DavinciGAN: Unpaired Surgical Instrument Translation for Data Augmentation

Kyungmoon Lee[*] KYUNGMOON@POSTECH.AC.KR
Min-Kook Choi MKCHOI@HUTOM.CO.KR
Heechul Jung HEECHUL@HUTOM.CO.KR

Abstract

Recognizing surgical instruments in surgery videos is an essential process to describe surgeries, which can be used for surgery navigation and evaluation systems. In this paper, we argue that an imbalance problem is crucial when we train deep neural networks for recognizing surgical instruments using the training data collected from surgery videos since surgical instruments are not uniformly shown in a video. To address the problem, we use a generative adversarial network (GAN)-based approach to supplement insufficient training data. Using this approach, we could make training data have the balanced number of images for each class. However, conventional GANs such as CycleGAN and DiscoGAN, have a potential problem to be degraded in generating surgery images, and they are not effective to increase the accuracy of the surgical instrument recognition under our experimental settings. For this reason, we propose a novel GAN framework referred to as DavinciGAN, and we demonstrate that our method outperforms conventional GANs on the surgical instrument recognition task with generated training samples to complement the unbalanced distribution of human-labeled data.

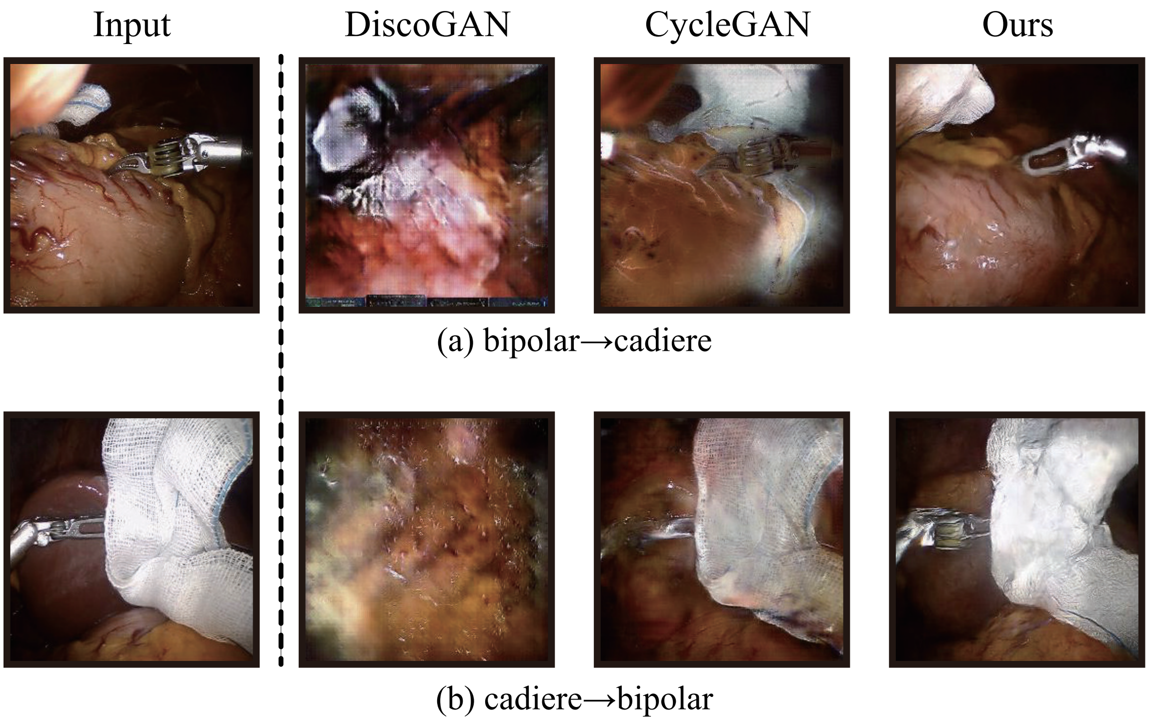
Keywords: Generative adversarial network (GAN), image-to-image translation, self attention, data augmentation.

1. Introduction

To help surgeon's decision making during the robotic surgery, providing surgical guidance like car navigation systems, based on the information extracted from the current surgery scene is necessary. Moreover, a surgery video should be analyzed to evaluate the robotic surgery after its operation. Recognizing surgical instruments is an essential process in such systems, and the information can be basically used for recognizing the current surgical phase (Twinanda et al., 2017).

In general, each surgical instrument is not used equally and uniformly in one operation. This leads to imbalance in terms of data collection, which is one of the critical problems in deep learning. In addition, since a certain tool is likely to show only in a specific environment, it can be said that context and background redundancy are high among the data for each tool. To address this issue, we propose an approach to translating an image. Generative adversarial network which is one of generative models is known for its ability to generate complex, high-dimensional data such as natural images (Goodfellow et al., 2014; Radford et al., 2016). With the advent of Conditional GAN, images can be created in the desired direction such as label-to-digit (Mirza and Osindero, 2014), text-to-image (Reed et al., 2016) and image-to-image (Isola et al., 2017; Choi et al., 2018).

[*] Work done at Hutom



(a) bipolar→cadiere

(b) cadiere→bipolar

Figure 1: Results of the conventional works (Kim et al., 2017; Zhu et al., 2017) and ours. DavinciGAN gives the result with appearance changes and identical background simultaneously. *(a)* bipolar→cadiere, *(b)* cadiere→bipolar. From left to right: input, DiscoGAN (Kim et al., 2017), CycleGAN (Zhu et al., 2017) and DavinciGAN (ours).

In particular, unpaired image-to-image translation (Zhu et al., 2017; Kim et al., 2017), which we address in this paper, has achieved impressive results recently.

