Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD27402

Improving the Robustness of Surgical Instrument Segmentation Models via GAN-Based Data Augmentation

Jiaheng Li

Arcadia High School, Arcadia, USA
45155@students.ausd.net

Abstract. In robotics-assisted surgery, accurate segmentation of surgical instruments from background pixels is crucial for the success of various downstream tasks. Currently, popular segmentation models such as the U-Net exhibit excellent performance under major surgical conditions. However, when benchmarking the segmentation models under unforeseen surgical complications such as overbleeding, concerns toward their robustness remain. In fact, due to the lack of training datasets for the surgical segmentation models under complications, these models often exhibit degraded performances and thus result in failures for conducting accurate segmentations at the pixel level. In this paper, we leverage the Generative Adversarial Network (GAN), specifically through generating realistic images that simulate the condition of overbleeding to enlarge the training datasets — a process known as data augmentation. With more diverse training data from different scenarios, segmentation models can be trained better and reduce the risk of errors. In addition, a dataset consisting of 17 mock endoscopic video sequences with binary masks indicating ground truths was used for training/testing. With the dataset, a U-Net architecture is used to examine performance through metrics such as the Normalized Surface Distance (NSD) and Dice Similarity Coefficient (DSC). As a result, integrations with GAN-based data augmentation allow segmentation models' performances to improve 9.07% for NSD and 6.63% for DSC. From the results, our GAN-based augmentation is proven to be effective, providing a novel direction for improving models' robustness during the training procedure, paving the way for future exploration and clinical translation of the cutting-edge deep learning models into real robotics surgery.

Keywords: Generative Adversarial Networks, Non-adversarial robustness, Surgical Tool Segmentation

1. Introduction

Robotics-assisted surgery has been identified as a pioneering field for novel applications of computer vision. The use of segmentation models in accurate robotic arm control has thereby become critical for the success of various downstream tasks. From 3D reconstructions of patients' anatomical structures to collision avoidance between robotics arms and human organs, accurate segmentation is an essential prerequisite before taking a step further. From U-Net [1] to Mask2Former [2], the existing models based on Feed-Forward Networks (FFN) have demonstrated

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD27402

desired, if not ideal, performances under major testing scenarios. These models, however, uniformly possess a common problem that originates in their model robustness. That is, when handling unseen and unpredictable surgical complications, the models are unable to segment as accurately as desired at the pixel level due to their overfitting nature in the training datasets. This inability can further disable the surgeons from making rational decisions instantaneously in such urgent situations. One complication is overbleeding, a frequently occurring complication during surgeries. As demonstrated in the Figure 1, overbleeding may corrupt some image features of the surgical arms and therefore degrade the segmentation accuracy. This uncertainty brought by unforeseen circumstances questions the reliability of existing models and restrains their clinical translation in the future. Without satisfactory segmentation, robotics arms turn their efficacy into interference for the surgeons, leading to a problematic outcome where the most fundamental surgical procedures are troubled in carrying out. To prevent such scenarios from occurring, this paper aims to raise awareness on model robustness while introducing possible solutions through generating training datasets adversarially.



Figure 1. Examples of overbleeding condition

As mentioned previously, the main methodology being leveraged is the Generative Adversarial Networks (GANs). In particular, we can perform data augmentation by taking images with a regular background to generate corresponding images with a corrupted overbleeding background. By introducing the reconstruction loss on the surgical arms in addition to the existing adversarial loss and carefully adjusting the weights during training, the generator is then expected to be able to generate images with blood features after taking in regular images when testing. This methodology has several advantages that justify its choice over the alternatives. The reconstruction loss on the surgical arms, for example, allows the generator to not only output a realistic photo corresponding to the input, but also gives an additional emphasis on the blood features that are essential in the augmented datasets. The combination of the two losses and a careful tuning of the weight ensures an optimization of both image realism and the presence of overbleeding, and this optimization is unachievable using either one of the two losses alone. Additionally, in comparison with many other generative models, such as the Stable Diffusion Model, GANs produce realistic photos while requiring less computational expense, thus becoming the ideal choice for the methodology.

With the augmented datasets, we can design an experiment where U-Net is utilized as the segmentation model. In addition, the U-Net architecture also contains several residual blocks for skip connections. Using this segmentation model architecture, three models (denoted Model 1,2,3) were set up. Model 1 is trained only on regular images, while Model 2 is trained on a combination of regular and blood images, and Model 3 is trained exclusively on blood images. After training, all three models are tested on blood images and evaluated using the Normalized Surface Distance (NSD) and Dice Similarity Coefficient (DSC). As a result, we found that in comparison with Model 1, Model 2 exhibits a 6.19% increase in NSD and a 2.53% increase in DSC; while Model 3 exhibits a 9.07% increase in NSD and a 6.63% increase in DSC. This result demonstrates an improvement in model robustness via the proposed methodology and thus highlights the effectiveness of GAN-based data augmentation. In addition, the paper finds that exposure to corrupted data during training

DOI: 10.54254/2755-2721/2025.LD27402

improves segmentation accuracy when exposed to similarly corrupted data during testing in the context of surgical complications. Overall, this paper provides a practical pipeline for surgical segmentation models to remain robust when encountering complications that may be underrepresented during model training.

