometric concepts are used, this algorithm differs from simple KNN-based object detection methods. By integrating these methods, the entire process of training for semantic segmentation of images can be automated. The proposed method is evaluated using laparoscopic images, which present significant detection challenges even for humans, to train ensemble semantic segmentation frameworks. By applying the proposed method to challenging images such as laparoscopic data, an accurate and reliable evaluation is conducted. Our findings demonstrate that the proposed method is both reliable and effective.

Thus, the contribution of the paper is listed as:

- **Self-Learning Method:** A Semi-Supervised approach for training neural networks for semantic image segmentation
- **Fully-automated Masking Method:** A method that integrates the physical and geometric concepts to mask a component within an image by only clicking or touching the component by user.
- **Unsupervised Methodology for Semantic Segmentation Models:** For the first time, an unsupervised method for developing neural networks for semantically segmenting and localization the objects and components, especially within the human body, is proposed.
- **Development of an Ensemble Neural Network Framework:** A framework is developed that is able to detect, segment, and localize human organs, such as ovary and uterine during a laparoscopic or robotic surgeries.

In Section 2, a literature review is presented. The proposed method is detailed in Section 3. In Section 4, the proposed method is evaluated using laparoscopic and robotic surgery images. The experiments demonstrate that the proposed method is reliable and results in accurate machine learning models. Finally, Section 5 provides a discussion of the results and outlines potential future research directions.

2. Literature Review

Semantic segmentation is a technique that enables computers to extract visual patterns from images and video signals to detect components within them by differentiating the pixels of the target component from non-relevant pixels. This method is widely used in various applications such as medical image processing, transportation, video games, and augmented reality. However, developing a semantic segmentation model using artificial neural networks requires a large amount of manually annotated images, making the process both challenging and time-consuming. Despite the fact they there are many user-friendly tools, Image annotation demands a high level of concentration and focus from humans, further complicating the model training process [18,20]. This paper proposes a reliable and interpretable unsupervised method for developing semantic segmentation models, addressing these challenges effectively.

ANNs are powerful models in computer science that can effectively learn patterns in data and apply them to various applications. They are widely used in medical and clinical procedures. Based on their robust detection abilities, a method to fully automate surgical operations using artificial intelligence-based systems is proposed in [12]. However, the proposed method require image annotation, a challenging and time-consuming task for humans and experts. Therefore, a fully automatic approach is needed for the development of reliable machine learning-based models in medical applications, particularly for semantic segmentation tasks. Medical and clinical applications can achieve greater reliability and interpretability if more data are utilized in their training process.

Some research has proposed semi-supervised semantic segmentation methods for detecting surgical instruments during operations [21-23]. The primary limitation of these methods lies in the inherent nature of semi-supervised learning, which still requires human intervention, thus posing significant challenges. Moreover, these methods are restricted to a very narrow scope, specifically for detecting surgical instruments. While segmentation of surgical instruments is undoubtedly important, it can be achieved with greater accuracy without using machine-learning methods. For instance, in robotic surgeries, the robot is aware of the exact position of each instrument. Consequently, the semantic segmentation of the instruments can be accomplished even without any camera, by geometrically calculating the precise position of the instruments and its geometric shape in relation to any point in space, such as a camera or the human eye. In this paper, an unsupervised semantic segmentation method is proposed that can be generalized for use in a wide range of tasks and applications, including various medical fields, transportation, biomedical engineering, biomechanical engineering, aerospace engineering, environmental science and engineering, face detection, and even social tasks such as fashion. This methodology plays a vital role in the development of computer vision and robotic vision tools and methods. Therefore, the proposed approach represents a significant advancement in the field of deep learning and computer vision learning.

Several studies have claimed to present unsupervised methods for semantic segmentation [24-26]. However, these methods have significant limitations. For instance, the approach proposed in [24] is designed exclusively for synthetic aperture radar (SAR) imagery, which contains easily extractable patterns. For analyzing SAR images, computational geometry-based algorithmic methods can outperform learning-based approaches, as these patterns often resemble geometric shapes such as lines and rectangles and can thus be detected more easily.

The method proposed in [25], being semi-supervised, still requires labeled data and is limited to images with orderly shapes. This task can be efficiently performed by employing a KNN-based algorithm to cluster pixels based on their color values. Similarly, in [26], the ImageNet-S dataset that contains masked images is provided. The method still necessitates a large number of annotated images. For example, 9,000 annotated images are used during training and 40,000 annotated images during the evaluation step in [26]. Consequently, these studies have not effectively addressed the challenge of training semantic segmentation models in an unsupervised manner.

Unsupervised learning refers to methods in which computers independently discover complex patterns in the data without any annotated data. Therefore, in the field of semantic segmentation, images must not be annotated or masked. From a learning perspective, the study in [24] can be considered a calculation-based method rather than a true learning method. For example, detecting moving objects which is used for transportation applications can be developed only by differentiating changing and unchanging pixels in sequential images of videos or real-time camera which is calculation-based and algorithmic method rather than learning-based method. Thus, the problem of unsupervised semantic segmentation remains an open challenge in computer vision, which is addressed for the first time in this paper.

Learning-based methodologies are more effective because one strategy can be applied to a wide range of applications and tasks. In contrast, calculation-based methods require considerable effort from human experts to redefine concepts, reconfigure metrics, and investigate the possibility and adaptability of updating the overall system whenever there is a small change or update in the program's purpose. Therefore, there is an important risk that many important ideas might be simply left because the conceptualization and development strategy is considerably time-consuming for humans. These challenges are eliminated, or at least significantly minimized, in learning-based methods. Thus, an adaptive unsupervised semantic segmentation training method is proposed in this paper to maximize the efficacy of computerization processes in various areas.

While previous studies [24-26] have focused on patterned images containing easy-to-detect components, this research utilizes laparoscopic and robotic surgery images, which present considerably more complex patterns even for human observers. The edges and borders in these images are significantly ambiguous due to the compact arrangement of organs within the body. Additionally, the similar and dark colors of different organs, closed spaces, and undetectable shadows make it challenging to distinguish the components and organs.

In this paper, the proposed methodology for unsupervised learning of semantic segmentation models has been applied to laparoscopic images. The findings demonstrate that this methodology is both effective and reliable. Despite the absence of masked or annotated images in the process, the trained models successfully detected and segmented organs such as the uterus and ovary in laparoscopic and robotic surgery images.

Since the proposed methodology in this paper has passed strict evaluation strategies on laparoscopic data, it can be generalized to various tasks and applications with different objectives and can be utilized in a wide range of procedures, both medical and non-medical. From a transfer learning perspective, the models used in this paper were pre-trained on the ImageNet dataset solely to accelerate the process, given the limited computational resources available for this research. Nonetheless, our findings indicate that the proposed methodology remains effective even without the use of pre-trained models, with randomly assigned weights. Thus, the proposed methodology in this paper can be conducted in a fully-unsupervised learning manner without any annotated or masked images despite some researches such as [26].

3. Proposed Methodology

The purpose of this paper is to propose a learning methodology for computers to learn from visual data in a manner analogous to how a human child learns from their mother or how a medical student learns when their professor introduces human organs solely by pointing to them. It is therefore crucial to investigate the psychology and neurology of human learning. Subsequently, the process of human learning can be computerized by integrating machine learning methods and deep neural networks. From the perspective of human neurology, when a student or infant intends to learn what an object is, a hint is initially provided by the mother or professor. The individual then examines the object, grasps it, and rotates it to view it from different orientations. During this process, the individual integrates visual signals with an understanding of physical and geometric properties. For instance, humans have an innate understanding that objects generally possess a continuous and coherent structure. From a geometric perspective, humans recognize that most objects in real life have smooth, rather than jagged, surfaces. These physical and geometric concepts, when combined with visual signals from the eyes, enable the child or student to learn to identify the object and distinguish it from other entities.

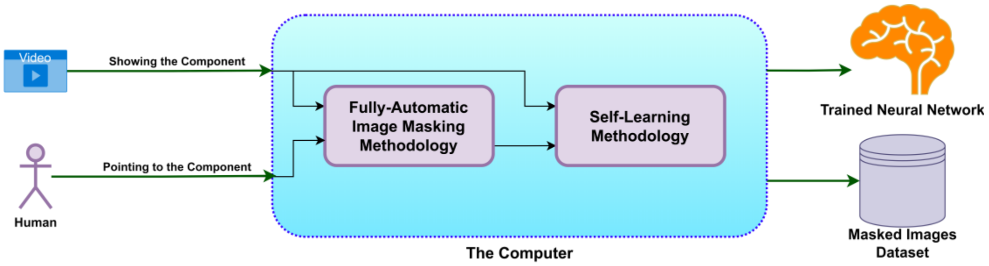
Video
Showing the Component
Human
Pointing to the Component
Fully-Automatic Image Masking Methodology
Self-Learning Methodology
The Computer
Trained Neural Network
Masked Images Dataset

Figure 1, Overall strategy of the proposed method.

This paper proposes a learning method where a video is provided to a computer, and a user points to the component of interest in the first frame of the video without annotating or masking the component. The computer then independently learns to detect and semantically segment the component of interest in all subsequent frames, as well as in other videos. The proposed approach is illustrated in Figure 1. Initially, Self-Learning, a semi-supervised approach for training neural networks for semantic segmentation is introduced in Section 3.1. In the Self-Learning method, the computer can independently learn how to detect, and semantically segment a component of interest in the images using only two masked images. Following this, a fully automated object detection approach is proposed in Section 3.2, which integrates physical and geometric concepts to automatically mask an image. Finally, by combining these two methods, the entire process of training for semantic segmentation can be conducted in an unsupervised manner. Section 3.3 presents a computational approach using distributed computing to enhance the proposed method, thereby developing more powerful intelligent frameworks more efficiently.

3.1. Self-Learning Methodology

In this section, the concept of Self-Learning is detailed. Self-Learning is a semi-supervised method that leverages convolutional neural networks (CNNs) and uses only two masked images from a video to enable the computer to detect and segment the component of interest across all frames of the video. The video is assumed to display the component of interest from various orientations and positions, inspired by an infant observing an object while rotating it.

The Self-Learning method begins with a large number of sorted images extracted from a video, with only two frames at the beginning of the video masked by a human expert. These two masked images are referred to as *LearningFrame1* and *LearningFrame2*. These frames should be close to each other in terms of time. In this paper, it is suggested that *LearningFrame1* and *LearningFrame2* be the first and tenth frames of the video, respectively. The CNN is then trained exclusively on these masked data. The resulting CNN model can then detect the component of interest in images that are very similar to the masked images. However, the resulting model is flexible to very small camera movements as different frames are provided to train the model ($LearningFrame1 \neq LearningFrame2$). Consequently, the model can segment all frames that are, in terms of time, between *LearningFrame1* and *LearningFrame2* because these unmasked frames are similar to these two masked frames (*LearningFrame1* and *LearningFrame2*), and the motion of the component of interest in the images is as minimal as is within the model's flexibility.

In the second step, all segmented frames are added to the training set, and the training process is repeated. At the end of the second step, the model's ability to detect the component of interest is enhanced, and its flexibility to the motion of the component of interest is improved. Thus, the model can accurately segment the frames that are, in terms of time, near and after Learning_Frame_2. For example, ten subsequent frames after Learning_Frame_2 can be meticulously segmented by the model. As a result, ten new frames can be automatically masked and added to the training set. Our findings show that repeating this strategy enables CNN to mask all of the frames accurately and also detect and segment the components of interest in other videos. The overall Self-Learning methodology is demonstrated in Figure 2. The initial step, executed only once at the beginning of the self-learning process, is shown on the left. The middle section displays the iterative learning process, which is executed multiple times. The final results after the self-learning method are depicted on the right.

