Journal of *Imaging*

MDPI

Article

Enhanced U-Net with GridMask (EUGNet): A Novel Approach for Robotic Surgical Tool Segmentation

Mostafa Daneshgar Rahbar [1,*] and Seyed Ziae Mousavi Mojab [2]

[1] Department of Electrical and Computer Engineering, Lawrence Technological University, Southfield, MI 48075, USA
[2] Department of Computer Science, Wayne State University, Detroit, MI 48202, USA; mousavi@wayne.edu
* Correspondence: mrahbar@ltu.edu

Abstract: This study proposed enhanced U-Net with GridMask (EUGNet) image augmentation techniques focused on pixel manipulation, emphasizing GridMask augmentation. This study introduces EUGNet, which incorporates GridMask augmentation to address U-Net's limitations. EUGNet features a deep contextual encoder, residual connections, class-balancing loss, adaptive feature fusion, GridMask augmentation module, efficient implementation, and multi-modal fusion. These innovations enhance segmentation accuracy and robustness, making it well-suited for medical image analysis. The GridMask algorithm is detailed, demonstrating its distinct approach to pixel elimination, enhancing model adaptability to occlusions and local features. A comprehensive dataset of robotic surgical scenarios and instruments is used for evaluation, showcasing the framework's robustness. Specifically, there are improvements of 1.6 percentage points in balanced accuracy for the foreground, 1.7 points in intersection over union (IoU), and 1.7 points in mean Dice similarity coefficient (DSC). These improvements are highly significant and have a substantial impact on inference speed. The inference speed, which is a critical factor in real-time applications, has seen a noteworthy reduction. It decreased from 0.163 milliseconds for the U-Net without GridMask to 0.097 milliseconds for the U-Net with GridMask.

Keywords: minimally invasive surgery; convolutional neural network; U-Net; data augmentation; surgical tools segmentation; computer vision; image processing

Citation: Daneshgar Rahbar, M.; Mousavi Mojab, S.Z. Enhanced U-Net with GridMask (EUGNet): A Novel Approach for Robotic Surgical Tool Segmentation. *J. Imaging* **2023**, *9*, 282. https://doi.org/10.3390/jimaging9120282

Academic Editors: Gerardo Cazzato and Francesca Arezzo

Received: 22 October 2023
Revised: 13 December 2023
Accepted: 15 December 2023
Published: 18 December 2023

1. Introduction

Minimally invasive surgery (MIS), often referred to as laparoscopic surgery or minimally invasive procedures, offers several notable advantages over traditional open surgery. These advantages include smaller incisions, resulting in less visible scarring [1], quicker patient recovery times, and shorter hospital stays [2]. MIS is associated with reduced postoperative pain and a decreased need for pain medication [3]. Additionally, the smaller incisions reduce the risk of surgical site infections, as there is less exposure to external contaminants [4]. Moreover, MIS procedures lead to reduced intraoperative and postoperative blood loss, which is particularly advantageous in cases where blood conservation is critical [5]. The minimized incision size also reduces the risk of wound complications, such as dehiscence, hernias, and evisceration [6], and results in better cosmetic outcomes [7]. Furthermore, MIS often enables patients to return to a regular diet more quickly, promoting early recovery, and can result in cost savings for both patients and healthcare systems [8].

MIS involves performing surgical procedures through small incisions rather than a large open incision. The surgical process begins with the surgeon creating several small incisions, typically ranging from 0.5 to 1.5 cm in length, near the surgical site [9]. These incisions serve as entry points for specialized surgical instruments and a camera. Subsequently, trocars, long, slender instruments, are inserted through these small incisions, providing access for the surgical instruments and the camera [10]. A laparoscope, a thin

tube equipped with a camera and light source at its tip, is inserted through one of the trocars, providing high-definition images of the surgical area displayed on an operating room monitor. This real-time visual feedback guides the surgeon throughout the procedure. Specialized surgical instruments, such as scissors, graspers, cautery devices, and suturing tools, are inserted through the other trocars and are designed for various surgical tasks, such as cutting, cauterizing, and suturing. The surgeon manipulates these instruments from outside the body while monitoring the live video feed on the monitor, allowing for precise control and fine movements of the instruments [11]. Depending on the specific surgical procedure, the surgeon may manipulate tissues, remove or repair damaged structures, and complete the necessary steps to address the medical condition. Throughout the surgery, the surgical team continuously monitors the patient's vital signs, such as heart rate, blood pressure, and oxygen saturation, to ensure the patient's safety and well-being. After completing the surgical procedure, the surgeon removes the surgical instruments and trocars. The small incisions may be closed using sutures, surgical glue, or adhesive strips, or, in some cases, they may not require closure at all. The patient is then transferred to a recovery area, where they are monitored as they wake up from anesthesia. MIS finds applications in various medical specialties, including general surgery, gynecology, urology, orthopedics, and more [12].

While MIS offers numerous benefits, it also presents several challenges stemming from the intricate nature of the surgical procedure, limited field of view, complex hand-eye coordination requirements, and the involvement of human assistants. These challenges can lead to increased surgical time and costs. The complexity of performing intricate procedures through small incisions, sometimes with limited tactile feedback compared to open surgery, can be especially challenging when dealing with delicate anatomical structures. The use of small incisions and an endoscope reduces the surgeon's field of view compared to open surgery, making it challenging to navigate and manipulate tissues effectively [13]. MIS necessitates that surgeons develop and maintain complex hand-eye coordination skills, translating external movements into precise actions within the body, often using instruments with articulating tips [14]. In many MIS procedures, a human assistant operates the endoscope's camera, providing the surgeon with visual feedback, which introduces an element of dependency on the assistant's skills and can affect surgery efficiency [15]. To address these challenges, there has been a noticeable surge of interest in the research domain of computer- and robot-assisted minimally invasive surgery (RAMIS) over the past few years. This heightened focus aims to enhance the capabilities of minimally invasive procedures, mitigate associated difficulties, and further improve patient outcomes, with researchers, clinicians, and healthcare institutions investing in exploring RAMIS's potential benefits and advancements [16,17].

Advancements in this domain revolve around providing surgeons with effective tools to address and mitigate these challenges and open up exciting possibilities in surgical skill assessment, workflow optimization, and training of junior surgeons. One prominent approach involves harnessing information related to the positions and movements of surgical instruments during a procedure, often utilizing innovative tracking methods that enable real-time monitoring of surgical tools. These tracking methods, relying on technologies such as electromagnetic- or infrared-based systems or the strategic attachment of external markers to instruments, precisely capture instrument locations and movements within the surgical field, enhancing the surgeon's capabilities and offering insights into surgical practice [18–21]. Presently, the primary focus in this domain is on utilizing visually derived data from endoscopic video streams, aligning with the advancements in deep-learning-based techniques within image processing [22,23].