2. Related work

Various FFN-based models have been established for surgical instrument segmentation tasks. These models, developed throughout the years, exhibit distinct architectural characteristics across different aspects, ranging from multi-scale feature learning [3] to the inclusion of the self-attention mechanism [4]. Part of such considerations in model designs [5-7], however, is gradually shifting towards robustness concerns against variations and getting more exposed through relevant research. Robustness [8] examines a model's performance under unforeseen or corrupted situations, which is critical for achieving reliable results, especially in the context of surgical data science. In recent years, robustness enhancement has been pioneered by the techniques of data augmentation [9-12], where various transformations are applied to sample images. These techniques have been proven effective during the testing, where the models maintain a satisfactory segmentation performance throughout [13]. While traditional augmentation techniques have significantly mitigated the issues of robustness, a measurable gap to excellence remains. This is due to their oversimplicity and inability to generate photo-realistic data to facilitate the training process [14]. Generative models , on the other hand, creates a new possibility for robustness to persistently develop. Generative models are capable of learning to synthesize more photo-realistic data representations, where specifically outlined conditions or image characteristics can be modeled with the provision of initializing datasets. This capability not only distinguishes itself from traditional augmentation techniques, but is also useful in resolving numerous downstream tasks [16-17]. In this work, we mainly discuss existing data augmentation methods in improving segmentation models' robustness to unforeseen and corrupted data, specifically examining the role of generative models among them.

Surgical Instrument Segmentation: Since U-Net [1] first introduced the role of segmentation models in surgical data science with an encoder-decoder architecture to skip connections, developments in relevant segmentation scenarios thrive. Models such as TernausNet [18], DeepLabV3+ [19], and HRNets [20] incorporate pre-trained backbones [21-22] such as VGG11 and ResNet101 to demonstrate the advantages of transfer learning in biomedical segmentation tasks. Unsatisfied with the incompatibility between object detection and pixel-wise semantic segmentation, models like Mask R-CNN [23] introduced novel mask prediction branches. Meanwhile, ISINet [24], MATIS [25], and Mask2Former [2] refined segmentation performance with attention-based improvements, culminating in advanced models like SurgicalSAM [26], which adapts to surgical settings through prototype-based prompting and contrastive learning. Extending further in accuracy, Attention U-Net [27] was modified and improved based upon traditional U-Net, where irrelevant features are omitted dynamically. These characteristics help by acting as useful indicators in analyzing their model robustness later in this work.

Model Robustness & Data Augmentation: As models' segmentation accuracy has been enhanced, their adaptability to unforeseen situations, or model robustness, becomes increasingly critical. A lack of robustness limits a model's likelihood to be generalized for public use, as there may exist countless other unaddressed possibilities in the real world. Currently, research trending in robustness enhancement primarily lies in data augmentation, a technique that artificially expands the training dataset by applying various transformations to the existing data. Traditional data augmentation techniques, such as AutoAugment [9] and RandAugment [10], have been proven effective in

DOI: 10.54254/2755-2721/2025.LD27402

diversifying training data through basic transformations like geometric changes [11], color adjustments [12], and elastic distortions. Additionally, methods including Mixup [28], CutMix [29], and StyleMix [30] further introduced interpolation-based techniques into experimentation as breakthroughs proceed. Interpolation-based techniques, in particular, select two random images with the labels and then linearly interpolate them. This will result in a blended image during training, encouraging models to learn generalized features and smoother decision boundaries. While interpolation-based techniques enhance model adaptability by creating new blended samples, other strategies focus on other dimensions of robustness. Masking and erasure techniques [31-33] aim to prevent models from overly relying on specific features, while complex composite methods like AugMix [34] and MixMatch [35] explore a multi-step transformation approach to further boost generalization and stability across unseen distributions.

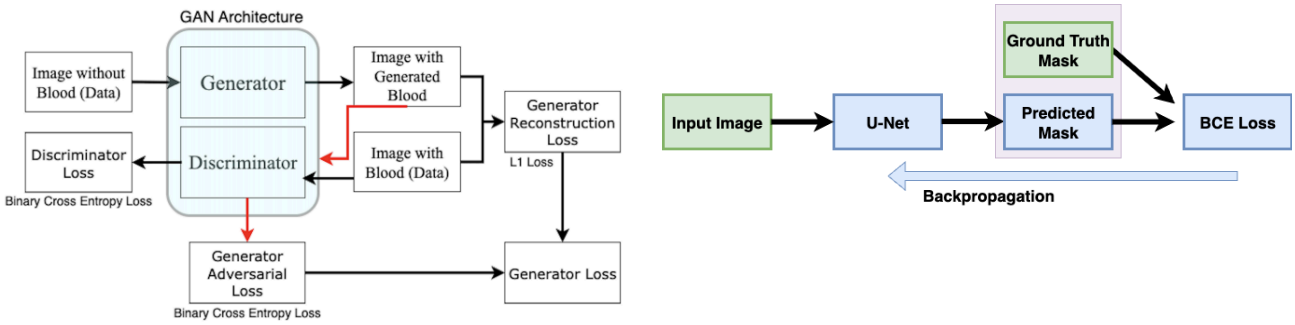
Generative Model: In data augmentation, the utilization of Adversarial training and generative models [15] is especially effective for improving robustness against various perturbations. Specifically, adversarial training [36] introduces targeted perturbations, preparing models for unexpected variations, especially crucial in robotics-assisted surgery. With GANs being used, data augmentation is capable of not only augmenting existing images but also generating new images in the datasets with unique features using a random noise vector. In image-to-image translation tasks, GANs such as Pix2Pix [37] and CycleGAN [38] are the most popular. Pix2pix requires supervised, or one-to-one paired data, to generate precise images, usually with colorization, segmentation, or translation effect. On the other hand, CycleGAN doesn't require both domains to be exactly paired. It's trained using Consistency Loss, which examines whether the original image and the one being generated in the other domain and when generated back. Generative models are useful methods to augment data and improve the robustness of segmentation models overall.