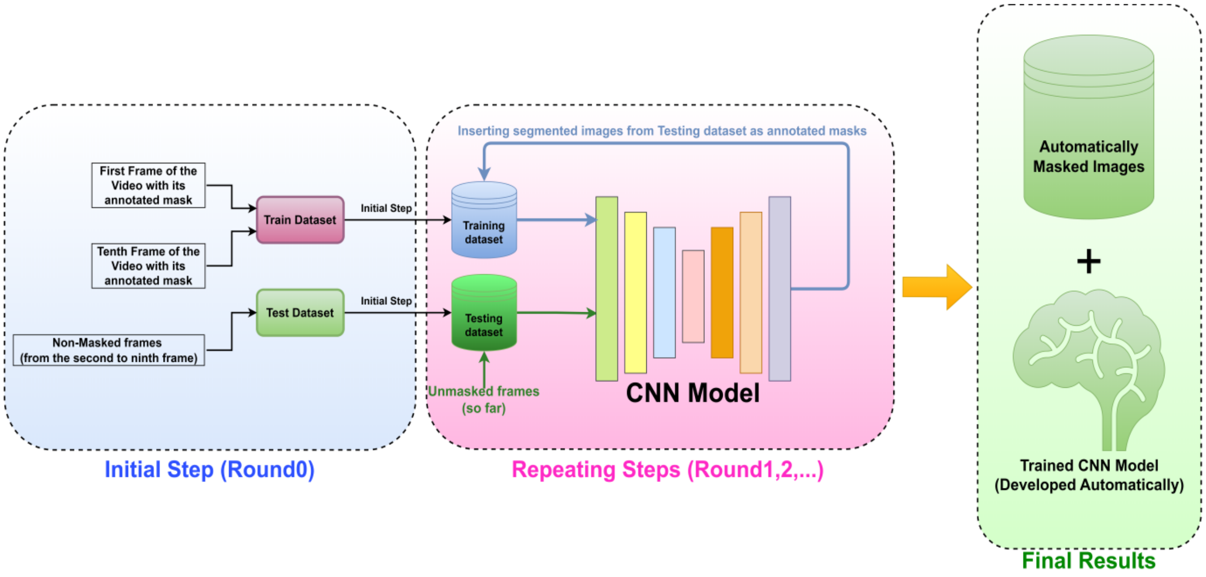
First Frame of the Video with its annotated mask
Tenth Frame of the Video with its annotated mask
Non-Masked frames (from the second to ninth frame)
Train Dataset
Test Dataset
Initial Step
Initial Step
Initial Step (Round0)
Inserting segmented images from Testing dataset as annotated masks
Training dataset
Testing dataset
Unmasked frames (so far)
CNN Model
Repeating Steps (Round1,2,...)
Automatically Masked Images
+
Trained CNN Model (Developed Automatically)
Final Results

Figure 2, Overall design of the proposed Self-Learning methodology for semantic segmentation.

In the context of Self-Learning, several concepts require definition. A non-interrupted video is conceptualized as a book containing continuous pages, with each page representing one frame of the video. Thus, P_t is the frame of the video represented at time t. A window (W_e^s) is defined as a set of sequential pages starting from page achieved at the time s of the video (P_s) and ending with the page achieved at the time e of the video (P_e). It is important that the pages in a window be extracted from sequential frames of the video. Size of window is defined as the number of pages in the window. In this paper, two windows are defined including Learning Window (LW_e^s) and Testing Window (TW_e^s) which are utilized as training dataset for training the model and as test dataset for evaluation of the model, respectively .

Learning Window and Testing Window are not constant during the proposed methodology. The Learning Window starts with 2 pages but is updated by inserting newer masked pages during the process. The act of transferring newly segmented images as masked pages of Testing Window to the Learning Window is defined as Injection and can be described as Eq.1in which $Round_i$ refers to the process where the CNN is trained and evaluated on the learning window, and subsequently used to segment the pages of Testing Window. During the proposed method, one CNN model should be repeatedly trained in different rounds in order to get more flexible to the camera motions.

$$Inject_i\!: \; LW\ (Round_{t+1}) = LW\ (Round_t) + TW\ (Round_t) = LW_{le}^{ls} + TW_{te}^{ts} = LW_{te}^{ls} \qquad \text{Eq.1}$$

Denoise is an important module of the proposed methodology that integrates the physical and geometric concepts to handle unexpected errors in the masked pages of Testing Window. This function is inspired by the ability of human to realize the physical concept of objects which helps them to better learning. Since most of the objects in the real world, especially in human and animal bodies, have smooth surface, the predicted pages by the CNN model should be blurred using Gaussian Blur function. Also, since the particles of any object in the world are connected to each other, objects have a coherent structure. Thus, if there is a small area belonging to a label like *L1* surrounded by another wider area that belongs to another label like *L2*, the smaller area should be detected as noise and its label should be replaced by *L2*. This calculation removes big unpredicted noises. As the result, the function of *Denoise* module can be described as Eq.2. In this paper, only two labels are defined: inside the component of interest and outside the component of interest.

$$Denoise(P_t) = GaussianBlur(P_t) + \ Removal\ of\ small\ sorounding\ subareas(P_t) \qquad \text{Eq.2}$$

The calculations described in Eq.2 is detailed in the pseudocode 1. In this algorithm, first the image is denoised by applying GaussianBlur function (Line 1). This function will remove small noises in the image. Then, for all of the black areas with small surface which are surrounded by another area with the white color, the color of the inner area is changed to white (Lines 2-4). In this concept, black and white color represent the negative and positive labels predicted by the neural network. In some cases, there might be some white noises. Thus, this algorithm should be also applied for white small areas which are surrounded by black areas.

Pseudocode 1, The proposed algorithm for denoising images using physical and geometric information.

```
    Denoise Algorithm (noisyImage):
    Input: The RAW prediction of the neural network
    Output: Denoised image using physical and geometric concepts

1   noisyImage = GaussianBlug( noisyImage ) ;
2   **Foreach** area **in** noisyImage **do** {
3           **If** (area→color == Black **AND** area is surrounded from
                more than 6 directions **AND** area→surface is small) **do** {
4                   area→color = White ;
            }
4   }
5   **Return** noisyImage ;
```

Since the flexibility of the CNN model grows step-by-step, It is necessary that Learning Window and Testing Window be completely neighbored with each other as is described by Eq.3. Therefore, the CNN model can perfectly mask the pages belonging to the Testing Window.

$$|\ ts - le\ | = 1\ or\ |\ te - ls\ | = 1 \quad if\ LW\ =\ LW_{le}^{ls}\ and\ TW\ = TW_{te}^{ts} \qquad \text{Eq.3}$$

The Self-Learning method can now be described using the defined annotations. In the initializing round ($Round_0$) ten sequential pages from the beginning of the videos should be selected. The first and tenth images ($P_1\ and\ P_{10}$) should now be masked by human expert and be considered as LW. This is the exceptional round in which Learning Window contain non-sequential pages. Then, the CNN model should be trained using LW as the training set. Due to the limited size of the training set, the training process in $Round_0$ is significantly time-consuming. After the model is trained perfectly on the LW, the pages inside TW can be masked by the trained model. Since the LW and TW are in the neighborhood of each other, the content of the pages inside them are as similar as the CNN model can perfectly segment the pages of TW. Finally, LW can be updated and improved by moving the segmented pages from TW to LW. Consequently, the $Round_0$ can be defined as Eq.4.

$$Round_0 = Training(\ LW_{10}^{1}\ -\ \sum_2^9 P_t\ ) +\ Predict(\ LW_9^2\ ) \quad \text{Eq.4}$$

In the cases where there is an intense change in the pages, decreasing the size of windows can be a solution. Decreasing the size of windows cause increase in the similarity of pages and decrease in the flexibility needed by the CNN model. This point is important when the whole process need to be executed with minimal possibility of supervision.

After the initializing round($Round_0$), 10 images with their masks are available which can be used for re-training the model. Thus, LW_{10}^{1} and TW_{20}^{11} can be used as the training set and testing set, respectively. As the result, other rounds can be described as Eq.5. In Eq. 5, 10 pages are used as the test dataset for segmentation and masking at each round. However, our findings indicate that this number grows exponentially as the number of rounds increases. This leads to a rapid growth in the knowledge, flexibility, and power of the CNN model used as the training system in the proposed methodology.

$$Round_i = Training(\, LW_{10\,i}^{1}\,) + \; Predict(\, LW_{10\,i+10}^{10\,i+1}\,) \quad \text{Eq.5}$$

The phases of each learning round are demonstrated in Figure 3. As shown, the neural networks begin with training the learning window. At this step, the pages that belong to the learning window provide knowledge to the network, making it more flexible, powerful, and accurate in detecting and segmenting the component of interest. Consequently, the enhanced model can reliably segment the pages of the testing window. The segmented images are then denoised using physical and geometric concepts. Finally, the new masked data are injected into the learning window. This strategy simulates the psychology of the learning process in the humans. Consequently, the proposed method is not only reliable but also interpretable.

Figure 4 illustrates the Learning and Testing Windows during the first four rounds of the proposed methodology. The content of the windows changes with each round. As the flexibility and detection power of the model increase with the size of the Learning Window, the size of the Testing Window can also be increased, accelerating the process of masking all the pages. The learning rounds should continue until all available pages are masked. If there is a memory limitation, the size of the Learning Window can remain constant by inserting new masked pages at the tail of the window and removing old pages from the head. However, in this situation, the number of testing windows should be controlled carefully.

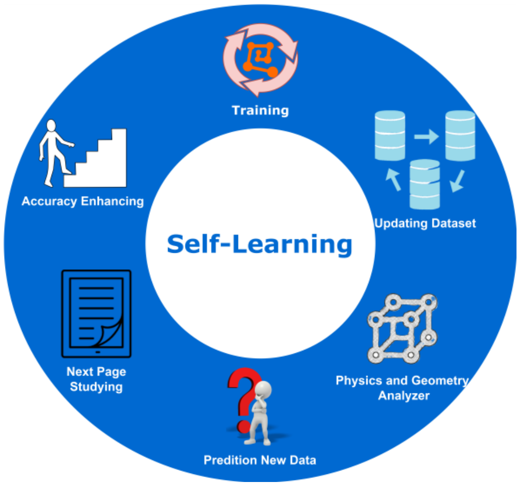

Figure 3, The overall methodology for each learning round in the Self-Learning methodology.

In the proposed Self-Learning method, by masking only two images of a video, all the images are accurately and automatically masked. Additionally, the CNN model used in this methodology is now powerful enough to detect and segment all the frames of the video. When the component of interest is shown from different orientations and positions, the CNN model is not only able to detect the component in the provided video but also in other related videos.