The task at hand involves identifying and tracking surgical instruments within the surgical field, typically accomplished using object detection methods that locate instruments by enclosing them within bounding boxes. Once instruments are initially detected, tracking methods follow them across multiple video frames, ensuring consistent monitoring and recording of instrument positions and movements [24,25]. While these methods offer

quick processing times, they may lack precision, particularly for instruments that extend from the bottom corners toward the center of the image. A significant advancement in surgical instrument detection and localization is achieved through segmentation methods, which provide a finer level of detail by predicting instrument shapes pixel by pixel. Unlike bounding box methods, segmentation offers a more accurate representation of instrument contours and boundaries [26–29]. A range of studies have explored surgical tool segmentation on the Johns Hopkins University (JHU) and Intuitive Surgical, Inc. (Sunnyvale, CA, USA. ISI) Gesture and Skill Assessment Working Set, JIGSAWS dataset, for autonomous image-based skill assessment. Papp [30] achieved promising results using TernausNet-11, while Ahmidi [31] reported high accuracy for gesture recognition techniques. Funke [32] demonstrated the feasibility of deep-learning-based skill assessment using 3D convolutional networks (ConvNets), and Lajkó [33] proposed a 2D image-based skill assessment method. Nema [34] and Jin [35] discussed the use of instrument detection and tracking technologies, with Jin introducing a new dataset and method for tool presence detection and spatial localization. Attia [36] achieved high accuracy in surgical tool segmentation using a hybrid deep convolutional neural network–recurrent neural network (CNN-RNN) auto-encoder–decoder. These studies collectively highlight the potential of various techniques for surgical tool segmentation and skill assessment on the JIGSAWS dataset.

This paper focuses on the application of the U-Net architecture, a CNN structure originally designed for precise pixel-wise image classification in medical image analysis tasks [26–28,37,38]. TernausNet, a U-Net architecture with a VGG11 pre-trained encoder, has shown superior performance in image segmentation tasks, particularly in the medical and satellite imaging domain [39]. This approach has been further extended to TernausNetV2, which allows for instance segmentation in high-resolution satellite imagery. The U-Net architecture, in general, has been widely adopted in medical image analysis, with various modifications and improvements proposed. These include the double U-Net, which combines two U-Net architectures and has shown improved performance in medical image segmentation [40], and UNet++, a nested U-Net architecture that has achieved significant gains in medical image segmentation tasks [41]. The U-NetPlus, a modified U-Net architecture with a pre-trained encoder, has also demonstrated superior performance in semantic and instance segmentation of surgical instruments from laparoscopic images [42]. The U-Net architecture's potential is further enhanced by the integration of residual skip connections and recurrent feedback with a pre-trained EfficientNet encoder, resulting in improved segmentation performance [43]. However, the U-Net architecture has some common limitations, including limited receptive field, susceptibility to overfitting, challenges with imbalanced class distributions, difficulty handling irregular shapes, computational complexity, and issues with adapting to multiple data modalities [44–49]. To address these limitations and increase segmentation accuracy for surgical instruments, this paper proposes the integration of advanced augmentation techniques like MixUp, CutMix, or GridMask, with a specific focus on the GridMask technique.

In summary, MIS offers numerous advantages but presents unique challenges, leading to increased interest in computer RAMIS. Advancements in this field aim to enhance surgical capabilities, optimize workflows, and train junior surgeons. Tracking and real-time monitoring of surgical instruments play a crucial role in addressing these challenges, with segmentation methods offering higher precision. This paper focuses on the U-Net architecture's application to surgical instrument segmentation, aiming to overcome its limitations by incorporating advanced augmentation techniques like GridMask.

2. Materials and Methods

Image augmentation techniques that rely on image erasure typically involve the removal of one or more specific portions within an image. The fundamental concept behind this approach is to substitute the pixel values in these removed regions with either constant or randomized values. In a study by DeVries and Taylor in 2017, they introduced a straightforward regularization method called "cutout." This method entails

randomly masking square regions within input data during the training of convolutional neural networks (CNNs). Cutout has been shown to enhance the resilience and overall performance of CNNs [50]. Another technique, proposed by Singh and others in 2018, is known as "Hide-and-Seek" (HaS). HaS involves randomly concealing patches within training images, encouraging the network to seek out other relevant information when crucial content is hidden [51]. In 2020, Zhong and his colleagues introduced the "random erasing" method. This technique randomly selects a rectangular area within an image and replaces its pixel values with random data. Despite its simplicity, this approach has demonstrated significant performance improvements [52]. A more recent development, presented by Chen et al. in 2020, delves into the necessity of information reduction and introduces a structured method called "GridMask". GridMask is also based on the removal of regions within input images [53]. Unlike Cutout and HaS, GridMask does not eliminate continuous regions or randomly select squares; instead, it removes uniformly distributed square regions, allowing for control over their density and size. To address the trade-off between object occlusion and information retention, FenceMask, as proposed by Li et al. in 2020, employs a strategy simulating object occlusion. This method adds to the array of techniques based on the deletion of specific regions within input images [54].

GridMask is a data augmentation technique designed to enhance the performance of deep learning models, particularly in computer vision tasks [53]. The method involves overlaying a grid on the image and masking (or dropping out) certain regions, forcing the model to learn from a partial view of the data. The idea is akin to the way dropout works for neurons but is applied spatially on input images. Just like dropout prevents neurons from co-adapting and helps in regularization, GridMask ensures that the model does not overly rely on specific local features or pixels of the input. By dropping out certain sections of the image, the network is forced to learn more robust and generalized features. By masking out portions of the image, the model is pushed to use the surrounding context to make predictions about the masked regions. This is especially useful for segmentation tasks where understanding the context can be crucial. In real-world scenarios, the objects or regions of interest in images may be partially occluded. GridMask trains the model to handle such occlusions, making it more robust. U-Net, with its large number of parameters, can be prone to overfitting, especially when the dataset is limited. Data augmentation techniques like GridMask can effectively increase the size of the training dataset by providing varied versions of the same image, helping to reduce overfitting. GridMask forces the U-Net to focus on both local and global features. While the local features within the unmasked regions become more pronounced, the model also tries to infer global context from the available parts of the image. Data augmentation techniques often help in better convergence during training. By providing more varied data, GridMask can potentially smoothen the loss landscape and assist in more stable training. GridMask allows for random rotations, resizing, and shifting of the grid, leading to a wide range of augmentations from a single image. This ensures that the network encounters varied challenges during training, pushing it to learn a broader set of features. Incorporating GridMask into the U-Net training pipeline can be straightforward. It is essential to ensure that the augmentations are applied consistently to both input images and their corresponding masks/annotations during training, especially for segmentation tasks. As with any augmentation technique, it is advisable to monitor validation performance to ensure the augmentations lead to genuine improvements and not just make the task harder without yielding benefits.