3. Methodology

3.1. Overview

In this section, we will discuss the model architecture of the GANs and U-Net, as well as the steps while training each model. Overall, the GANs consist of a generator and a discriminator, two convolutional neural networks that are trained adversarially with a well-designed architecture tailored for image-to-image translation. Specifically, the generator architecture contains an initial series of convolutions, two downsampling layers, six residual blocks, two upsampling layers, and a final output layer. In addition, the discriminator architecture contains five layers of convolutions that are PatchGAN-styled since the final convolutional layer is a 13×13 feature map before flattening. In addition, the prototype of the segmentation model uses U-Net, which contains a series of upsampling and downsampling of convolutional layers. Below, Figure 2a displays the flowchart of training GANs, and Figure 2b displays the training flowchart of the U-Net-based segmentation models.

DOI: 10.54254/2755-2721/2025.LD27402



(a) Flowchart of GANs training (b) Flowchart of U-Net training

Figure 2. GANs and U-Net training flowchart

3.2. GANs architecture

The architecture of the Generative Adversarial Network mainly comprises two networks: a generator and a discriminator. The generator is a neural network that creates realistic fake data to mimic the real data, while the discriminator is another network that tries to distinguish between real data and the fake data produced by the generator.

Generator: As shown in Figure 3a, the generator contains a convolutional encoder-decoder architecture with the addition of six residual blocks and additive connections. It maps an input image $x \in \mathbb{R}^{C\times H\times W}$ to a generated image $y \in \mathbb{R}^{C\prime\times H\times W}$ where H represents the image height, W represents the image width, and $C, C\prime$ each represents the number of image channels. And $C \times H \times W$, $C\prime\times H \times W$ represent the tensor dimensions of x, y.

The generator mainly comprises three convolutional layers, two down-sampling, two up-sampling, and a set of six ResNet-based residual blocks utilizing skip connections. In each of the three convolutional layers, the convolutional block is preceded by a reflection padding and followed by Instance Normalization and ReLU Activation. The reflection padding, as opposed to the typical zero-padding, performs better in terms of reducing edge artifacts and preserving edge textures. And the Instance Normalization normalizes the instances in the batch independently, which exhibits robustness with small batch sizes and stronger style-invariance for training. Overall, the convolutional layers extract local features without reducing spatial dimensions.

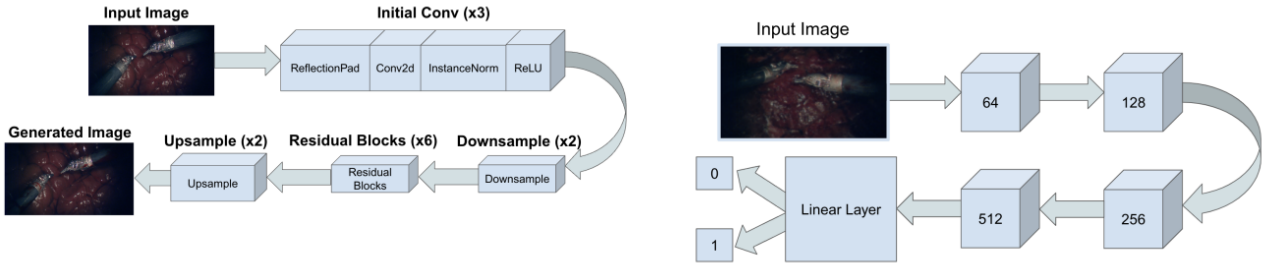
The initial convolutions are followed by two down-sampling layers. In down-sampling, the spatial dimensions of the input are reduced and thus enabling the generator to capture global textures. After down-sampling, six residual blocks were used to perform nonlinear transformations. Each residual block applies two convolutional layers with instance normalization and a ReLU activation in between, followed by an element-wise addition with the input. This helps to preserve the image content while facilitating a stable training.

The residual blocks are followed by two up-sampling layers that utilize additive connections with outputs of downsample 1 and initial convolution. In each step, the feature map is upsampled and then added to the output element-wise. This step helps to recover some fine spatial details and enables easier gradient flow. Lastly, the final convolution layer maps the feature map with 64 channels outputted by the second upsampling layer to the desired number of channels. It utilizes the Sigmoid activation function to normalize all image pixel values in the interval of $[0,1]$.