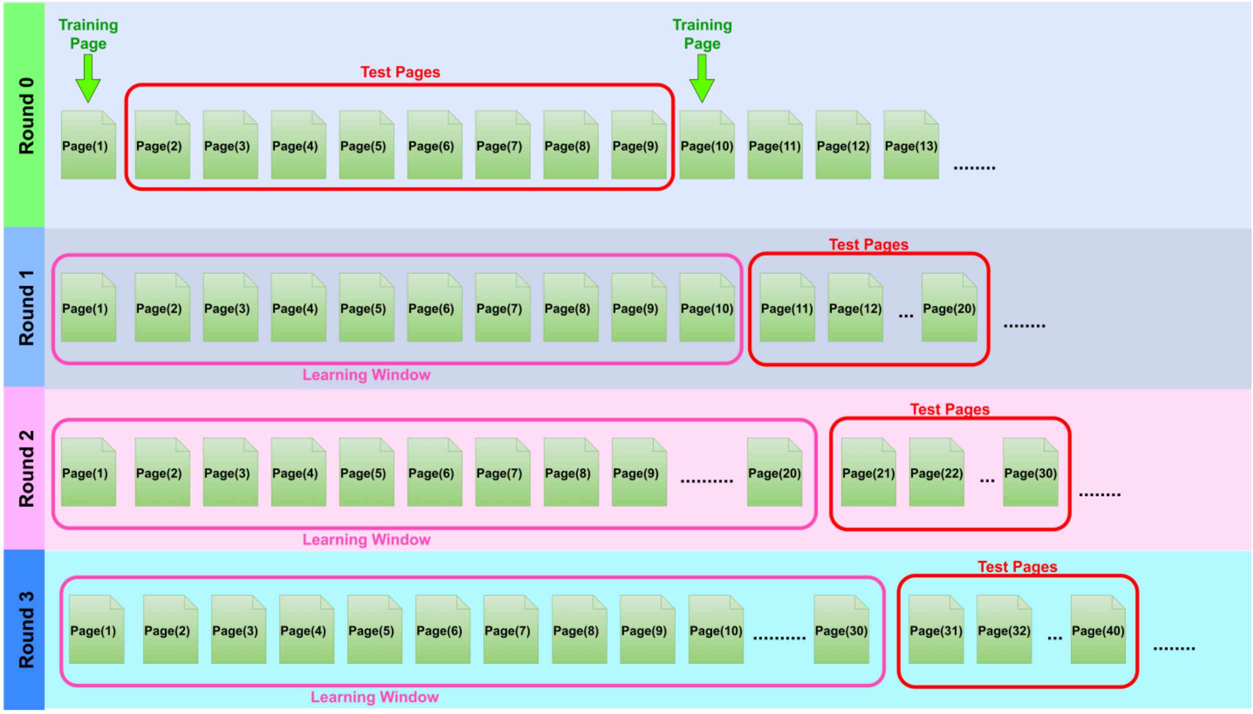

Figure 4, The distribution of pages for each round of learning.

In this research, laparoscopic images are used to evaluate the proposed method. Our findings, detailed in Section 4, demonstrate that the proposed methodology is effective and reliable. Since laparoscopic images are challenging to interpret even for humans, it is expected that the proposed method can be generalized to other applications where the images have sharper edges and are easier to detect and learn.

3.2. Fully-Automatic Image Masking Methodology

Using the proposed semi-supervised method for the development of semantic segmentation frameworks can significantly facilitate training and data engineering tasks for human experts. However, since two or a few images must be masked in the proposed method, the approach presented in Section 3.1 is still considered semi-supervised. In this section, the psychology and neurology of human learning are inspired to propose a fully automated masking application. In real life, humans do not segment components to introduce them to others. Instead, humans simply point to the components of interest, and the human learner distinguishes the component from other objects independently.

The detection capabilities of humans and animals are not limited to vision alone. They integrate physical concepts such as edge detection, coherent object structures, color similarity in object particles, and even the three-dimensional geometry of components. In many living creatures, including humans, having a pair of eyes enables three-dimensional visualization of the world. This ability, combined with other senses such as touch and proprioception, allows humans to learn about objects and components simply by being pointed out by a mother, teacher, professor, or any other trainer.

Therefore, the process of object detection can be computerized by integrating physical and geometric concepts into visual signals. This will enable us to fully automate the masking of components within an image without the need for meticulous segmentation, annotation or masking. As a result, the initial image masking step in the proposed Self-Learning method can be automated, allowing the training of semantic segmentation models to be performed entirely in an unsupervised manner. In this section, a methodology for automating the masking of components of interest in an image is proposed.

To automatically mask an image, the human user is first asked to point to the component of interest within the image. They are also asked to point to locations outside the component of interest. This delineates coordinates inside and outside the component of interest. Using these defined coordinates, the center position of the component of interest can be estimated. Additionally, the red, green, and blue color values of the pixels at these coordinates can be extracted from the image. Using this information, the K-Nearest Neighbor (KNN) algorithm can be applied to classify all pixels as either inside or outside the component of interest.

Following the KNN algorithm, all pixels of the image are compared with labeled pixels. To achieve this, a distance function is defined to measure the distance between any two pixels. Unlabeled pixels can then be classified based on the category of their nearest samples. In this paper, color values are used as the metric for distance measurement, as colors are one of the most important factors in object detection.

Geometric features are also important in object detection. When pixels are classified based on their color values, some geometric computations, such as edge detection, are implicitly performed. However, it is possible to add more geometric details to the KNN algorithm. The physical distance between the coordinates of the pixels can serve as an additional feature for the KNN algorithm. Therefore, the distance between each pixel and the center position of the component of interest should be calculated and provided as an additional value for each pixel. Consequently, the distance between any pair of pixels located at coordinates *(x, y)* with color value *(r, g, b)* can be calculated using Eq. 6, where (x_c, y_c) represents the coordinates of the estimated center of the component of interest.

$$SinaDistance(P_A, P_B) = \sqrt{0.3\ \times DistanceFromCenter + (r_A - r_B)^2 + (g_A - g_B)^2 + (b_A - b_B)^2} \quad \text{Eq.6}$$

The $DistanceFromCenter$ is calculated using Eq. 7. Our findings show that this feature significantly enhances the accuracy and performance of the algorithm compared to other metrics, such as comparing the distance along each axis independently. Since most real-life objects, especially human organs like the uterus and ovaries, have a round shape, the direct distance value performs better than other distance measurements, such as Manhattan or Chebyshev metrics. However, the distance metric is considered a secondary informative feature because color information is more informative than the distance from the center of the component of interest. Therefore, it is weighted at 0.3 in Eq. 6.

$$DistanceFromCenter(P_A) = \sqrt{(x_A - x_c)^2 + (y_A - y_c)^2} \quad \text{Eq.7}$$

Using the proposed Distance metric in Eq. 6, all pixels in an image can be classified to determine whether they are located inside or outside of the component of interest. As the method meticulously considers the details of the pixels, the computer-generated output is more accurate than human-masked images. However, there is still a risk of noise in the output. For instance, the similarity of colors in human organs can cause erroneous detections. Our findings show that these noises can be effectively eliminated by incorporating additional physical and geometric concepts. Consequently, the output of the proposed KNN-based algorithm should be denoised using our proposed denoising function in Eq. 2. Finally, the automated masking method is described in Eq. 8.

$$AutomaticMasking(P_t) = Denoise\big(KNN_{SinaDistance}(P_t)\big) \quad \text{Eq.8}$$

Since many objects and components in the real world have a convex shape, the proposed methodology for automating the masking and annotating of components of interest within images can be improved by integrating computational geometry-based concepts with the proposed KNN-based algorithm. In this research, it is suggested that after calculating the mask of the component of interest using Eq. 8, the convex hull of the masked image be computed. The convex hull represents the smallest convex shape that can enclose the component. However, this technique should be used carefully. For example, in the proposed ensemble framework for semantic segmentation of the ovary, detailed in Section 4.1, calculating the convex hull of the denoised masks significantly enhances the efficacy of the process. However, because another organ was located on the uterus within the laparoscopic images in this research, the convex hull technique could not be applied to the proposed ensemble framework for uterus segmentation. Nevertheless, the overall method remains reliable for both ovary and uterus segmentation frameworks. The automatically masking the component of interest by calculating the convex hull of the positive-labeled pixels is defined in Eq.9. The algorithm for solving the convex hull problem which is effective in this method is proposed by [27].

$$AutomaticMasking(P_t) = Convex\ Hull\left(Denoise\left(KNN_{SinaDistance}(P_t)\right)\right) \quad \text{Eq.9}$$

The fully automatic masking methodology presented in this section can significantly facilitate data engineering processes and, consequently, promote image processing procedures. The proposed method eliminates the need for meticulous annotation and masking of components in images and simulates the process of visual conceptualizing in humans. Thus, it can be integrated with the proposed Self-Learning methodology to train a semantic segmentation model in a completely unsupervised manner. The results of the methods proposed in this paper include not only masked images and semantic segmentation frameworks but also the potential for training classification, generation, and in-depth analysis. Consequently, the rate of the machine learning training process can be significantly accelerated, resulting in more intelligent systems in real-world applications.

3.3. Integrating Distributed Computing for Faster Computations

In this section, the concept of distributed computing is integrated with the proposed method, wherein several virtual machines running on local computers or cloud-based virtual machines can perform the proposed computations in parallel. Since there is no need to mask images manually, each video gathered from a patient can be analyzed in parallel by a virtual machine. Consequently, each virtual machine masks the images within its dataset, allowing a distributed computing system to mask all the images. This approach generates a rich dataset with minimal effort from human experts.

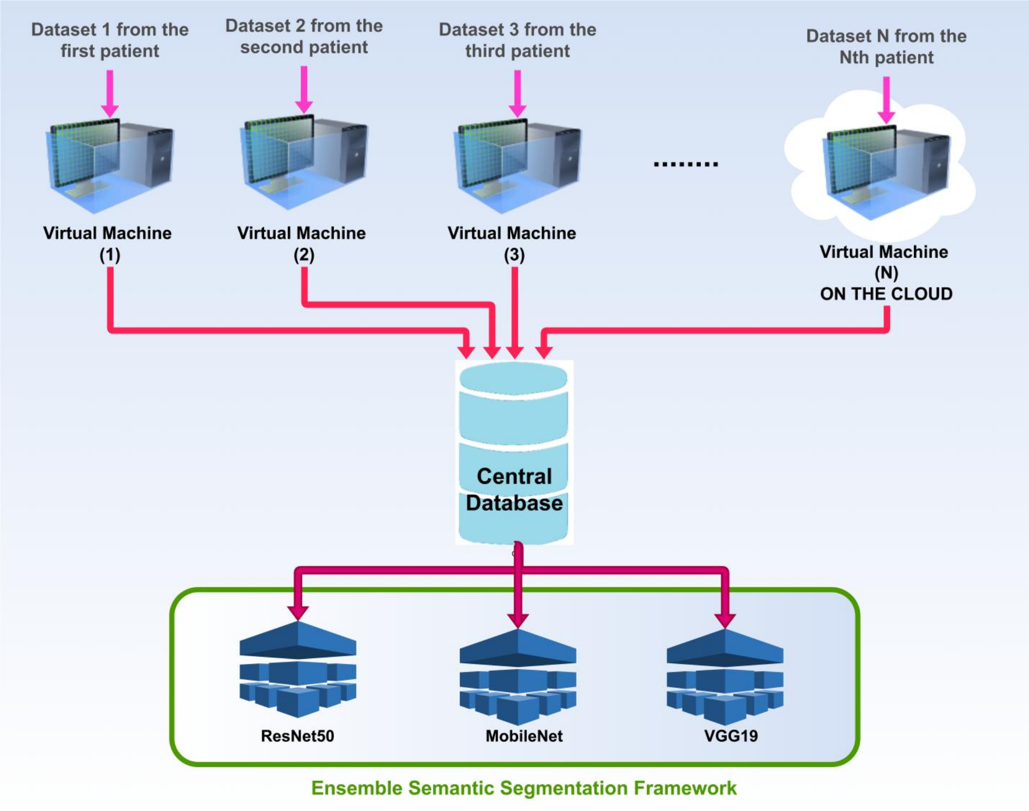
Dataset 1 from the first patient
Dataset 2 from the second patient
Dataset 3 from the third patient
Dataset N from the Nth patient
Virtual Machine (1)
Virtual Machine (2)
Virtual Machine (3)
....
Virtual Machine (N) ON THE CLOUD
Central Database
ResNet50
MobileNet
VGG19
Ensemble Semantic Segmentation Framework

Figure 5, The proposed distributed architecture utilizing the unsupervised learning methodology.