2.1. Enhanced U-Net with GridMask (EUGNet) Architecture

To address the inherent challenges associated with the traditional U-Net architecture and to harness the potential benefits of the GridMask augmentation technique, we introduce the enhanced U-Net with GridMask (EUGNet) architecture. The following subsections detail the components and innovations of EUGNet:

1. Deep Contextual Encoder: To capture distant contextual information, which the traditional U-Net might miss, our encoder is deepened and incorporates dilated con-

volutions. This enhancement allows for a broader receptive field without a significant increase in computational complexity.

2. Residual Connections: To mitigate the loss of fine-grained spatial details during the downsampling and upsampling processes, we have integrated residual connections between corresponding encoder and decoder layers. This integration ensures the preservation of spatial information, aiding in more accurate segmentation output reconstruction.
3. Class Balancing Loss: Considering the frequent challenge of imbalanced class distributions in medical image analysis, our architecture employs a class-balancing loss function. This adjustment ensures that the model remains unbiased towards the majority class, providing equal importance to all classes during training.
4. Adaptive Feature Fusion: To better handle objects of irregular shapes, we introduce an adaptive feature fusion mechanism within the decoder. This mechanism adaptively weighs features from the encoder and the upsampled features from the preceding decoder layer, allowing the model to focus on the most pertinent features for segmentation.
5. GridMask Augmentation Module: The GridMask technique is directly integrated into our training pipeline. Before the images are input into the encoder, they undergo the GridMask module, ensuring the model consistently trains with augmented data, enhancing its robustness and reducing overfitting tendencies.
6. Efficient Implementation: To address U-Net's computational demands, our architecture employs depthwise separable convolutions where feasible. This approach reduces the parameter count without compromising the model's learning capacity.
7. Multi-Modal Fusion: For tasks that involve multiple data modalities, EUGNet introduces a fusion layer post-encoder. This layer is designed to effectively fuse features from different modalities before they are passed to the decoder. Figure 1 depicts a visual representation of EUGNet.

2.2. GridMask Algorithm

GridMask is a straightforward, versatile, and effective technique. When provided with an input image, our algorithm randomly eliminates certain pixels from it. In contrast to other approaches, our algorithm's removal region is distinct in that it does not consist of continuous pixel clusters or randomly scattered pixels as in dropout. Instead, it removes a region made up of disconnected sets of pixels. The setting can be expressed as follows [53]:

$$\widetilde{X} = X \times M \tag{1}$$

where $X \in R^{H \times W \times C}$ represents the input image, $M \in \{0,1\}^{H \times W}$ is the binary mask that stores pixels to be removed, and $\widetilde{X} \in R^{H \times W \times C}$ is the result produced by our algorithm. $R^{H \times W \times C}$ represents a 3-dimensional space for an image where H stands for the height of the image in pixels, W stands for the width of the image in pixels, and C stands for the number of channels in the image. For the binary mask M, if $M_{i,j} = 1$, we keep pixel (i, j) in the input image; otherwise, it will be removed. The algorithm is applied after the image normalization operation.

The shape of M looks like a grid, as shown in Figure 2. Four numbers $(r, d, \delta_x, \delta_y)$ are used to represent a unique M. Every mask is formed by tiling the units, as shown in Figure 3. Here, r is the ratio of the shorter gray edge in a unit, and d is the length of one unit. δ_x and δ_y are the distances between the first intact unit and the boundary of the image.

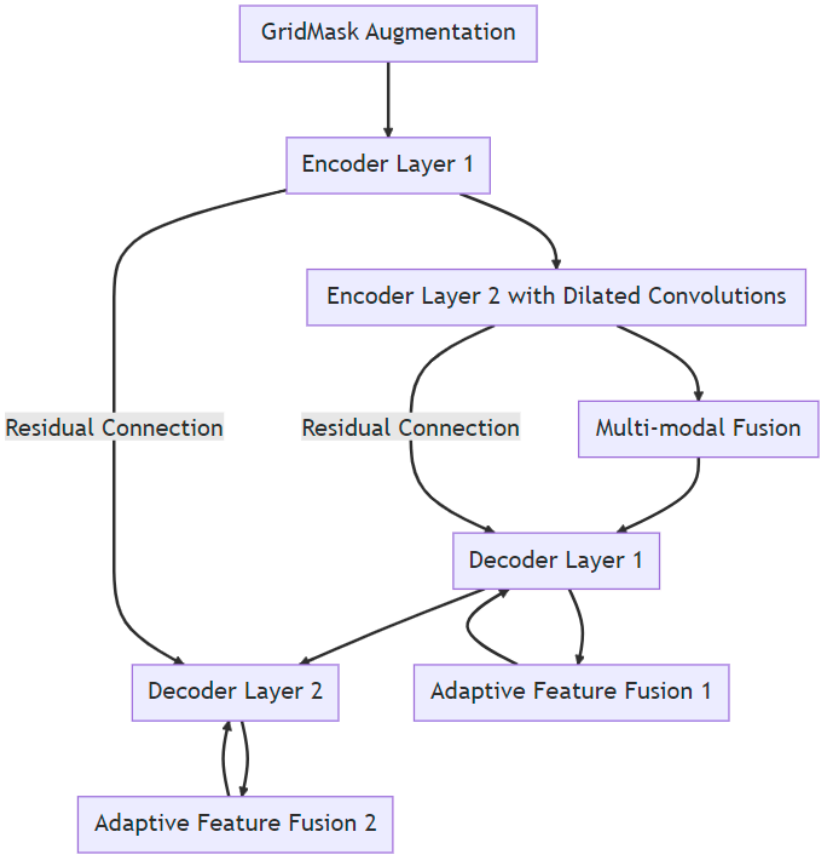

Figure 1. Visual representation of EUGNet.

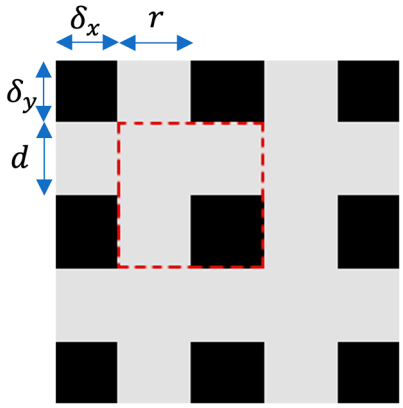

Figure 2. A single cell of the GridMask is illustrated by the red-dashed square.