Discriminator: The discriminator is a convolutional neural network designed for distinguishing real and fake data, often produced by the generator. It takes in an image $x \in \mathbb{R}^{C\times H\times W}$ and outputs

DOI: 10.54254/2755-2721/2025.LD27402

$y \in [0,1]$. The architecture of the discriminator in Figure 3b mainly comprises several convolutions, LeakyReLU activations, and Instance Normalization. The architecture gradually reduces spatial size through strided convolutions, capturing both local and global image details. After the final convolution, the output is flattened into a fully connected layer, followed by a sigmoid activation that calculates a single probability score indicating whether the input image is real. The discriminator architecture is also styled like a PatchGAN since it uses smaller receptive fields to make decisions and outputs a 13×13 grid to predict on a patch-wise basis.



(a) Generator architecture (b) Discriminator architecture

Figure 3. Generator and discriminator architectures

3.3. GANs training

The training loop of the GANs is executed over multiple epochs. In each epoch, we can perform the following operations.

First of all, in the generator's forward pass, we let the generator G take a regular image x and generate a synthetic image $y = G(x)$. With the synthetic image y , the discriminator D evaluates it to compute the adversarial loss. Meanwhile, a reconstruction loss that specifically targets the surgical arms regions with binary mask m is computed to compare y and the ground truth $y\prime$. Note that the adversarial loss is a binary cross-entropy loss while the reconstruction loss is an L1 loss. The total generator loss is a combination of the two losses, where λ is the assigned weight for reconstruction loss. Then, the gradients accumulate, and the generator is updated for every interval of accumulation in backpropagation. The discriminator receives both the ground truth $y\prime$ and the generated image $y = G(x)$. It computes a binary classification loss using y and $y\prime$. Similar to the generator, the discriminator updates its gradients per accumulation interval. The generator and discriminator losses are recorded, and GANs' weights are saved every 20 batches as well as at the end of each epoch.

The generator adversarial loss is given by:

$$\mathcal{L}_{adv} = -\frac{1}{N}\sum_{i=1}^{N}(\log D(G(x))) \tag{1}$$

The generator reconstruction loss is given by:

$$\mathcal{L}_{recon} = \frac{\sum_{i,j}|G(x)_{i,j} - y\prime_{i,j}| \cdot m_{i,j}}{\sum_{i,j} m_{i,j}} \tag{2}$$

The combined generator loss is given by:

DOI: 10.54254/2755-2721/2025.LD27402

$$\mathscr{L}_{\mathrm{G}} = \mathscr{L}_{\mathrm{adv}} + \lambda \cdot \mathscr{L}_{\mathrm{recon}} \tag{3}$$

The discriminator loss is given by:

$$\mathscr{L}_{\mathrm{D}} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \left(-\log \mathrm{D}\left(\mathrm{y}^{(\mathrm{i})}\right) - \log\left(1 - \mathrm{D}\left(\mathrm{G}\left(\mathrm{x}^{(\mathrm{i})}\right)\right)\right)\right) \tag{4}$$

3.4. U-Net architecture

To evaluate the effectiveness of leveraging GANs in image segmentation, U-Net, a popular convolutional neural network designed for biomedical image segmentation, is utilized as the prototype for testing.

The U-Net architecture, as shown in Figure 4, consists of an encoder, bottleneck, and decoder with several stages of forward pass. First, the encoder contains a series of convolutional layers denoted as DoubleConv, where each is followed by a ReLU activation and max pooling to downsample the input image. Each layer aims to reduce the spatial dimensions while doubling the number of feature channels to increase feature abstractions. When the convolutions reach the deepest part of the network, also known as the bottleneck, it upsamples the layers through transposed convolutions. In contrast to downsampling, upsampling involves halving the channels and increasing the spatial dimension. During each stage of upsampling, a skip connection that links the encoder and decoder layers is utilized to help preserve spatial details that are potentially lost during down-sampling. This prevents losing fine-grained details and enables a high-level semantic understanding to conduct accurate segmentation.

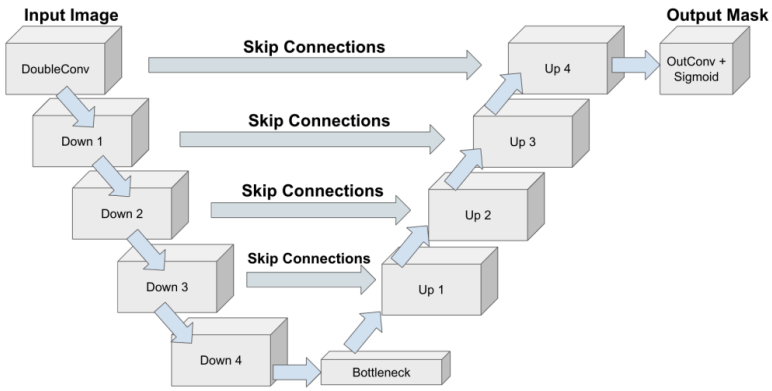


Figure 4. U-Net architecture

3.5. U-Net training

The U-Net is trained using a supervised learning approach, especially through the binary cross-entropy loss and optimized using the Adam optimizer. In each iteration of training, the U-Net processes the batches of input image pairs with their corresponding ground truth masks. For each batch, the segmentation model predicts a batch of masks and calculates the loss between the predicted masks and the ground truth masks. Then, the model updates its parameters using the loss via backpropagation. Using a predefined scheduler, the learning rate is adjusted in every epoch to help stabilize the loss convergence. The loss is recorded, and the model's states are checkpointed. The same process iterates for other epochs.