Additionally, the neural networks used in each virtual machine can be retrained in parallel using the resulting dataset, leading to the development of a powerful and interpretable ensemble framework. This capability is feasible because different architectures can be utilized in the proposed method. For instance, in this research, two ensemble frameworks were developed following the proposed unsupervised methodology to semantically detect and segment the uterus and ovaries during laparoscopic and robotic surgery. By employing different architectures, unexpected errors are minimized, while reliability, accuracy, and interpretability are maximized [11]. The proposed computing strategy is demonstrated in Figure 5.

4. Evaluation and Findings

In this section, the proposed unsupervised method for training semantic segmentation neural networks is evaluated. Two experiments have been designed to teach the computer how to semantically segment human inner organs, such as the ovary and uterus, during laparoscopic and robotic surgeries. The choice of medical image processing for evaluating the proposed method stems from the challenges it presents, which surpass those in many other fields. Human inner organs have similar colors, complex vessel structures, and intricate edges, making object detection in these images extremely difficult, even for advanced image editing software like Adobe Photoshop.

If the proposed unsupervised application proves technically feasible for these experiments, it can be expected to work across a wide range of applications. For instance, detecting objects such as birds or airplanes in the sky or humans and cars in the streets is easier due to their sharper edges and more distinct color differences from their surroundings, facilitating their detection and differentiation.

The experiments designed to develop a semantic segmentation framework for detecting the uterus and ovaries during laparoscopic and robotic surgeries are medically and surgically valuable. These experiments pave the way for effectively utilizing medical and surgical data, such as that from laparoscopic or robotic surgeries, in machine learning and computer science research and development. Many clinical centers conduct multiple robotic or laparoscopic surgeries daily, generating vast amounts of data that often remain unused. A significant reason for this phenomenon is the considerable time, energy, and concentration required from human experts to mask surgical images and videos.

By implementing the proposed unsupervised methodology for developing semantic segmentation models, the knowledge gained from each laparoscopic and robotic surgery can be easily saved, extracted by computers, and compiled. This will result in more reliable medical applications, as they will be trained on a larger dataset. As explained in Section 3.3, knowledge from different clinical centers can be integrated by developing a distributed database and training all stored data from various hospitals on a single machine learning framework. Consequently, the experiences and knowledge from each surgery will not be lost but will be readily available for use by computers. This will exponentially enhance the effectiveness of fully automated surgeries performed by computers and artificial intelligence models, as introduced by [12].

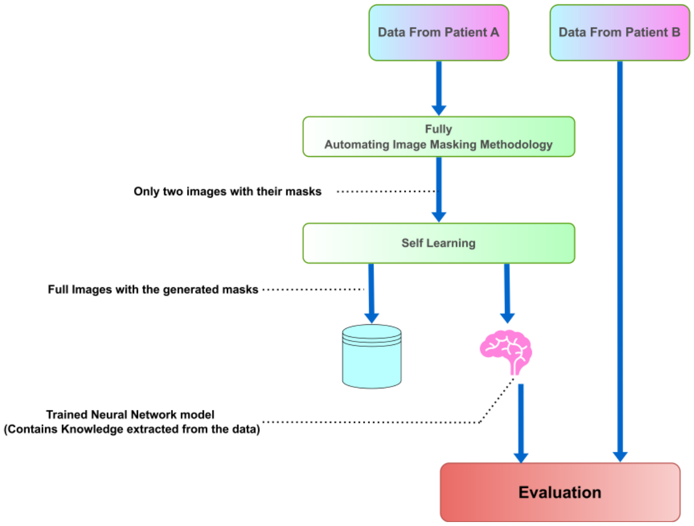
Data From Patient A
Data From Patient B
Fully
Automating Image Masking Methodology
Only two images with their masks
Self Learning
Full Images with the generated masks
Trained Neural Network model
(Contains Knowledge extracted from the data)
Evaluation

Figure 6, The evaluation procedure for the proposed unsupervised semantic segmentation methodology.

In this section, laparoscopic data from two patients are utilized in each experiment. Initially, the proposed automatic masking technique is applied, automatically masking the component of interest (ovary in the first experiment and uterus in the second) in two images per experiment. Next, using the proposed self-learning methodology, all images for the patient in the experiment are masked, leveraging only two masked images from the proposed KNN-based method. Finally, the deep learning model used in the self-learning process for each experiment is employed to detect and segment the component of interest in laparoscopic data from another patient. The models successfully detect and segment the target organ. For each experiment, U-Net models with different architectures are used as the backbone, demonstrating the effectiveness of the proposed method across various models and neural network architectures. The evaluation methodology is illustrated in Figure 6.

The rationale for not employing the proposed automatic KNN-based masking method for automatically masking all images can be attributed to two factors. Firstly, the psychology of human learning underpins the unsupervised method proposed in this paper. Given the critical role of memory and memorized knowledge in fostering the conceptualization of components in humans and animals, a neural network-based method is employed for the proposed self-learning approach. Our findings demonstrate that the neural network-based model significantly outperforms the KNN method in learning the concepts and components of interest, rather than merely categorizing colors.

Secondly, the use of CNNs in the proposed self-learning methodology inherently facilitates the calculation of numerous appearance-based measurements and features. Consequently, the KNN-based method is recommended only at the initial stage of the process when neither learned knowledge nor a trained CNN is available to select the most informative features from the images.

4.1. Ovary Semantic Segmentation Methodology Evaluation

In this section, following the proposed methodology, an ensemble neural network is developed to semantically segment the ovary within laparoscopic or robotic surgeries. Initially, the ovary is automatically masked using the proposed KNN-based algorithm. To facilitate the masking process, a web-based application is designed and developed, allowing the user to pinpoint the component of interest by clicking or dragging the mouse on some points on the ovary. The computer then detects pixels belonging to the ovary and distinguishes the ovary from other components within the image by executing the KNN-based algorithm. As a result, the ovary is automatically masked without the need for meticulous annotation by human experts. Figure 7 demonstrates the input and output of the proposed algorithm. As shown, the ovary is meticulously detected and masked within the image.

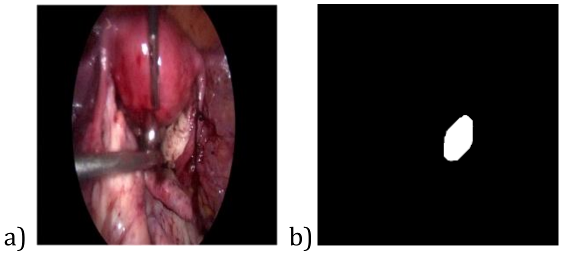

Figure 7, Fully-automatically masking of ovary by only clicking on the organ inside the image.

The features used in the proposed KNN-based algorithm are selected through thorough investigation. Our findings indicate that color is the most important feature for differentiating between different components. While geometry is also important and informative, it is considered secondary. Various scenarios were investigated in this research, including not considering geometric concepts such as the distance between pixels and the estimated central position of the component of interest, allocating the same weights to distance and color-based values, and using different distance metrics such as Manhattan. The proposed methodology detailed in Section 3.2 shows significantly the highest efficacy. For instance, Figure 8 illustrates a case where the geometric distance between pixels and the estimated central position of the component of interest is not considered in the proposed KNN-based method.

Automatically masking a component in an image simulates the process by which a teacher, mother, or professor introduces a component by simply pointing to it. In humans, many factors, including physical metrics like colors, elasticity, and coherent structure of the component, combine with three-dimensional geometric features. Therefore, humans can distinguish the component from neighboring components at first glance, a process simulated by the proposed KNN-based algorithm in this research. Subsequently, humans take a deep look at the component of interest, investigating it from different angles and orientations, touching and holding the component to observe it from as many different directions as possible. This process of deep investigation is simulated by the proposed self-learning method.

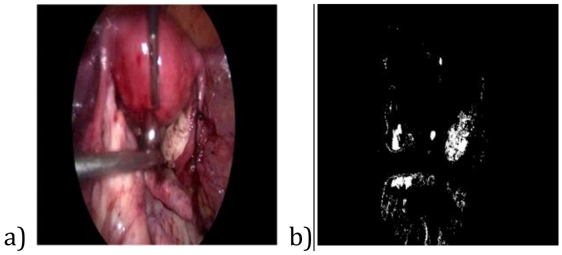

Figure 8, Fully-automatically masking of ovary by only clicking on the organ inside the image.

After automatically masking only two images from the laparoscopic videos, the proposed self-learning method is used to train CNN models that can semantically segment the ovary within laparoscopic images. This task is vital in the computerization of surgical procedures, as detailed in [12]. The self-learning method not only masks all the images but also results in a well-trained deep neural network framework that is highly powerful. For example, the resulting model is flexible to any changes in the lighting, location, or orientation of the camera because it has seen the component of interest from various perspectives [12].

In this section, two neural network architectures (ResNet50 and DenseNet201) are utilized as the backbones of U-Net models to develop semantic segmentation models for detecting and segmenting ovaries during laparoscopic and robotic surgeries using the proposed method. Since the purpose of this section is to evaluate the method, the self-learning process is independently repeated for each U-Net model. However, in other tasks, since self-learning is computationally demanding, it is sufficient to follow the self-learning methodology only once, as the generated masks can be used to train other neural network models, as detailed in Section 3.3.

In the first experiment, a U-Net model with ResNet50 architecture is utilized to investigate and evaluate the proposed self-learning method. The model is pre-trained on the ImageNet dataset. In the first round (Round 0), two images with their paired masks, produced by the proposed KNN-based algorithm, are used as the training dataset. Thus, the Learning Window has a size of two and contains the images and masks of the first and tenth frames of a laparoscopic surgery captured directly by the laparoscope. Since the amount of data in this phase is very low, it takes a few minutes for the model to be trained, requiring 1617 epochs of learning.

In this paper, Learning Speed is defined as the ratio of the number of rounds to the number of epochs needed for the complete learning process during certain rounds. The Learning Speed is described by Eq. 9, where k represents the number of rounds for a specific set of continuous rounds. The Learning Speed grows exponentially as the self-learning process progresses. Thus, the computational demand of the proposed self-learning methodology is only challenging at the beginning of the process. This phenomenon is due to two main factors: first, the model retains its extracted knowledge from previous rounds and only updates itself in newer rounds; second, the amount of training data increases, providing the model with more data to learn from in each round.

$$Learning\ Speed = \frac{k}{\sum_k Number\ of\ epochs\ for\ Round_k} \quad \text{Eq.9}$$

At the end of the Round 0 process, the model has been trained on the same content in the video with only minor changes in the camera's position. As a result, the model can detect the ovary in images that are very similar to those used in the training dataset, as it is flexible only to small changes in the camera's position. Consequently, the model can successfully detect and segment the ovary in all the images of the Test Window, as the variations in the Test Window images are smaller or equal to the changes between the first and tenth images of the video, which were selected as the Learning Window.