2.3. Data Collection for Algorithm Evaluation

To show the robustness and generalization ability of the proposed framework for robotic instrument segmentation, a dataset with different robotic surgical scenarios and instruments has been used to validate the proposed architectures. This dataset consists of training and testing data for ex vivo robotic surgery with different articulated instruments:

(1) Da Vinci robotic (DVR) dataset [55]: The training set contains four ex vivo 45 s videos. The testing set consists of four 15 s and two 60 s videos. The test video features contain two instruments. Articulated motions are present in all the videos. The ground truth masks are automatically generated using joint encoder information and forward kinematics. Hand-eye calibration errors are manually corrected. All the videos have been recorded with the da Vinci Research Kit (dVRK) open-source platform [56]. The

frames have a resolution of 720 × 576, and the videos run at 25 frames per second. This means that we have a total of (60 + 15) × 25 = 1875 frames (images).

(2) We obtained the recorded videos for testing our algorithm from open sources on the Internet, including the U.S. National Library of Medicine [57] (video links are available upon request). The videos showed various surgical procedures, such as midline lobectomy, right superior line lobectomy, thoracotomy, thoracoscopic lung surgery, and prostatectomy. Each video showed splash-like bleeding. The frames have a resolution of 720 × 576, and the videos run at 25 frames per second. The total duration of these videos is 2 min, which means 2 × 60 × 25 = 3000 frames (images).

(3) The binary segmentation EndoVis 17 [58] dataset, comprising 600 images, was utilized for both testing and training purposes. It consists of 10 sequences from abdominal porcine procedures recorded with da Vinci Xi systems. The dataset was curated by selecting active sequences that exhibited substantial instrument motion and visibility, with 300 frames sampled at a 1 Hz rate from each procedure. Frames where the instrument motion was absent for an extended period were manually excluded to maintain a consistent sequence of 300 frames. For the purpose of training, the first 225 frames from 8 sequences were made available, while the remaining 75 frames of these sequences were reserved for testing. Additionally, 2 sequences, each with a complete set of 300 frames, were exclusively allocated for testing.

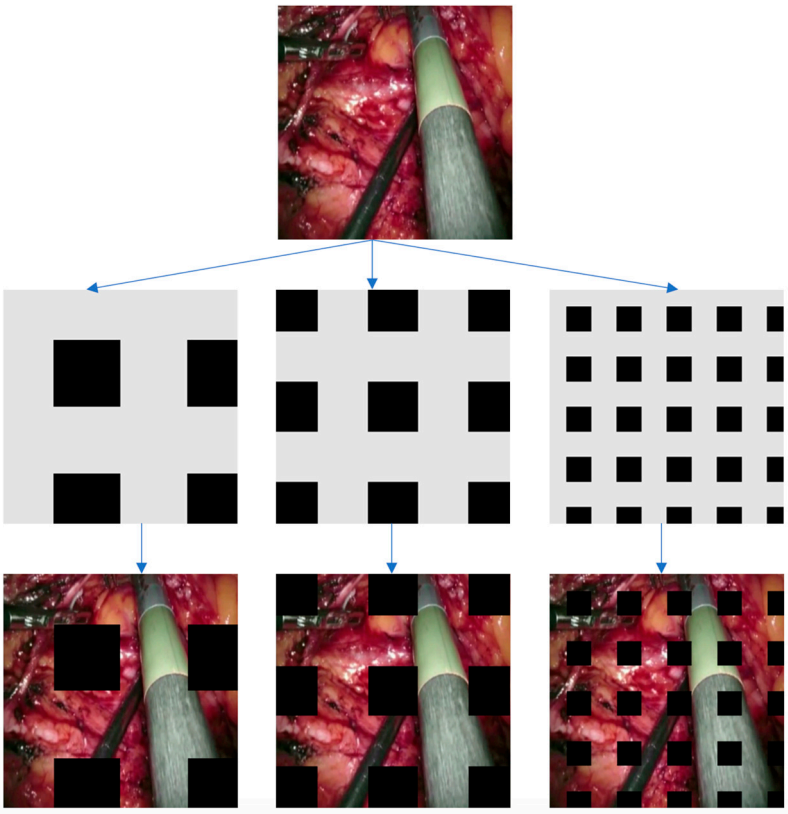

Figure 3. This illustration displays instances of GridMask in action. Initially, we generate a mask based on the specified parameters (r, d, δ_x, δ_y). Subsequently, we apply this mask to the input image through multiplication. The outcome is presented in the final image.

2.4. Baseline Method and Evaluation Protocol

U-Net is a widely adopted tool in the field of medical image analysis, especially for segmenting surgical instruments in medical images and videos. To establish a baseline for comparison, we employ an advanced version of the U-Net architecture, known for its precision in segmenting robotic surgical tools. This choice is natural because the fine-grained U-Net architecture represents one of the cutting-edge convolutional models for this

specific task. The frames are randomly chosen during training to present the networks with varying input data, as we are mostly interested in comparing the proposed architecture to the baseline method rather than achieving the highest scores, and GridMask data augmentation is performed. Because transfer learning carries the risk of transferring biases present in the source dataset to our target task, transfer learning is neglected. Since biases are undesirable for your application, training from scratch was preferred. In our experiments, the cyclical learning rate (CLR) bounds for the U-Net network are set to (10^{-4}; 10^{-2}. The quantitative metrics of choice to evaluate the predicted segmentations are mean Intersection over union (mean *IoU*) and mean Dice similarity coefficient (mean *DSC*):

$$\overline{IoU}(\hat{y}, y) = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k}{TP_k + FP_k + NF_k} \tag{2}$$

$$\overline{DSC}(\hat{y}, y) = \frac{1}{K}\sum_{k=1}^{K}\frac{2TP_k}{2TP_k + FP_k + NF_k} \tag{3}$$

where $K = 2$ ($k = 0$ background, $k = 1$ foreground), and TP_k, FP_k, and NF_k represent true positives, false positives, and false negatives for class k, respectively.

All networks were trained and tested (including inference times) on a computer with a 13th generation Intel Core™ i9-13900KF processor (E-cores up to 4.30 GHz and P-cores up to 5.40 GHz) CPU and a NVIDIA GeForce RTX™ 4080 16GB GDDR6X GPU. The inference time was calculated, including data transfers from CPU to GPU and back, and averaged across 1000 inferences.

3. Results

Comparing the results of our proposed augmented data employing GridMask with the U-Net in the DVR testing set, we observed improvements of 1.6, 1.7, and 1.7 percentage points in balanced accuracy (foreground), mean intersection over union (IoU), and mean DSC, respectively (see Table 1). This has a significant impact on inference speed, which is reduced from 0.163 ms for the U-Net without GridMask to 0.097 ms or the U-Net with GridMask, becoming a viable real-time instrument–tissue segmentation method for robotic surgery with the da Vinci platform.

Table 1. Quantitative results for segmentation of non-rigid robotic instruments in testing set videos. IoU stands for intersection over union, and DSC for Dice similarity coefficient. The means are performed over classes, and the results presented are averaged across testing frames.