DOI: 10.54254/2755-2721/2025.LD27402

4. Experiment

4.1. Overview

In this section, we design an experiment that uses the designed GANs to generate blood image datasets. With the dataset, we then set up three U-Net segmentation models that are architecture-wise identical, yet differ in the content of training datasets. Specifically, Model 1 gets trained only on regular images, while Model 2 is trained on a combination of both regular and generated blood images, and Model 3 is trained exclusively on generated blood images. For each segmentation model, we evaluate it using the Normalized Surface Distance (NSD)and the Dice Similarity Coefficient (DSC) to assess performance under the unforeseen complication of overbleeding.

The designed experimental procedures are shown in Figure 5.

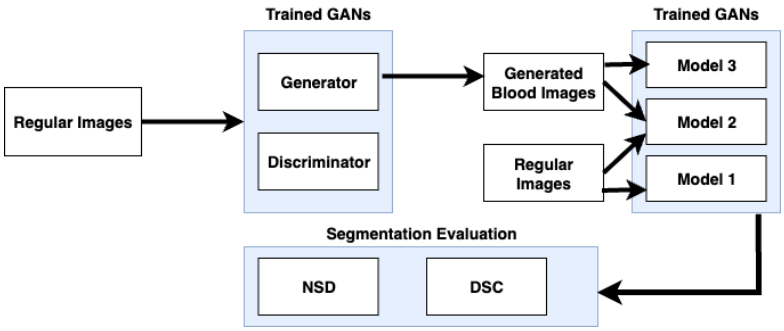


Figure 5. Flowchart of experimental design

4.2. Implementation details

The experiments are implemented on the Google Colaboratory using an A100 GPU. Specifically, the model setup, training, and evaluation are all carried out via the PyTorch Framework.

To test the hypothesis, we set up three segmentation models sharing the same U-Net architecture as the prototype but trained with different datasets:

• Model 1: trained only on regular images with no blood features.

• Model 2: trained on regular images as well as generated blood images. The ratio of the split is $2:3$.

• Model 3: trained only on generated blood images, all containing blood features.

Each of the three models is tested on the regular images and real blood images from the dataset.

4.2.1. Hyperparameters

Hyperparameters are declared prior to model training and adjusted during the experiment to optimize the value. After tuning the hyperparameters carefully through the experiments to achieve optimized segmentation, we obtain the following table.

DOI: 10.54254/2755-2721/2025.LD27402

Table 1. GANs training hyperparameters

Hyperparameter	Value
Training Hyperparameters	
Number of Epochs (num_epochs)	9
Batch Size (batch_size)	1
Accumulation Steps (accumulation_steps)	1
Shuffle (shuffle)	True
Optimizer Hyperparameters	
Generator Learning Rate (lr)	0.0003
Discriminator Learning Rate (lr)	0.0001
Optimizer	Adam
Model Hyperparameters	
Generator Input Channels (input_nc)	3
Generator Output Channels (output_nc)	3
Discriminator Input Channels (input_nc)	3

4.3. Evaluation metrics

After training, the segmentation models were tested on a dataset with all images containing blood features. This is used to simulate an unforeseen scenario where overbleeding occurs.

Specifically, the metrics used to evaluate the segmentation model's performance are the Normalized Surface Distance and the Dice Similarity Coefficient, two popular metrics used in biomedical segmentation tasks. Both metrics are naturally bounded within the interval from 0 to 1, and a large value indicates better segmentation performance. The specific formulation of each is broken down below.

Dice Similarity Coefficient:

The DSC can be computed through the following formula:

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{5}$$

In this case, TP represents the number of true positive pixels in the segmented image, FP represents the number of false positives, and FN represents the number of false negatives. This effectively measures the overall precision of segmentation.

Normalized Surface Distance:

The NSD can be computed through the following formula:

$$\mathrm{NSD} = \frac{\left|\mathrm{S_A}\cap\mathscr{B}_{\mathrm{B}}^{(\tau)}\right|+\left|\mathrm{S_B}\cap\mathscr{B}_{\mathrm{A}}^{(\tau)}\right|}{|\mathrm{S_A}|+|\mathrm{S_B}|} \tag{6}$$

Specifically, A and B represent the the segmentation masks, $\mathrm{S_A}$ and $\mathrm{S_B}$ represent the segmentation boundaries, and $\mathscr{B}_{\mathrm{B}}^{(\tau)}$ and $\mathscr{B}_{\mathrm{A}}^{(\tau)}$ represent the boundaries of A and B with extended tolerance τ. The normalized surface distance is used to complement the Dice similarity

DOI: 10.54254/2755-2721/2025.LD27402

coefficient. It specifically focuses on the segmentation performance on the boundaries of surgical tools.