In Round 1, the segmented images from Round 0 are added to the training dataset, and the model is re-trained on the updated dataset. As the model observes images with more diversity in the camera's position, its flexibility increases, allowing it to successfully segment images that are less similar to those in the training dataset. Thus, the Test Window consists of the next 10 images in the laparoscopic video. At the end of Round 1, the model can successfully segment the ovary in the images of the Test Window. Therefore, the segmented images can be used as new masked data for re-training the model in subsequent rounds.

The details of this experiment are provided in Table 1, where the number of pages or frames starts at 0 and ends at 211. The process is conducted under in-depth supervision and investigation, demonstrating that the proposed methodology in this paper is not only feasible but also effective and reliable.

Table 1, The details for training ovary segmentation U-Net model with ResNet50 backbone.

Round Number	Learning Window			Test Window			Learning Speed	Self-Learning rate	Model's Flexibility
	Start Page	Ending Page	Window Size	Start Page	Ending Page	Window Size			
0	0000	0009	2	0001	0008	8	1/1617	10	Flexible to very small motions in the camera
1	0000	0009	10	0010	0019	10	1/239	20	
2	0000	0019	20	0020	0029	10	1/101	30	
3	0000	0029	30	0030	0039	10	1/58	40	Flexible to small motions in the camera
4	0000	0039	40	0040	0049	10	1/17	50	
5	0000	0049	50	0050	0069	20	1/10	70	
6	0000	0069	70	0070	0119	50	1/6	120	Flexible to any motions in the camera
7	0000	0119	120	0120	0159	30	1/3	150	
8	0000	0159	160	0168	0210	43	1/2	203	

As discussed, the model's flexibility grows as the self-learning process progresses. Our findings show that this growth in the model's flexibility and power is not linear but increases exponentially. This exponential rise in flexibility can be seen in the size of the Test Window. Despite the small size of the Test Window in the initial rounds, it can grow exponentially. Consequently, the self-learning methodology enables computers to learn from big data within a small number of rounds, making the methodology both effective and fast.

The experiment that followed the proposed methodology to develop a semantic segmentation model using a ResNet50 architecture was repeated with another neural network architecture. Since flexibility is a key feature of CNN models in this paper, DenseNet201 was used to repeat the experiment, as it is expected that deeper networks can learn image details more effectively and, as a result, achieve higher flexibility in each round. The results of the second experiment for developing an ovary segmentation model demonstrate that the proposed methodology is effective across diverse neural network architectures, although deeper neural networks may have higher efficacy.

In this experiment, the physical and geometric denoiser is further enhanced using a computational geometry algorithm. Given the convex shape of the ovary, calculating the convex hull of the segmented images can improve the effectiveness of the process. This approach simulates processes occurring in the human brain. As discussed in Section 1, human understanding relies not only on vision but also on the physical and geometric concepts embedded in the brain. Thus, calculating the convex hull effectively combines visual signals with the physical and geometric concepts, simulating the human cognitive processes.

By enhancing the physical and geometric denoiser system and employing a deeper neural network, the self-learning process is significantly accelerated. The details of developing the segmentation model to detect the ovary using the DenseNet201 architecture are provided in Table 2. As shown, in just ten rounds, more than one thousand and one hundred of images are automatically masked due to the substantial growth in the size of the Test Window.

In the laparoscopic video containing the ovary used in this section, there is noise from page numbers 0160 to 0170, where the entire image appears completely black due to a sudden outage of the laparoscope camera. In the experiment using the ResNet50 model, this period was removed during the data preprocessing steps. However, in the DenseNet201 experiment, this period was not removed. To be cautious, the size of the Test Window is designed to be small. Our findings show that the model can successfully handle the noisy frames in the surgical videos. As a result, the proposed methodology is effective even in cases where there are intense changes with unexpected noises.

Table 2, The details for training ovary segmentation U-Net model with DenseNet201 backbone.

Round Number	Learning Window			Test Window			Learning Speed	Self-Learning rate	Model's Flexibility
	Start Page	Ending Page	Window Size	Start Page	Ending Page	Window Size			
0	0000	0009	2	0001	0008	8	1/1397	10	Flexible to very small motions in the camera
1	0000	0009	10	0010	0019	10	1/891	20	
2	0000	0019	20	0020	0087	68	1/317	88	
3	0000	0029	30	0088	0167	80	1/86	168	Flexible to small motions in the camera
4	0000	0169	170	0168	0175	8	1/13	176	
5	0101	0175	75	0176	0349	174	1/37	350	
6	0178	0349	172	0350	0449	100	1/52	450	Flexible to any motions in the camera
7	0350	0549	200	0450	0876	427	1/47	877	
8	0700	0876	177	0877	1169	293	1/63	1170	

By integrating the resulting neural networks that can reliably detect and segment the ovaries during laparoscopic and robotic surgeries, an ensemble neural network framework is developed in this section to semantically segment the ovaries in surgical images and videos. The ensemble framework is then evaluated using 10 new images from another patient in which the ovaries are visible. The findings demonstrate that the developed framework can effectively detect the ovaries, as the IoU accuracy for the test images is measured at 93.11%. However, by utilizing more surgical videos and following the proposed methodology, more powerful frameworks can be developed. Since the proposed method is an unsupervised approach, elevating the developed framework with more videos is considerably simplified for human experts.

Using the segmentation framework developed in this section, a localization system can also be created by reusing the trained model. This allows for a meticulous evaluation to assess whether the proposed method results in reliable and accurate trained neural networks. The findings demonstrate that the proposed method is both reliable and effective, as the localization system for detecting the ovary in laparoscopic videos shows high accuracy and reliability, even during intense movements of the ovary and neighboring organs. The video demonstrating ovary localization by the developed system is available in the link detailed in Section 7.1.

4.2. Uterine Semantic Segmentation Methodology Evaluation

This section aims to develop an ensemble neural network for the semantic segmentation of the uterus during laparoscopic or robotic surgery. The approach leverages the proposed unsupervised methodology for training semantic segmentation models. Precise detection of the uterus is crucial for many surgical systems, including the method proposed in [12] for uterine endometriosis treatment.

Our focus on the uterus in this section is due to its color similarity to neighboring organs, such as the peritoneum, in many laparoscopic or robotic surgeries. This similarity poses a significant challenge for semantic segmentation models, making it difficult to distinguish the uterus from adjacent parts. By subjecting the proposed method to these challenging conditions, the performance and effectiveness of the proposed method can be rigorously assessed. If the method successfully passes this evaluation, it can be considered reliable and technically feasible for a wide range of applications, both in medical and non-medical projects. For this purpose, the scenario described in Section 4.1 is pursued for developing a uterus segmentation framework.

First, the proposed KNN-based algorithm and the developed web-based graphical user interface are utilized to automatically mask two images of the dataset. In this procedure, the laparoscopic image is displayed in the web-based application, and the user is asked to pinpoint the uterus by clicking or dragging the mouse over it. While clicking on the component of interest, some pixels inside the uterus are detected by the computer. The computer can then distinguish the uterus by following the proposed KNN-based algorithm. One of the results of the automatic uterus masking is illustrated in Figure 9. As shown, the computer carefully masks the uterus without requiring meticulous annotation by human experts.

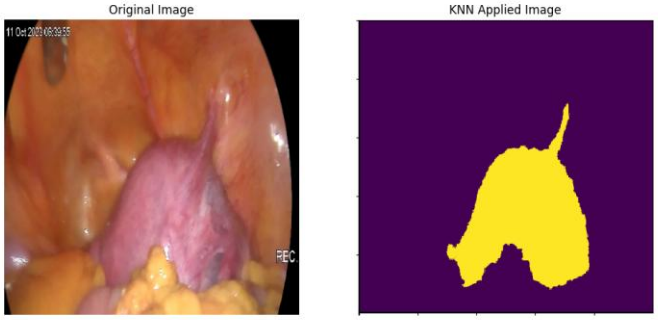

Figure 9, Automatically masking uterus by the proposed KNN-based algorithm.

After automatically masking two images from the laparoscopic video, a U-Net model with MobileNet architecture, pre-trained on the ImageNet dataset, is used. This experiment demonstrated that the proposed methodology is effective; without any manual annotation, the model meticulously segmented and masked more than six hundred images. The details of the experiment are provided in Table 3.

As shown in Table 3, during Rounds 6 to 8, the size of the Test Window was designed to be very small. During these rounds, the model was processing frames numbered 2602 to 2607. These frames belong to a scene in the laparoscopic video where intense camera movement occurred, making the test images relatively different from those the model had encountered previously. This phenomenon, also observed in shocked humans and animals, occurs when an individual encounters new conditions different from previous experiences.

To address this challenge, the learning speed was decreased by reducing the size of the Test Window in the related rounds. This allowed the model to encounter new images slowly, and analyze them more effectively. This experiment demonstrates that this strategy enables the model to learn patterns from sudden changes without any assistance from human experts, provided the changes are not too acute. Decreasing the size of the Test Window simulates the psychological strategy where the learner is given more time to conceptualize the component of interest and the complexity of the task is reduced by the trainers.

Table 3, The details for training uterus segmentation U-Net model with MobileNet backbone.

Round Number	Learning Window			Test Window			Learning Speed	Self-Learning rate	Model's Flexibility
	Start Page	Ending Page	Window Size	Start Page	Ending Page	Window Size			
0	2482	2491	2	2483	2490	8	1/553	10	Flexible to very small motions of the camera
1	2482	2490	10	2492	2501	10	1/83	20	
2	2482	2501	20	2502	2521	20	1/1	40	
3	2482	2521	40	2522	2541	20	1/1	60	Flexible to small motions of the camera
4	2482	2541	60	2542	2581	40	1/1	100	
5	2482	2581	100	2582	2601	20	1/1	120	
6	2482	2601	120	2602	2605	4	½	124	Earning Flexibility for intense motions of the camera
7	2482	2605	124	2606	2606	1	1/1	125	
8	2482	2606	125	2607	2607	1	1/2	126	
9	2482	2607	126	2608	2621	14	1/3	140	
10	2482	2621	140	2622	2701	80	1/1	220	Flexible to any motion of the camera
11	2482	2701	220	2702	2801	100	1/5	320	
12	2482	2801	320	2802	2951	150	1/3	470	
13	2632	2951	320	2952	3110	159	½	628	

While the training of the model can be continued by decreasing the size of the Test Window in specific periods, utilizing deeper neural networks is another solution. The experiment of training a segmentation model for the uterus was repeated using a U-Net model with a ResNet50 backbone. Since ResNet50 is deeper than MobileNet, it can learn more complex patterns from the data and achieve higher flexibility. The details of the experiment in which the ResNet50-based model was developed for uterus segmentation are provided in Table 4. As shown, the sudden change in the camera's position did not cause any obstacles. Additionally, the size of the Test Window exponentially grew during the self-learning process. The same exponential growth was observed in the Learning Speed and Self-Learning rate. As a result, the proposed unsupervised methodology for semantic segmentation in this paper is technically feasible, effective, and reliable.