Network	Inference Time (ms/fps)	Balanced Accuracy (fg.)	Mean IoU	Mean DSC
U-Net without GridMask	62.1/16.1	82.5%	78.2%	84.2%
U-Net with GridMask	34.2/29.2	86.3%	80.6%	89.5%

The results of the U-Net with the GridMask method for data augmentation show an improvement over the U-Net without GridMask of 4.6, 2.9, and 6.2 percentage points in balanced accuracy (foreground), mean IoU, and mean DSC, respectively (see Table 1). The inference speed is also real-time, approximately 29 fps. The qualitative results in Figure 4 show how our proposed architecture respects the borders of the tooltip of left-handed surgical tools more.

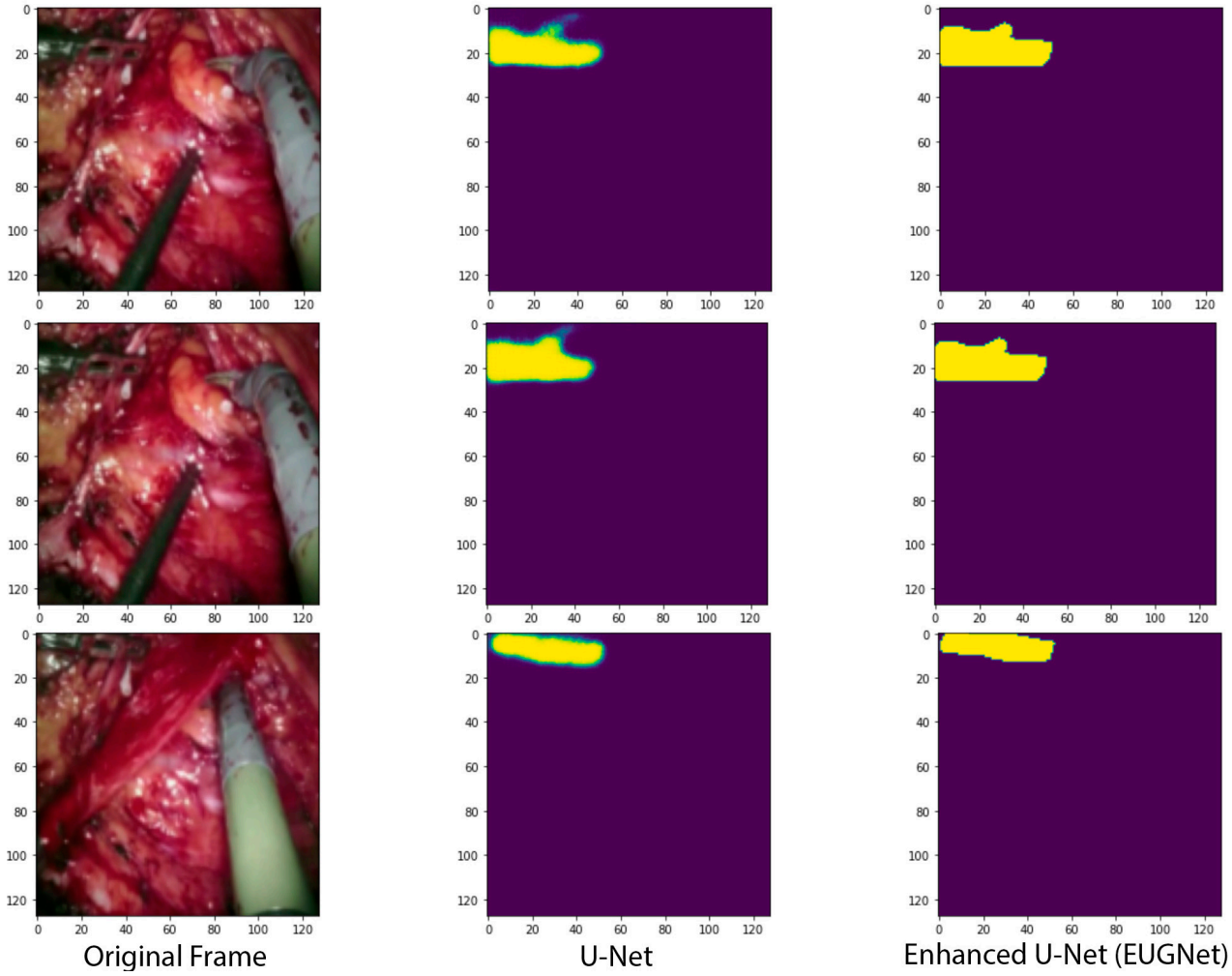

Figure 4. Qualitative comparison of our proposed convolutional architectures with GridMask data augmentation versus the state-of-the-art U-Net network.

As can be seen in Figure 5, the application of GridMask data augmentation in the training of an enhanced U-Net model has demonstrated notable improvements across various metrics. Specifically, the graphs provided illustrate that incorporating GridMask results in a more stable and generally lower training loss over 40 epochs, suggesting better generalization and less overfitting compared to training without GridMask. Accuracy metrics also show an improvement, with the model achieving comparable or slightly higher accuracy when trained with GridMask, implying that the model's predictions are more reliable. Furthermore, the Dice coefficient, which is crucial for evaluating the model's performance in segmentation tasks, shows a clear enhancement when GridMask is utilized. The model with GridMask maintains a consistently higher Dice coefficient, indicating superior overlap between the predicted and ground-truth segmentation masks. These results collectively suggest that the use of GridMask data augmentation can significantly bolster the performance of enhanced U-Net architectures in learning tasks.