4.4. Results

4.4.1. GANs training loss

The overall GANs are trained for 9 epochs. The loss of discriminator stabilizes around 1.2-1.3, while the loss of generator initially decreases, then converges to 4.0-5.0. The overall trend can be reflected in the following diagram.

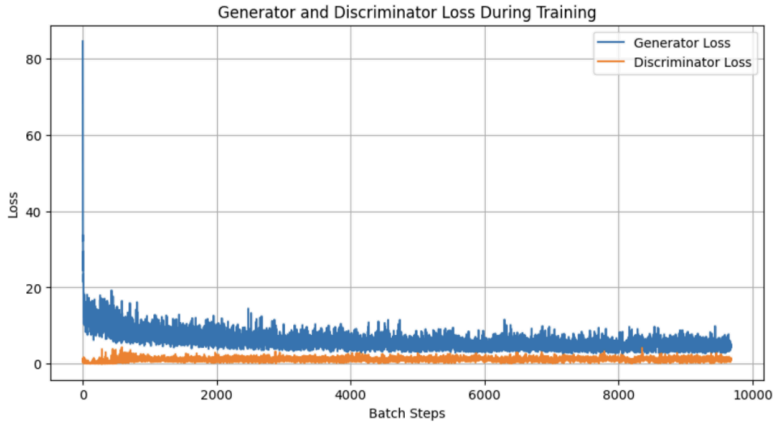


Figure 6. Generator loss over training

4.4.2. Images with generated blood samples

We select eight sample images with blood features generated by the trained GANs, displayed below.

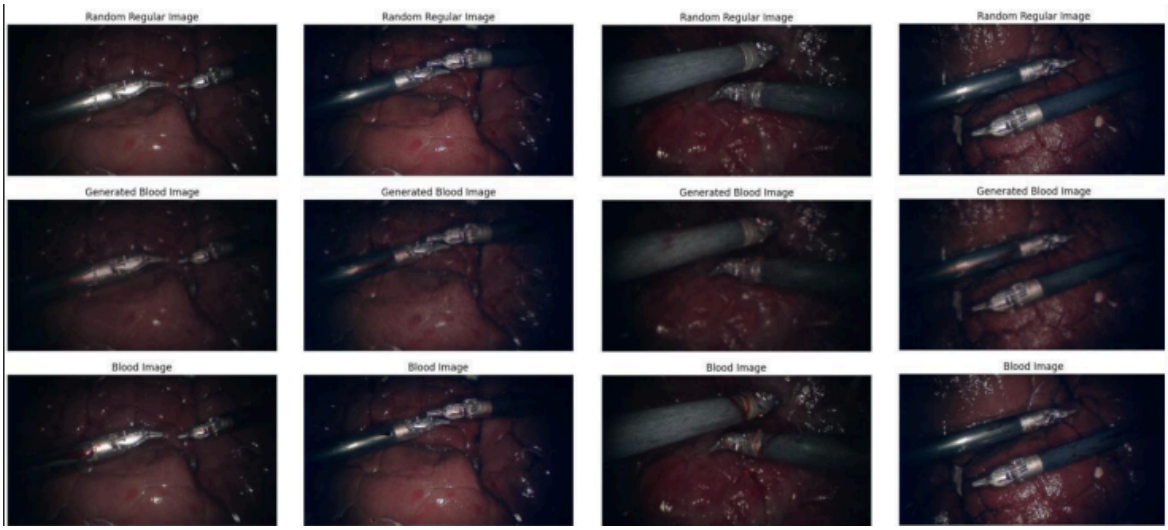


Figure 7. Regular/generated blood/real blood samples

4.4.3. Qualitative illustration of Model 1,2,3 on blood images

For each of the Model 1, Model 2, and Model 3, we can visualize their segmentation performance by inputting images with generated blood and obtaining outputs of segmented masks. In masks

DOI: 10.54254/2755-2721/2025.LD27402

displayed in Figure 15, the orange region represents the pixels labeled true by ground truth but not in the mask produced by the model (false negatives), the red region represents the pixels labeled true by model's produced mask but not in the ground truth (false positives), and the green region (true positives) represents the overlay of orange and red.