Finally, by integrating the two developed U-Net models, an ensemble neural network framework is created that can meticulously detect and segment the uterus in laparoscopic and robotic surgeries. Using the developed ensemble framework, images from another patient, in which the uterus is visible, were provided for segmentation. The ensemble method was able to meticulously segment the uterus with an IoU of 96.14%. However, the accuracy can be further improved by following the distributed approach described in Section 3.3, which involves producing a larger dataset and consequently feeding more surgical videos to the framework.

The ensemble frameworks for segmenting the uterus and ovary during surgery are useful in computer-based surgical assistant systems and the full automation of surgeries. The developed ensemble neural network frameworks in this research can be used as modules to denoise surgical images and maneuver robotic arms within the patient's body when the entire surgery is conducted fully automatically by artificial intelligence-based methods. For instance, these frameworks can be used in the proposed system for fully automating endometriosis surgery, as detailed in [12]. Using multiple neural networks for the same task allows the models to extract and learn the same patterns while learning different noises. Thus, integrating multiple neural networks for one purpose can reduce noise and unexpected errors, thereby maximizing the system's reliability [11]. Consequently, the proposed ensemble frameworks are reliable for use in medical and surgical applications.

Table 4, The details for training uterus segmentation U-Net model with ResNet50 backbone.

Round Number	Learning Window			Test Window			Learning Speed	Self-Learning rate	Model's Flexibility
	Start Page	Ending Page	Window Size	Start Page	Ending Page	Window Size			
0	2482	2491	2	2483	2490	8	1/1508	10	Flexible to very small motions of the camera
1	2482	2491	10	2492	2501	10	1/517	20	
2	2482	2501	20	2502	2521	20	1/88	40	
3	2482	2521	40	2522	2541	20	1/7	60	Flexible to small motions of the camera
4	2482	2541	60	2542	2581	40	1/1	100	
5	2482	2581	100	2582	2601	20	½	120	
6	2482	2601	120	2602	2745	143	1/1	263	Flexible to any motion of the camera
7	2482	2745	263	2746	2956	211	1/1	474	
8	2482	2956	474	2957	3110	154	1/1	628	

Since the medical and surgical images utilized for the evaluation of the proposed methodology are challenging to process even for humans, the proposed method is expected to be effective and reliable. Another notable feature of the proposed method is its resulting models' flexibility to changes in the camera's position, such as rotation, zooming in and out, illumination, darkening, and adjustments in camera settings like contrast, color warmth, and white balance. This flexibility is achieved because the model has encountered the component of interest through images with various configurations.

Enhancing the model's flexibility through the proposed method is crucial, as it promotes the system's naturalness by mirroring the learning and detection process in humans. Furthermore, this feature enhances the system's reliability, as such variations are common in real-life multimedia data.

Similar to Section 4.1, a localization system has been created using the ensemble neural network frameworks developed in this section. Our findings indicate that the localization system developed from the ensemble frameworks can accurately and reliably detect the uterus in laparoscopic videos. Consequently, the proposed unsupervised semantic segmentation methodology proves effective not only for semantic segmentation but also for localization purposes and other tasks. The results from the localization system for the uterus are available with the link provided in Section 7.1.

5. Discussion and Future Works

In this study, an unsupervised method for training and developing semantic segmentation CNN models is proposed which is adaptive for various tasks. First, the self-learning method is proposed in which a semantic segmentation model can be trained on a dataset containing only two annotated images. Next, a KNN-based algorithm is proposed that enables human users to automatically mask any component within an image by simply clicking on it, without the need for meticulous annotation or masking. By integrating the proposed methods, the entire process of semantic segmentation can be automated in an unsupervised approach.

The aim of this research is to simulate the learning process when teachers, professors, or caregivers introduce a component by simply pointing to it. Given that physical and three-dimensional geometric concepts are vital in human learning, the proposed methodology for unsupervised semantic segmentation training is enhanced by algorithms that integrate these physical and geometric concepts with neural networks. Our findings show that this integration significantly enhances the learning procedure, allowing the model to independently learn how to detect and segment the component of interest in the videos without any supervision or guidance from humans.

In this research, laparoscopic images and videos were investigated to evaluate the proposed methodology. Our findings show that the proposed method is reliable for medical images. Given that medical images are more challenging compared to other fields, the proposed methodology is expected to be reliable for a wide range of applications and tasks. However, further investigation is needed to explore how the proposed method can be applied to other fields such as biology, transportation, social environments, and even fashion.

The geometric algorithms proposed in this paper have been validated as they can significantly enhance the effectiveness of the segmentation by denoising the outputs of the neural networks. However, additional research is needed to develop faster algorithms that can detect and address different types of noise, such as smoothing sharp edges of detected components, converting binary images into Doubly Connected Edge Lists (a data structure used in computational geometry to represent maps of areas), and determining whether an area has an abnormal boundary shape. Achieving these goals can significantly improve the quality of the segments produced by the proposed method. However, they are scheduled to be investigated in future works.

In this paper, the learning progress direction is investigated in one time direction, as illustrated in Figure 2. However, the self-learning process can be conducted in both directions, with two different Test Windows—one moving forward in time and the other moving backward. In such cases, the initial phase (Round 0) should start from the middle of the dataset. This method can double the speed of the entire self-learning process, minimize computations, and result in green computing. However, this scenario requires further investigation and is scheduled to be studied in future work.

6. Appendix

In this section, the architectures of the neural networks used in this research are detailed to maximize the transparency of the paper. CNNs are deep neural networks designed to process visual data such as images or videos. They include convolutional layers that extract features from the image, visualized as a mathematical matrix. Convolutional layers divide the matrix into smaller submatrices to search for specific features, such as edges or colors, which are essential for analyzing the image. Consequently, the convolutional layers discover informative details from the image and provide this information to other network layers. Pooling layers summarize the extracted features from the convolutional layers, and after these analyses, the discovered information is sent to fully-connected layers, which consist of standard neurons that receive outputs from all neurons in the previous layer [28-30].

U-Net models are a class of CNNs used for image semantic segmentation purposes. They contain two phases: the encoder (or backbone) and the decoder. In the backbone, the network extracts visual details from the image. In U-Net models, at each level of feature extraction in the backbone, the image size decreases while its depth increases, meaning the image gets smaller but more informative with visual details. This process is repeated multiple times in the backbone, and then reversed in the decoder phase. Finally, the image is segmented at the output layer of the network. The combination of the encoder, which condenses the image, and the decoder, which restores the image to its original size, generates a U-shaped architecture. This design is why the model is named the U-Net model [12].

Neurons are mathematical simulations of the cells that shape the brains of humans and animals. Each neuron in the human brain is biologically a cell that can receive messages from neighboring neurons and generate a message to other neurons if the incoming message is sufficiently important. This concept is modeled in artificial neural networks, where each neuron receives multiple numbers, each containing a message. To summarize these messages, a weighted sum is performed. Finally, an activation function determines if the summarized message is important enough to generate a signal for the next neurons.

The brain of humans and animals is a complex network of interconnected neurons. Consequently, by creating and connecting neurons in software, it is expected that computers could behave as intelligently as humans. Numerous studies demonstrate that this strategy is indeed feasible, as computers can analyze various data for different tasks, such as classification, segmentation, and even generation [12, 38-40]. There are different types of neurons in humans and animals, each responsible for detecting and managing specific phenomena. For example, in vision, some neurons are responsible for detecting colors, while others detect brightness. Consequently, CNNs are utilized in this paper because their convolutional layers simulate visual neurons effectively. Therefore, the proposed learning method in this paper is not only accurate and reliable but also natural and interpretable, as it meticulously simulates the real process of visual learning and conceptualization in humans.

In this Section, the convolutional layerare abbreviated as 'Conv2D' or 'Conv', the batch normalization layer as 'Batch', and the addition layer that computes the sum of two matrices as 'Add'. The zero-padding layer that expands a matrix by adding zeros around it is referred to as 'Zero' or 'Zero Pad', the average pooling layer that averages its input values as 'Average' or 'Avg', and the concatenation layer that links two arrays as 'Concatenate' or 'ConCat' or 'Concat'. Additionally, 'Global', 'Reshape', 'Depth', 'Act', 'Up', and 'Fix' denote the layers of global average pooling 2D, reshape, depthwise convolution 2D, activation layer, up sampling layer, and fixed dropout, respectively.

Additionally, the function of the 'product of a module' indicates its repeated application. For instance, 'L1 × 2' signifies that the output of one L1 layer is fed into another L1 layer sequentially. Similarly, the function of 'addition' involves sending the output of the left module as the input to the right layer. For example, 'M1 × 3 + M2' means that there are three sequential M1 modules, where the output of each one is fed into the next M1 module, and the output of the final M1 module is fed into the M2 module. These definitions simplify explaining the architecture of the neural networks.

Figure 10 demonstrates the architecture of the U-Net model with DenseNet201 as the backbone. This model is used in this paper for developing the ovary segmentation system. As shown, the neural network is deep in processing layers and has a U-shaped architecture. The modules used in this architecture are illustrated in Figures 11, 12, 13, 14, and 15.

Figure 16 illustrates the design of the U-Net model with a ResNet50 backbone. This depiction synthesizes the components outlined in Figures 17 through 21. The ResNet architecture innovatively incorporates shortcut connections between layers to mitigate the vanishing gradient problem encountered in deeper network layers. These shortcuts are visible in modules L3 and L4. In the L3 module, Output 1 connects to the network's deeper layers in the decoder, while Output 2 links directly to the subsequent layer in the encoder. Similarly, the architecture of the U-Net model with a MobileNet backbone is shown in Figure 22, with its inner modules illustrated in Figures 23 through 28.

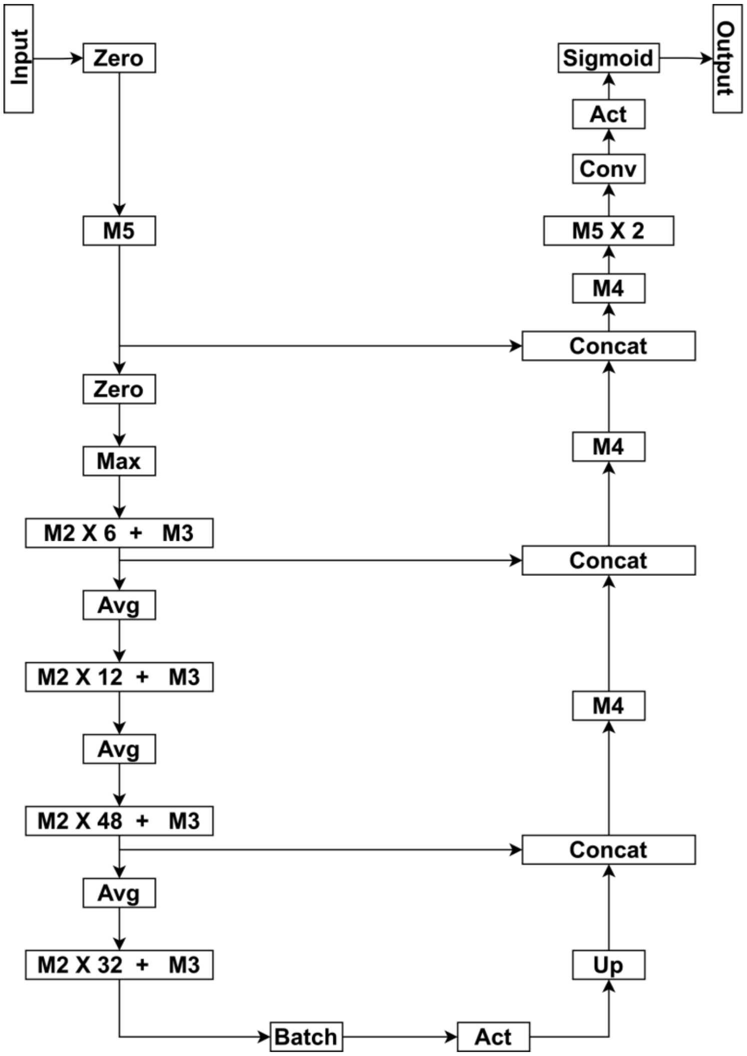

Figure 10, The architecture of the U-Net model with DenseNet201 backbone.