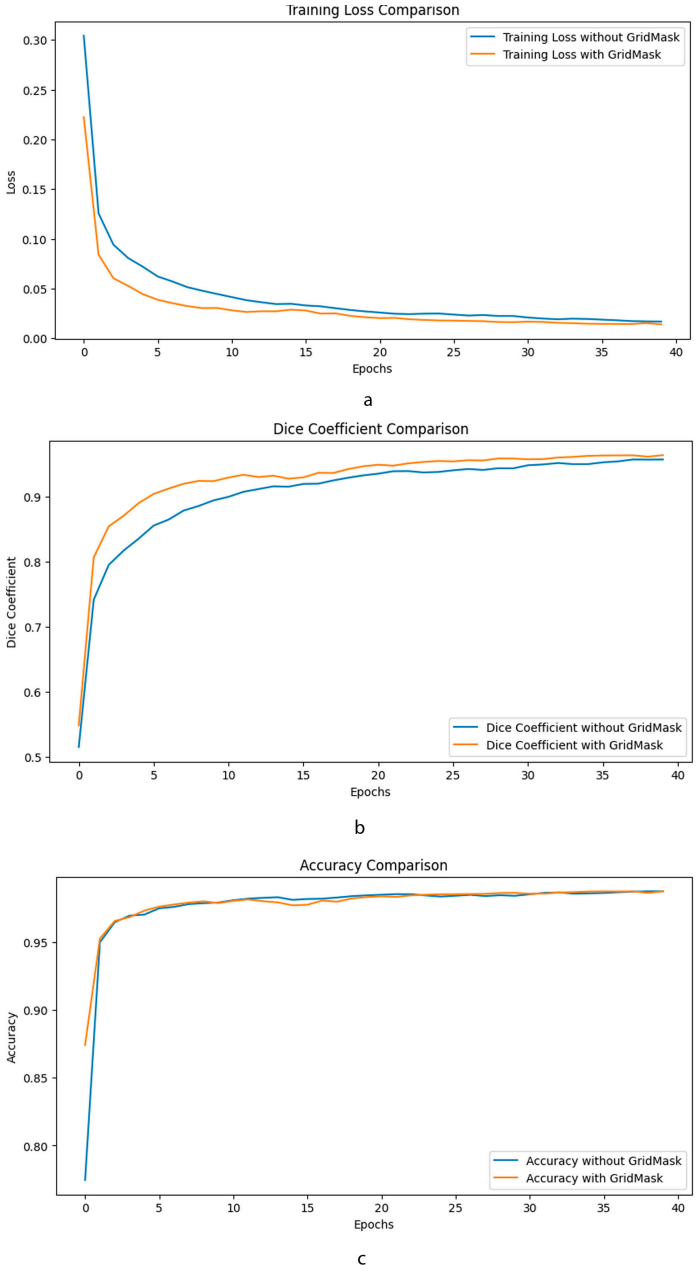

Figure 5. Comparative performance metrics of enhanced U-Net with GridMask data augmentation. This composite image showcases three key performance indicators—training loss (**a**), Dice coefficient (**b**), and accuracy (**c**)—over 40 epochs, comparing the outcomes of an enhanced U-Net.

4. Discussion

In the forthcoming stages of our research, we aim to refine our approach to using GridMask for data augmentation. Instead of uniformly applying GridMask to all training images, we intend to implement a more precise strategy. This refined approach involves overlaying masks exclusively onto the segmented regions of the images while excluding the background. By doing so, we aim to create a more targeted and effective data augmentation process. Furthermore, we plan to conduct an in-depth evaluation and comparison of the effectiveness of alternative augmentation techniques such as MixUp and CutMix. This comparative analysis will help us determine which augmentation strategy yields the most favorable results for our specific task. In our ongoing research, we will also explore the potential advantages of transfer learning, potentially harnessing the capabilities of deeper neural network architectures. Transfer learning involves leveraging pre-trained models

to expedite training and potentially improve the performance of our model in surgical instrument segmentation.

The GridMask data augmentation technique led to performance improvement in the segmentation of surgical tools by introducing structured occlusion in training images. This encourages the U-Net model to learn more robust features by (a) forcing contextual learning: by partially occluding the surgical tools, the network must learn to infer the shape and position of tools from the visible context. (b) Improving generalization: GridMask helps in reducing overfitting by preventing the network from relying on specific visual cues that are only present in the training data. (c) Enhancing feature learning: it encourages the network to learn more comprehensive feature representations by not depending on any particular region of the tool, leading to a more versatile understanding of the tool's appearance. (d) Simulating real-world occlusions: in a surgical environment, tools may be occluded by various objects, including human hands, tissues, or other tools.

By integrating these challenging scenarios during training, GridMask effectively enhances the robustness of the U-Net model for surgical tool segmentation tasks.

It should be emphasized that in the training of U-Net, standard data augmentation techniques are employed. These include rotation, flipping, scaling, translation, elastic deformation, brightness adjustment, noise injection, and cropping. The intent of our study is to evaluate the effectiveness of GridMask data augmentation in comparison to these widely used techniques. During the data preparation phase of training, we implement the aforementioned common augmentation strategies before proceeding to train the model with GridMask techniques. Therefore, our comparisons are not made against models without any data augmentation but against those that have been trained with the standard set of augmentation methods, excluding GridMask.

Additionally, we are considering the application of adversarial training as part of our research. This technique, which has demonstrated promising outcomes in broader semantic labeling applications, may play a role in further enhancing the precision and accuracy of our robotic surgical instrument segmentation model. One particularly captivating avenue for future investigation is the three-dimensional (3D) localization of segmented surgical instruments. With the growing prevalence of stereo endoscopes, we have the opportunity to segment both images produced by these devices. This dual-image segmentation enables us to align instrument pixels between the stereo images. By ensuring proper camera calibration, we can leverage these matched points for dense geometry-based triangulation, ultimately providing precise 3D localization information for surgical tools. This advancement could greatly enhance the spatial understanding and guidance of surgical procedures.

5. Conclusions

We have improved the current state-of-the-art U-Net architecture by employing GridMask data augmentation techniques. When we compare the outcomes of our novel data augmentation approach, which incorporates GridMask, with the U-Net model on the DVR testing dataset, we observe notable enhancements. Specifically, there are improvements of 1.6 percentage points in balanced accuracy for the foreground, 1.7 points in IoU, and 1.7 points in mean DSC. These improvements are highly significant and have a substantial impact on inference speed. The inference speed, which is a critical factor in real-time applications, has seen a noteworthy reduction. It decreased from 0.163 milliseconds for the U-Net without GridMask to 0.097 milliseconds for the U-Net with GridMask. This substantial improvement in inference speed positions our model as a viable real-time instrument–tissue segmentation method for robotic surgery, especially when deployed on the da Vinci platform.

Moreover, when we examine the results of the U-Net model with GridMask for data augmentation, we witness even more impressive performance gains compared to the U-Net with GridMask alone. There are remarkable improvements of 4.6 percentage points in balanced accuracy for the foreground, 2.9 points in mean IoU, and 6.2 points in mean

DSC, as highlighted in Table 1. These improvements further underscore the efficacy of our approach.

Additionally, the inference speed remains in real time, operating at an approximate rate of 29 frames per second (fps). This fast processing rate is essential for the dynamic and time-sensitive nature of robotic surgical procedures.

Author Contributions: Conceptualization, M.D.R. and S.Z.M.M.; methodology, M.D.R. and S.Z.M.M.; software, M.D.R.; validation, M.D.R.; formal analysis, M.D.R. and S.Z.M.M.; investigation, M.D.R. and S.Z.M.M.; resources, M.D.R.; data curation, M.D.R.; writing—original draft preparation, M.D.R. and S.Z.M.M.; writing—review and editing, M.D.R. and S.Z.M.M.; visualization, M.D.R.; supervision, M.D.R. and S.Z.M.M.; project administration, M.D.R.; funding acquisition, M.D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported with funding from the College of Engineering Research Seed Grant Program at Lawrence Technological University, award No. 5000.