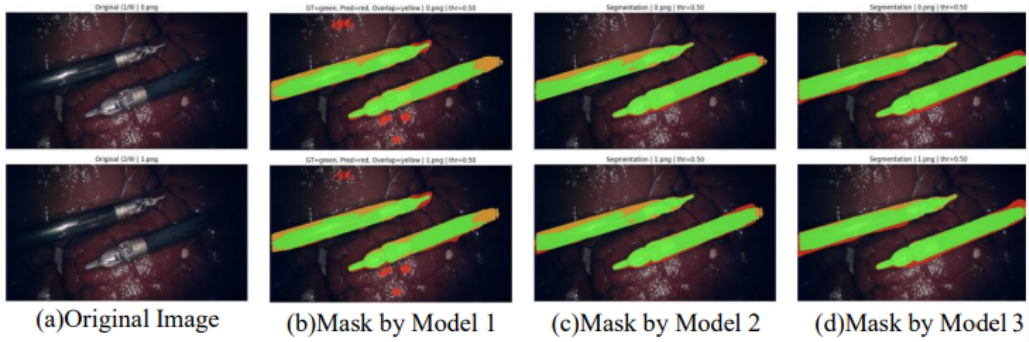


Figure 8. Binary mask for surgical instrument segmentation

4.4.4. NSD and DSC of Model 1,2,3 on blood images

The primary results are the metrics evaluating the segmentation performance. For each of the three models, the corresponding results are listed below in the form of mean (% of increase) ± standard deviation.

Table 2. NSD and DSC for Model 1, 2, 3

	Model 1	Model 2	Model 3
NSD	0.4880 ± 0.0674	0.5183 (+6.19%) ± 0.0958	0.5352 (+9.07%) ± 0.0962
DSC	0.6658 ± 0.1330	0.6826 (+2.53%) ± 0.1515	0.7099 (+6.63%) ± 0.1425

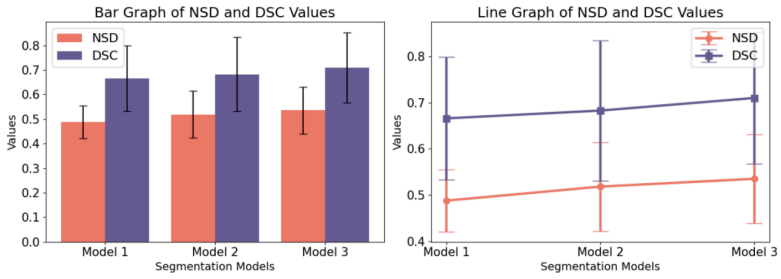


Figure 9. Bar and line graph for NSD and DSC

From the tables and graphs, we can observe an increase in the mean values of both NSD and DSC, suggesting an exceeding segmentation performance for Models 2 and 3 compared to Model 1. Specifically, Model 2's NSD increases by 6.19% while its DSC increases by 2.53% compared to Model 1. Additionally, Model 3's NSD increases by 9.07% while DSC increases by 6.63% . This implies that with more augmented datasets, the performance is better in overbleeding circumstances. Furthermore, the generated blood images are blurry on fine details and textures due to the downsampling architecture, suggesting that the results of Models 2 and 3 may be improved further without this situation.

DOI: 10.54254/2755-2721/2025.LD27402

4.4.5. NSD and DSC of Model 1,2,3 on regular images

In addition to testing the three models on generated blood images, we also test them on the regular images. For each of the three models, the corresponding results are listed below in the form of mean (% of increase) ± standard deviation.

Table 3. NSD and DSC for Model 1, 2, 3

	Model 1	Model 2	Model 3
NSD	0.8359 ± 0.0805	0.8026 ± 0.0759	0.7853 ± 0.0702
DSC	0.8907 ± 0.0645	0.8833 ± 0.0762	0.8769 ± 0.0551

The evaluation scores demonstrate that Model 1 performs the best among regular images, followed by Model 2, then Model 3. Despite being outperformed by Model 1, the percentage of increase when being tested with blood images compensates the percentage of decrease when being tested with regular images. The improvements of scores under overbleeding condition outweighs the minor score lowerings under regular surgical conditions.

5. Discussion

5.1. Limitations

There are several limitations of current research to address. For example, the current dataset on which the data augmentation is applied is insufficient to reflect the overall improvements in model robustness. There are various other types of surgery in addition to endoscopic surgery, such as open, laparoscopic, and orthopedic surgery. The lack of a dataset specific to these types limits the ability to generalize this GAN-based robustness enhancement technique across different surgical environments. Another limitation can be the overwhelming amount of computational expenses needed. This gets especially problematic when the scope of unforeseen complications expands beyond simply overbleeding, as it is practically unrealistic to train the model on every possible rare surgical scenario individually.

5.2. Future work

This conclusion inspires many possibilities for improving segmentation robustness. For instance, the proposed approach of data augmentation through GANs can be applied to similar surgical complications, such as smoke and low luminosity. This allows the models to be adaptable to a more diverse scope of scenarios, thus more reliable. In addition, models with enhanced robustness help introduce higher instrument tracking accuracy and more precise robotic movement, improving the quality of surgical processes in general.

One other aspect to consider is the integration of the segmentation model with multimodal data. Examples such as combining RGB and depth information or endoscopic video and ultrasound data can enhance the model's adaptability under different scenarios, providing a direction for future work.

For years, model robustness has been a determining factor before transforming the surgical process to be fully automated, especially in the context of surgical data science. Surgeries are full of unforeseen risks, uncertainties, and complications. By proposing and implementing segmentation models capable of handling these problems, it inspires future research in robotic-assisted surgery

DOI: 10.54254/2755-2721/2025.LD27402

through the utilization of computer vision techniques and develops a general trust in these models among clinicians.