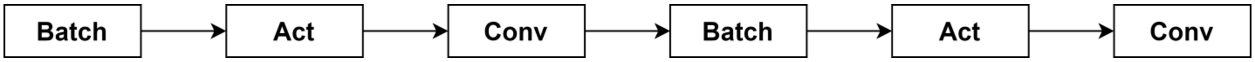

Figure 11, The structure of the M1 module used in the U-Net with DenseNet201backbone.

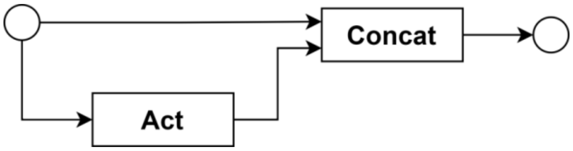

Figure 12, The structure of M2 module used in the U-Net model with DenseNet201 backbone.

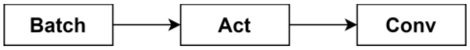

Figure 13, The structure of M3 module used in the U-Net model with DenseNet201 backbone.

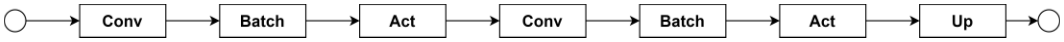

Figure 14, The structure of M4 module used in the U-Net model with DenseNet201 backbone.

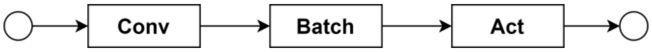

Figure 15, The structure of M5 module used in the U-Net model with DenseNet201 backbone.

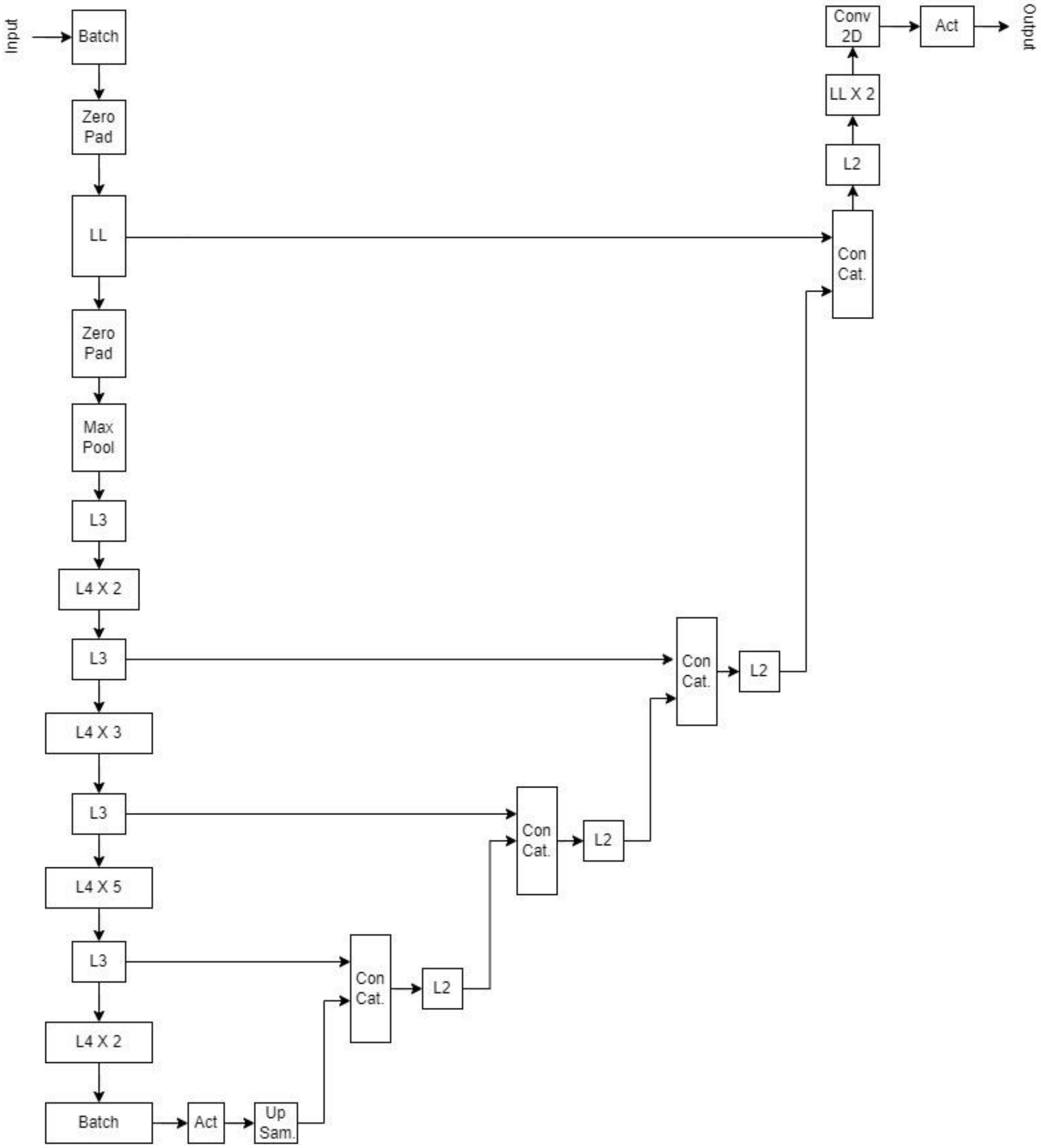

Figure 16, Architecture of a U-Net model with ResNet50 backbone.

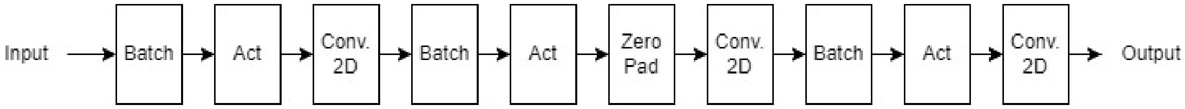

Figure 17, Structure of L1 module of U-Net model with ResNet50 backbone.

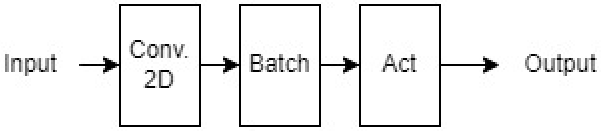

Figure 18, Structure of LL module of U-Net model with ResNet50 backbone.

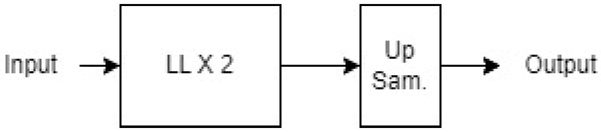

Figure 19, Structure of L2 module of the U-Net model with ResNet50 backbone.

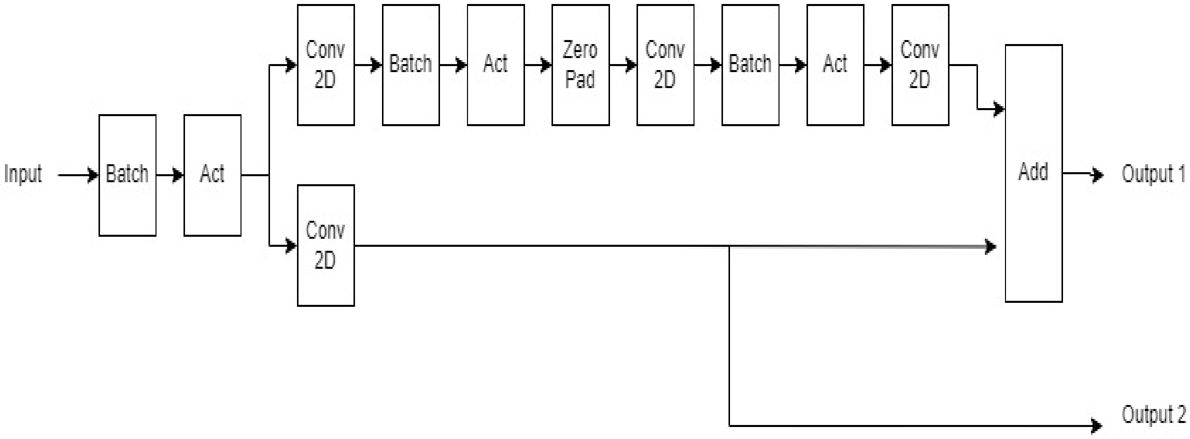

Figure 20, Structure of L3 module of the U-Net model with ResNet50 backbone.

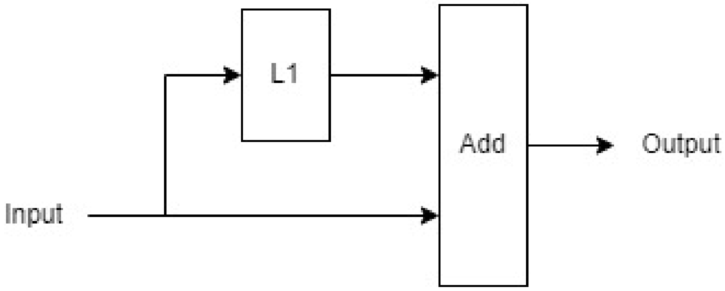

Figure 21, Structure of L4 module of the U-Net model with ResNet50 backbone.