Data Availability Statement: The data presented in this study are available at https://universe.roboflow.com/models/instance-segmentation (accessed on 12 September 2023) and https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6462551/figure/vid/ (accessed on 14 August 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gagner, M. Laparoscopic adrenalectomy in Cushing's syndrome and pheochromocytoma. *N. Engl. J. Med.* **1992**, *327*, 1033. [PubMed]
2. Berger, R.A.; Jacobs, J.J.; Meneghini, R.M.; Della Valle, C.; Paprosky, W.; Rosenberg, A.G. Rapid rehabilitation and recovery with minimally invasive total hip arthroplasty. *Clin. Orthop. Relat. Res.* **2004**, *429*, 239–247. [CrossRef] [PubMed]
3. Kehlet, H.; Wilmore, D.W. Evidence-based surgical care and the evolution of fast-track surgery. *Ann. Surg.* **2008**, *248*, 189–198. [CrossRef] [PubMed]
4. Darzi, A.; Mackay, S. Recent advances in minimal access surgery. *Bmj* **2002**, *324*, 31–34. [CrossRef] [PubMed]
5. Maurus, C.F.; Schäfer, M.; Müller, M.K.; Clavien, P.-A.; Weber, M. Laparoscopic versus open splenectomy for nontraumatic diseases. *World J. Surg.* **2008**, *32*, 2444–2449. [CrossRef] [PubMed]
6. Khorgami, Z.; Haskins, I.N.; Aminian, A.; Andalib, A.; Rosen, M.J.; Brethauer, S.A.; Schauer, P.R. Concurrent ventral hernia repair in patients undergoing laparoscopic bariatric surgery: A case-matched study using the National Surgical Quality Improvement Program Database. *Surg. Obes. Relat. Dis.* **2017**, *13*, 997–1002. [CrossRef] [PubMed]
7. Pollard, J.S.; Fung, A.K.-Y.; Ahmed, I. Are natural orifice transluminal endoscopic surgery and single-incision surgery viable techniques for cholecystectomy? *J. Laparoendosc. Adv. Surg. Tech.* **2012**, *22*, 1–14. [CrossRef] [PubMed]
8. Stefanidis, D.; Fanelli, R.D.; Price, R.; Richardson, W.; Committee, S.G. SAGES guidelines for the introduction of new technology and techniques. *Surg. Endosc.* **2014**, *28*, 2257–2271. [CrossRef]
9. Gifari, M.W.; Naghibi, H.; Stramigioli, S.; Abayazid, M. A review on recent advances in soft surgical robots for endoscopic applications. *Int. J. Med. Robot. Comput. Assist. Surg.* **2019**, *15*, e2010. [CrossRef]
10. Somashekhar, S.; Acharya, R.; Saklani, A.; Parikh, D.; Goud, J.; Dixit, J.; Gopinath, K.; Kumar, M.V.; Bhojwani, R.; Nayak, S. Adaptations and safety modifications to perform safe minimal access surgery (MIS: Laparoscopy and Robotic) during the COVID-19 pandemic: Practice modifications expert panel consensus guidelines from Academia of Minimal Access Surgical Oncology (AMASO). *Indian J. Surg. Oncol.* **2021**, *12*, 210–220. [CrossRef]
11. Vitiello, V.; Lee, S.-L.; Cundy, T.P.; Yang, G.-Z. Emerging robotic platforms for minimally invasive surgery. *IEEE Rev. Biomed. Eng.* **2012**, *6*, 111–126. [CrossRef] [PubMed]
12. Cohn, L.H.; Adams, D.H.; Couper, G.S.; Bichell, D.P.; Rosborough, D.M.; Sears, S.P.; Aranki, S.F. Minimally invasive cardiac valve surgery improves patient satisfaction while reducing costs of cardiac valve replacement and repair. *Ann. Surg.* **1997**, *226*, 421. [CrossRef] [PubMed]
13. Link, R.E.; Bhayani, S.B.; Kavoussi, L.R. A prospective comparison of robotic and laparoscopic pyeloplasty. *Ann. Surg.* **2006**, *243*, 486. [CrossRef] [PubMed]
14. Schijven, M.; Jakimowicz, J.; Broeders, I.; Tseng, L. The Eindhoven laparoscopic cholecystectomy training course—Improving operating room performance using virtual reality training: Results from the first EAES accredited virtual reality trainings curriculum. *Surg. Endosc. Other Interv. Tech.* **2005**, *19*, 1220–1226. [CrossRef] [PubMed]
15. Blavier, A.; Gaudissart, Q.; Cadière, G.-B.; Nyssen, A.-S. Comparison of learning curves and skill transfer between classical and robotic laparoscopy according to the viewing conditions: Implications for training. *Am. J. Surg.* **2007**, *194*, 115–121. [CrossRef] [PubMed]
16. Haidegger, T.; Speidel, S.; Stoyanov, D.; Satava, R.M. Robot-assisted minimally invasive surgery—Surgical robotics in the data age. *Proc. IEEE* **2022**, *110*, 835–846. [CrossRef]