6. Conclusion

In conclusion, this paper examines the effect of GAN-based data augmentation on enhancing segmentation performances under unforeseen circumstances, specifically overbleeding. Using a U-Net segmentation prototype and metrics such as NSD and DSC, we conclude that segmentation models 2 and 3, which were trained with the inclusion of generated blood images, exhibit better robustness under overbleeding condition while maintaining desired segmentation performance under regular conditions in comparison to model 1, which was trained purely on regular images. This proves the effectiveness of GANs-based data augmentation toward achieving robust surgical segmentation under surgical complications, providing insights to a novel direction for future exploration and clinical transition of the popular cutting-edge deep learning models into real robotics-assisted surgery.

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pages 234–241. Springer, 2015.

[2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1290–1299, 2022.

[3] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2117–2125, 2017.

[4] A Vaswani. Attention is all you need. Advances in Neural Information Processing Systems, 2017.

[5] Ying Yu, Chunping Wang, Qiang Fu, Renke Kou, Fuyu Huang, Boxiong Yang, Tingting Yang, and Mingliang Gao. Techniques and challenges of image segmentation: A review. Electronics, 12(5), 2023.

[6] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems, 34: 12077–12090, 2021.

[7] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. Advances in Neural Information Processing Systems, 35: 1140–1156, 2022.

[8] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8828–8838, 2020.

[9] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv: 1805.09501, 2018.

[10] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pages 702–703, 2020.

[11] Patrice Simard, David Steinkraus, and John Platt. Best practices for convolutional neural networks applied to visual document analysis. pages 958–962, 01 2003.

[12] Eun Kyeong Kim, Hansoo Lee, Jin Yong Kim, and Sungshin Kim. Data augmentation method by applying color perturbation of inverse psnr and geometric transformations for object recognition based on deep learning. Applied Sciences, 10(11), 2020.

[13] Misgana Negassi, Diane Wagner, and Alexander Reiterer. Smart (sampling) augment: Optimal and efficient data augmentation for semantic segmentation. Algorithms, 15(5): 165, 2022.

[14] Sanskruti Patel, Unnati Patel, Jay Nanavati, and Atul Patel. Analyzing generative models for realistic data augmentation across modalities and applications. In 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom), pages 1601–1606, 2024.

DOI: 10.54254/2755-2721/2025.LD27402

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 27. Curran Associates, Inc., 2014.
[16] Danush Kumar Venkatesh, Dominik Rivoir, Micha Pfeiffer, Fiona Kolbinger, and Stefanie Speidel. Synthesizing multi-class surgical datasets with anatomy-aware diffusion models. arXiv preprint arXiv: 2410.07753, 2024.
[17] Danush Kumar Venkatesh, Dominik Rivoir, Micha Pfeiffer, and Stefanie Speidel. Surgical-cd: Generating surgical images via unpaired image translation with latent consistency diffusion models. arXiv preprint arXiv: 2408.09822, 2024.
[18] Vladimir Iglovikov and Alexey Shvets. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. arXiv preprint arXiv: 1801.05746, 2018.
[19] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV), pages 801–818, 2018.
[20] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 43(10): 3349–3364, 2020.
[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.
[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014.
[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017.
[24] Cristina González, Laura Bravo-Sánchez, and Pablo Arbelaez. Isinet: an instance-based approach for surgical instrument segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 595–605. Springer, 2020.
[25] Nicolás Ayobi, Alejandra Pérez-Rondón, Santiago Rodríguez, and Pablo Arbeláez. Matis: Masked-attention transformers for surgical instrument segmentation. In 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), pages 1–5. IEEE, 2023.
[26] Wenxi Yue, Jing Zhang, Kun Hu, Yong Xia, Jiebo Luo, and Zhiyong Wang. Surgicalsam: Efficient class promptable surgical instrument segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 6890–6898, 2024.
[27] Ozan Oktay. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv: 1804.03999, 2018.
[28] Hongyi Zhang. mixup: Beyond empirical risk minimization. arXiv preprint arXiv: 1710.09412, 2017.
[29] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6023–6032, 2019.
[30] Minui Hong, Jinwoo Choi, and Gunhee Kim. Stylemix: Separating content and style for enhanced data augmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 14862–14870, 2021.
[31] Terrance DeVries. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv: 1708.04552, 2017.
[32] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In Proceedings of the AAAI conference on artificial intelligence, volume 34, pages 13001–13008, 2020.
[33] Pengguang Chen, Shu Liu, Hengshuang Zhao, Xingquan Wang, and Jiaya Jia. Gridmask data augmentation. arXiv preprint arXiv: 2001.04086, 2020.
[34] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. arXiv preprint arXiv: 1912.02781, 2019.
[35] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. Advances in neural information processing systems, 32, 2019.
[36] Yusuf Dalva, Hamza Pehlivan, Said Fahri Altındiş, and Aysegul Dundar. Benchmarking the robustness of instance segmentation models. IEEE Transactions on Neural Networks and Learning Systems, 2023.
[37] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1125–1134, 2017.

DOI: 10.54254/2755-2721/2025.LD27402

[38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycleconsistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 2223–2232, 2017.