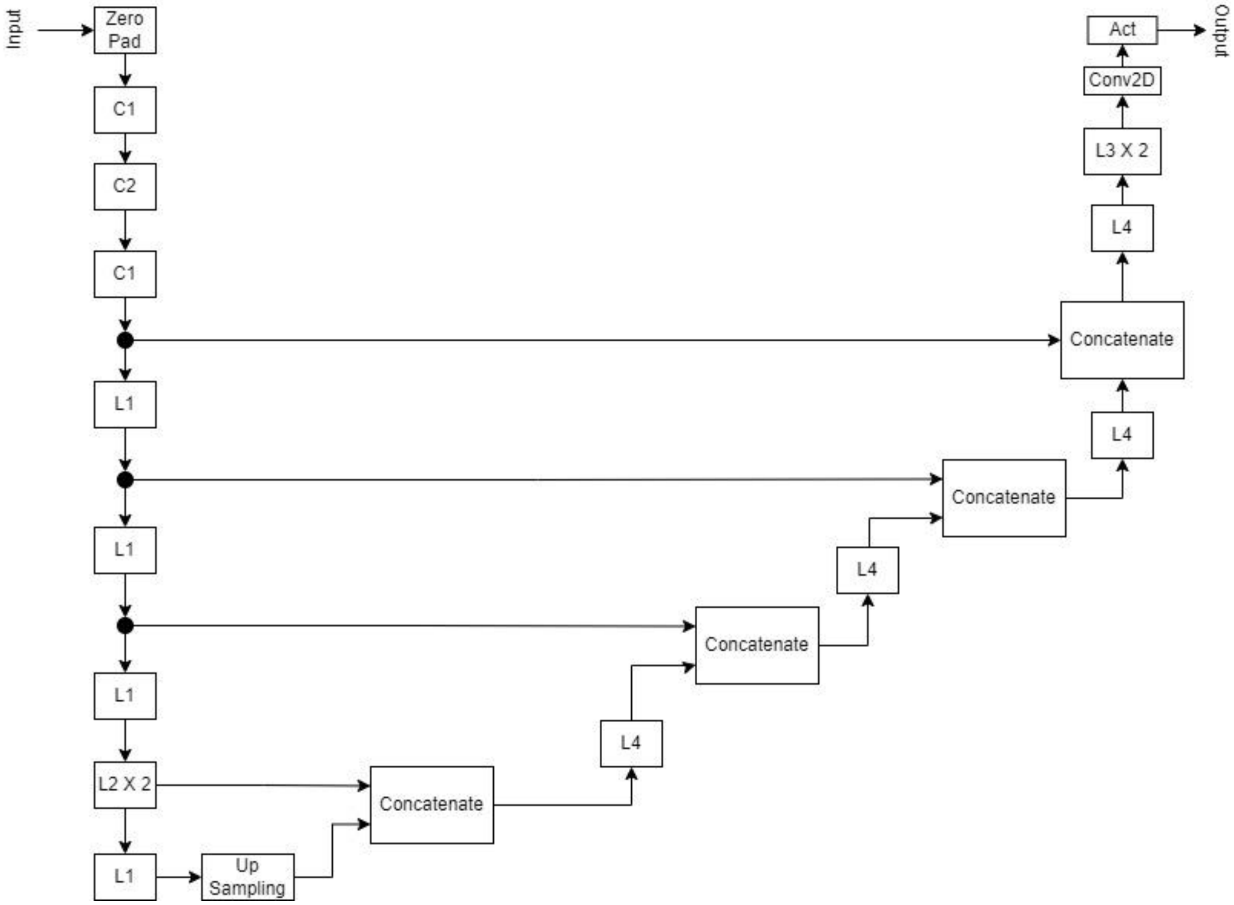

Figure 22, Architecture of U-Net model with MobileNet backbone.

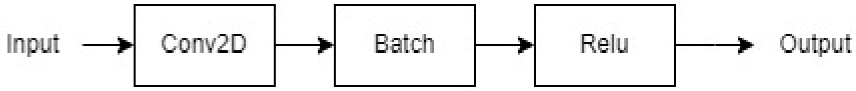

Figure 23, Structure of C1 module in U-Net model with MobileNet backbone.

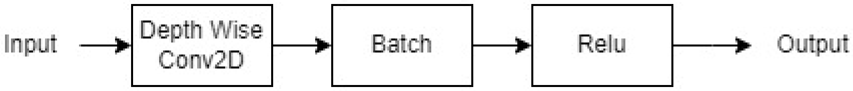

Figure 24, Structure of C2 module in U-Net model with MobileNet backbone.

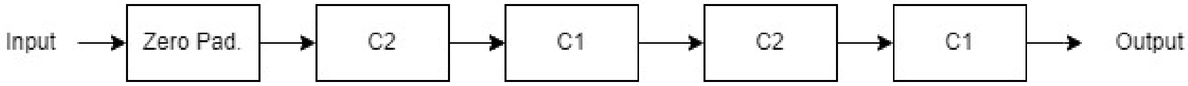

Figure 25, Structure of L1 module in U-Net model with MobileNet backbone.

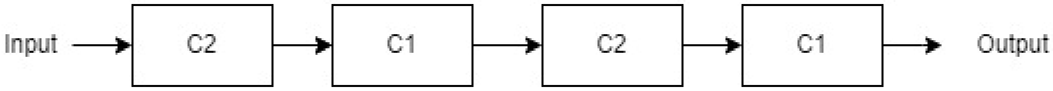

Figure 26, Structure of L2 module in U-Net model with MobileNet backbone.

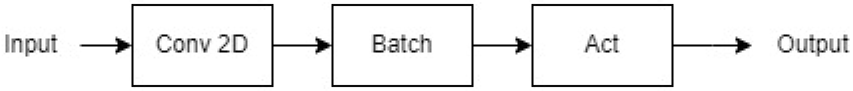

Figure 27, Structure of L3 module in the U-Net model with MobileNet backbone.

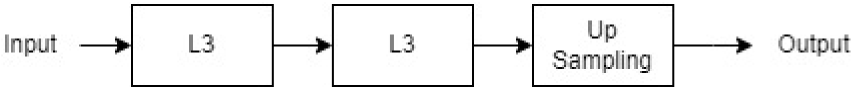

Figure 28, Structure of L4 module in U-Net model with MobileNet backbone.

7. Ethics

This study was conducted in accordance with the ethical standards. To ensure confidentiality and anonymity, personal identifiers were removed from all data, and participants were assigned unique codes. Data was stored securely and only accessible to the research team. Participants were informed that their participation was voluntary and that they could withdraw from the study at any time without any negative consequences.

7.1. Data Availability

The data from this research has been published on GitHub. You are invited to explore the repository to review the results of the detection systems developed in this study, showcasing their effectiveness, accuracy, and reliability. These results demonstrate that the proposed method in this paper is both effective and reliable. You can access the repository at the following link:
https://github.com/sinasaadati95/self-learning/

7.2. Acknowledgement

Special thanks for Dr. Maryam Hashemi, the assistant professor and laparoscopic surgeon at Isfahan University of Medical Science, Isfahan, Iran for providing the laparoscopic dataset.

References

[1] Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach. pearson.

[2] Burkov, A. (2020). Machine learning engineering (Vol. 1). Montreal, QC, Canada: True Positive Incorporated.

[3] Alpaydin, E. (2021). Machine learning. MIT press.

[4] Zaki, M. J., Meira Jr, W., & Meira, W. (2020). Data mining and machine learning: Fundamental concepts and algorithms. Cambridge University Press.

[5] Guo, Y., Liu, Y., Georgiou, T., & Lew, M. S. (2018). A review of semantic segmentation using deep neural networks. International journal of multimedia information retrieval, 7, 87-93. https://doi.org/10.1007/s13735-017-0141-z

[6] Razavi-Far, R., Ruiz-Garcia, A., Palade, V., & Schmidhuber, J. (Eds.). (2022). Generative adversarial learning: architectures and applications. Cham: Springer.

[7] Aggarwal, C. C. (2015). Data mining: the textbook (Vol. 1, No. 3). New York: springer.

[8] Barragán-Montero, A., Javaid, U., Valdés, G., Nguyen, D., Desbordes, P., Macq, B., ... & Lee, J. A. (2021). Artificial intelligence and machine learning for medical imaging: A technology review. Physica Medica, 83, 242-256.

[9] Naqvi, N. Z., Kaur, K., Khanna, S., & Singh, S. (2023). An overview of machine learning techniques focusing on the diagnosis of endometriosis. Machine Vision and Augmented Intelligence: Select Proceedings of MAI 2022, 61-84. https://doi.org/10.1007/978-981-99-0189-0_6

[10] Rezaei, Z. (2021). A review on image-based approaches for breast cancer detection, segmentation, and classification. Expert Systems with Applications, 182, 115204.

[11] Saadati, S., Sepahvand, A., & Razzazi, M. (2025). Cloud and IoT based smart agent-driven simulation of human gait for detecting muscles disorder. Heliyon. https://doi.org/10.1016/j.heliyon.2025.e42119

[12] Saadati, S., & Amirmazlaghani, M. (2024). Revolutionizing endometriosis treatment: automated surgical operation through artificial intelligence and robotic vision. Journal of Robotic Surgery, 18(1), 383. https://doi.org/10.1007/s11701-024-02139-7

[13] Larrañaga, P., Atienza, D., Diaz-Rozo, J., Ogbechie, A., Puerto-Santana, C. E., & Bielza, C. (2018). Industrial applications of machine learning. CRC press. https://doi.org/10.1201/9781351128384

[14] Datta, S., & Davim, J. P. (Eds.). (2021). Machine learning in industry. Springer Nature.

[15] Hassoun, M. H. (1995). Fundamentals of artificial neural networks. MIT press.

[16] Suzuki, K. (Ed.). (2013). Artificial neural networks: Architectures and applications. BoD–Books on Demand.

[17] Roberts, D. A., Yaida, S., & Hanin, B. (2022). The principles of deep learning theory (Vol. 46). Cambridge, MA, USA: Cambridge University Press.

[18] Pande, B., Padamwar, K., Bhattacharya, S., Roshan, S., & Bhamare, M. (2022, May). A review of image annotation tools for object detection. In 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 976-982). IEEE.

[19] Edalat, A. (2017). Self-attachment: A holistic approach to computational psychiatry. Computational neurology and psychiatry, 273-314. https://doi.org/10.1007/978-3-319-49959-8_10

[20] Qin, X., He, S., Zhang, Z., Dehghan, M., & Jagersand, M. (2018, March). Bylabel: A boundary based semi-automatic image annotation tool. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1804-1813). IEEE.

[21] Scheikl, P. M., Laschewski, S., Kisilenko, A., Davitashvili, T., Müller, B., Capek, M., ... & Mathis-Ullrich, F. (2020, September). Deep learning for semantic segmentation of organs and tissues in laparoscopic surgery. In Current directions in biomedical engineering (Vol. 6, No. 1, p. 20200016). De Gruyter.

[22] Zhao, Z., Jin, Y., Gao, X., Dou, Q., Heng, PA. (2020). Learning Motion Flows for Semi-supervised Instrument Segmentation from Robotic Surgical Video. In: Martel, A.L., et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science(), vol 12263. Springer, Cham. https://doi.org/10.1007/978-3-030-59716-0_65

[23] Wang, Z., Liu, X., Perreault, C., & Jarc, A. (2023). Automatic detection of out-of-body frames in surgical videos for privacy protection using self-supervised learning and minimal labels. Journal of Medical Robotics Research, 8(01n02), 2350002.

[24] P. Yu, A. K. Qin and D. A. Clausi, "Unsupervised Polarimetric SAR Image Segmentation and Classification Using Region Growing With Edge Penalty," in IEEE Transactions on Geoscience and Remote Sensing, vol. 50, no. 4, pp. 1302-1317, April 2012, doi: 10.1109/TGRS.2011.2164085

[25] Kalluri, T., Varma, G., Chandraker, M., & Jawahar, C. V. (2019). Universal semi-supervised semantic segmentation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 5259-5270).

[26] Gao, S., Li, Z. Y., Yang, M. H., Cheng, M. M., Han, J., & Torr, P. (2022). Large-scale unsupervised semantic segmentation. IEEE transactions on pattern analysis and machine intelligence, 45(6), 7457-7476.

[27] Saadati, S., & Razzazi, M. (2022). Natural way of solving a convex hull problem. arXiv preprint arXiv:2212.11999. https://doi.org/10.48550/arXiv.2212.11999

[28] Khan, S., Rahmani, H., Shah, S. A. A., Bennamoun, M., Medioni, G., & Dickinson, S. (2018). A guide to convolutional neural networks for computer vision.

[29] Venkatesan, R., & Li, B. (2017). Convolutional neural networks in visual computing: a concise guide. CRC Press.

[30] Michelucci, U. (2019). Advanced applied deep learning: convolutional neural networks and object detection. Apress.