17. Maier-Hein, L.; Eisenmann, M.; Sarikaya, D.; März, K.; Collins, T.; Malpani, A.; Fallert, J.; Feussner, H.; Giannarou, S.; Mascagni, P. Surgical data science–from concepts toward clinical translation. *Med. Image Anal.* **2022**, *76*, 102306. [CrossRef]
18. Bouarfa, L.; Akman, O.; Schneider, A.; Jonker, P.P.; Dankelman, J. In-vivo real-time tracking of surgical instruments in endoscopic video. *Minim. Invasive Ther. Allied Technol.* **2012**, *21*, 129–134. [CrossRef]
19. Mamone, V.; Viglialoro, R.M.; Cutolo, F.; Cavallo, F.; Guadagni, S.; Ferrari, V. Robust Laparoscopic Instruments Tracking Using Colored Strips. In Proceedings of the Augmented Reality, Virtual Reality, and Computer Graphics: 4th International Conference, AVR 2017, Ugento, Italy, 12–15 June 2017; Proceedings, Part II 4. Springer: Berlin/Heidelberg, Germany, 2017.
20. Sorriento, A.; Porfido, M.B.; Mazzoleni, S.; Calvosa, G.; Tenucci, M.; Ciuti, G.; Dario, P. Optical and electromagnetic tracking systems for biomedical applications: A critical review on potentialities and limitations. *IEEE Rev. Biomed. Eng.* **2019**, *13*, 212–232. [CrossRef]
21. Wang, Y.; Sun, Q.; Liu, Z.; Gu, L. Visual detection and tracking algorithms for minimally invasive surgical instruments: A comprehensive review of the state-of-the-art. *Robot. Auton. Syst.* **2022**, *149*, 103945. [CrossRef]
22. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.; Van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
23. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* **2016**, *316*, 2402–2410. [CrossRef] [PubMed]
24. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
25. Danelljan, M.; Häger, G.; Khan, F.; Felsberg, M. Accurate Scale Estimation for Robust Visual Tracking. In Proceedings of the British Machine Vision Conference, Nottingham, UK, 1–5 September 2014; BMVA Press: Durham, UK, 2014.
26. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18. Springer: Berlin/Heidelberg, Germany, 2015.
27. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; IEEE: Stanford, CA, USA, 2016.
28. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, 17–21 October 2016; Proceedings, Part II 19. Springer: Berlin/Heidelberg, Germany, 2016.
29. Kamnitsas, K.; Ledig, C.; Newcombe, V.F.; Simpson, J.P.; Kane, A.D.; Menon, D.K.; Rueckert, D.; Glocker, B. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* **2017**, *36*, 61–78. [CrossRef] [PubMed]
30. Papp, D.; Elek, R.N.; Haidegger, T. Surgical Tool Segmentation on the Jigsaws Dataset for Autonomous Image-Based Skill Assessment. In Proceedings of the 2022 IEEE 10th Jubilee International Conference on Computational Cybernetics and Cyber-Medical Systems (ICCC), Reykjavík, Iceland, 6–9 July 2022; IEEE: Reykjavík, Iceland, 2022.
31. Ahmidi, N.; Tao, L.; Sefati, S.; Gao, Y.; Lea, C.; Haro, B.B.; Zappella, L.; Khudanpur, S.; Vidal, R.; Hager, G.D. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 2025–2041. [CrossRef] [PubMed]
32. Funke, I.; Mees, S.T.; Weitz, J.; Speidel, S. Video-based surgical skill assessment using 3D convolutional neural networks. *Int. J. Comput. Assist. Radiol. Surg.* **2019**, *14*, 1217–1225. [CrossRef] [PubMed]
33. Lajkó, G.; Nagyne Elek, R.; Haidegger, T. Endoscopic image-based skill assessment in robot-assisted minimally invasive surgery. *Sensors* **2021**, *21*, 5412. [CrossRef] [PubMed]
34. Nema, S.; Vachhani, L. Surgical instrument detection and tracking technologies: Automating dataset labeling for surgical skill assessment. *Front. Robot. AI* **2022**, *9*, 1030846. [CrossRef] [PubMed]
35. Jin, A.; Yeung, S.; Jopling, J.; Krause, J.; Azagury, D.; Milstein, A.; Fei-Fei, L. Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; IEEE: Lake Tahoe, NV, USA, 2018.
36. Attia, M.; Hossny, M.; Nahavandi, S.; Asadi, H. Surgical Tool Segmentation Using a Hybrid Deep CNN-RNN Auto Encoder-Decoder. In Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, Canada, 5–8 October 2017; IEEE: Banff, AB, Canada, 2017.
37. Isensee, F.; Jaeger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]
38. Falk, T.; Mai, D.; Bensch, R.; Çiçek, Ö.; Abdulkadir, A.; Marrakchi, Y.; Böhm, A.; Deubner, J.; Jäckel, Z.; Seiwald, K. U-Net: Deep learning for cell counting, detection, and morphometry. *Nat. Methods* **2019**, *16*, 67–70. [CrossRef]
39. Iglovikov, V.; Shvets, A. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv* **2018**, arXiv:1801.05746.

40. Jha, D.; Riegler, M.A.; Johansen, D.; Halvorsen, P.; Johansen, H.D. Doubleu-Net: A Deep Convolutional Neural Network for Medical Image Segmentation. In Proceedings of the 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, MN, USA, 28–30 July 2020; IEEE: Rochester, MN, USA, 2020.
41. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A Nested U-Net Architecture for Medical Image Segmentation. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018; Proceedings 4. Springer: Berlin/Heidelberg, Germany, 2018.
42. Hasan, S.K.; Linte, C.A. U-NetPlus: A Modified Encoder-Decoder U-Net Architecture for Semantic and Instance Segmentation of Surgical Instruments from Laparoscopic Images. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; IEEE: Berlin, Germany, 2019.
43. Siddique, N. U-Net Based Deep Learning Architectures for Object Segmentation in Biomedical Images. Doctoral Dissertation, Purdue University Graduate School, West Lafayette, IN, USA, 2021.
44. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
45. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
46. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Jorge Cardoso, M. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. In Proceedings of the Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, 14 September 2017; Proceedings 3. Springer: Berlin/Heidelberg, Germany, 2017.
47. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
48. Havaei, M.; Davy, A.; Warde-Farley, D.; Biard, A.; Courville, A.; Bengio, Y.; Pal, C.; Jodoin, P.-M.; Larochelle, H. Brain tumor segmentation with deep neural networks. *Med. Image Anal.* **2017**, *35*, 18–31. [CrossRef] [PubMed]
49. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
50. DeVries, T.; Taylor, G.W. Improved regularization of convolutional neural networks with cutout. *arXiv* **2017**, arXiv:1708.04552.
51. Singh, K.K.; Yu, H.; Sarmasi, A.; Pradeep, G.; Lee, Y.J. Hide-and-seek: A data augmentation technique for weakly-supervised localization and beyond. *arXiv* **2018**, arXiv:1811.02545.
52. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random Erasing Data Augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
53. Chen, P.; Liu, S.; Zhao, H.; Jia, J. Gridmask data augmentation. *arXiv* **2020**, arXiv:2001.04086.
54. Li, P.; Li, X.; Long, X. Fencemask: A data augmentation approach for pre-extracted image features. *arXiv* **2020**, arXiv:2006.07877.
55. Pakhomov, D.; Premachandran, V.; Allan, M.; Azizian, M.; Navab, N. Deep Residual Learning for Instrument Segmentation in Robotic Surgery. In Proceedings of the Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, 13 October 2019; Proceedings 10. Springer: Berlin/Heidelberg, Germany, 2019.
56. Kazanzides, P.; Chen, Z.; Deguet, A.; Fischer, G.S.; Taylor, R.H.; DiMaio, S.P. An Open-Source Research Kit for the da Vinci® Surgical System. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–7 June 2014; IEEE: Hong Kong, China, 2014.
57. Novellis, P.; Jadoon, M.; Cariboni, U.; Bottoni, E.; Pardolesi, A.; Veronesi, G. Management of robotic bleeding complications. *Ann. Cardiothorac. Surg.* **2019**, *8*, 292. [CrossRef]
58. Allan, M.; Shvets, A.; Kurmann, T.; Zhang, Z.; Duggal, R.; Su, Y.-H.; Rieke, N.; Laina, I.; Kalavakonda, N.; Bodenstedt, S. 2017 robotic instrument segmentation challenge. *arXiv* **2019**, arXiv:1902.06426.