However, these prior works failed easily when there are geometric changes between domains or the resolution of an input is high. Our goal is to address these issues as well as data imbalance problem between surgical instruments. To this end, our method, given an image, captures a candidate tool (e.g., cadiere) and transforms it into a target tool (e.g., bipolar), as shown in Figure 1. There are similar works (Joo et al., 2018; Tang et al., 2018) that change gestures of a person while maintaining his/her identity, but they differ from ours in the sense that they deal with simple images without causing geometric changes.

Our main contributions are as follows:

1. We propose a new generative adversarial network, named as DavinciGAN that captures candidate daVinci instruments and transforms them into target daVinci instruments by making appearance changes.

2. We introduce background consistency loss using self-attention mechanism without ground truth mask data. With this loss, our network is encouraged to transform only the candidate tool to the target tool while maintaining background.

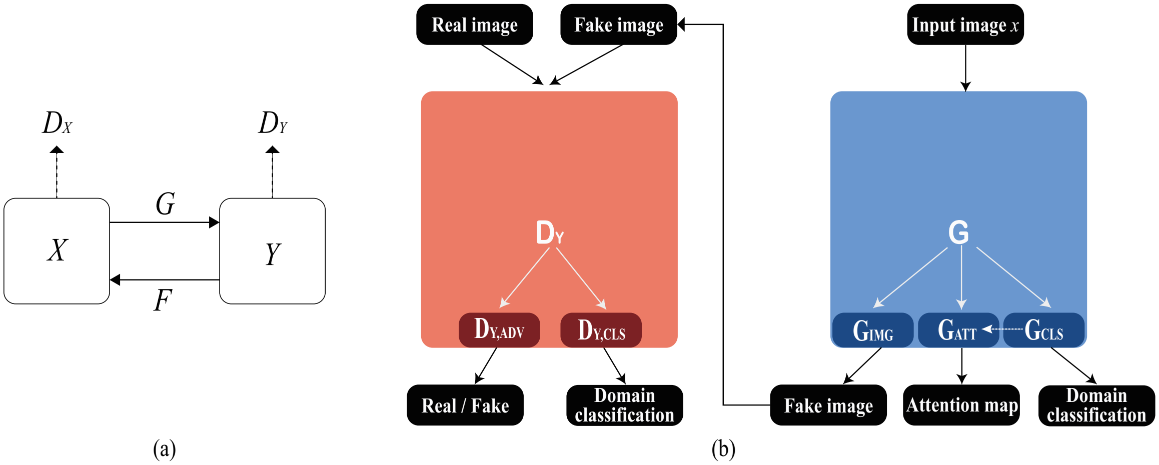


Figure 2: **Overall architecture of DavinciGAN.** (a) DavinciGAN consists of two generators G, F and two discriminators D_X, D_Y. (b) The figure on the left shows the discriminator D_Y, and the figure on the right shows the generator G. The F and D_X also follow this architecture.

3. We augment training data using GANs, and we show that the data augmentation using our DavinciGAN is the most effective to improving the instrument classification accuracies.
4. To the best of our knowledge, we first handle daVinci surgical instruments via image translation for data augmentation.

2. Methods

2.1. Architecture

Figure 2 illustrates the overall architecture of our DavinciGAN. Our goal is to find a mapping function to change the original surgical instrument in a source domain X to the desired surgical instrument in the target domain Y without background changes.

DavinciGAN consists of two generators $G : x \rightarrow \{G_{IMG}(x), G_{ATT}(x), G_{CLS}(x)\}$ and $F : y \rightarrow \{F_{IMG}(y), F_{ATT}(y), F_{CLS}(y)\}$, where $x \sim p_{data}(x)$ and $y \sim p_{data}(y)$. G_{IMG} represents a generator to generate fake image, and G_{ATT} produces an attention map computed from the predicted class score in domain classification and the feautre maps of G via weakly supervised learning technique (Zhou et al., 2016). Also, G_{CLS} is a classifier to classify domains. The F plays exactly the same role with G, but reverses two domains. Our DavinciGAN also has two discriminators $D_X : x \rightarrow \{D_{X,ADV}(x), D_{X,CLS}(x)\}$ and $D_Y : y \rightarrow \{D_{Y,ADV}(y), D_{Y,CLS}(y)\}$ and they also not only discriminate whether an input image is real or fake but also classify its domain.

2.2. Loss Function

We designed four loss functions such as adversarial loss, domain adversarial loss, background consistency loss and cycle consistency loss. The adversarial loss function is designed to make the generated image distribution indistinguishable to the real data distribution. Also, the domain adversarial loss function is introduced with an auxiliary classifier to derive efficient topological changes. With the background consistency loss, our method can maintain the background information while changing the appearance of the tool. Lastly, the cycle consistency loss is helpful to reduce the space of possible mapping functions and guarantee one-to-one mapping between two domains.

2.2.1. Adversarial Loss

We use an adversarial loss as a perceptual loss to enhance the naturalness of the generated images. By using this loss function, regardless of which output the auxiliary classifiers of discriminators give for domain classification, the generators learn to generate images indistinguishable from real images. In our work, we adopt LS-GAN loss (Mao et al., 2017) which is known to be advantageous for learning stability. We train G and F to maximize this objective and D_X and D_Y to minimize this objective.

$$\begin{aligned} L_{adv} = \mathbb{E}_{x\sim p_{data}(x)}[\|1 - D_{X,ADV}(x)\|_2 + \mathbb{E}_{y\sim p_{data}(y)}[\|D_{X,ADV}(F_{IMG}(y))\|_2 \\ + \mathbb{E}_{y\sim p_{data}(y)}[\|1 - D_{Y,ADV}(y)\|_2 + \mathbb{E}_{x\sim p_{data}(x)}[\|D_{Y,ADV}(G_{IMG}(x))\|_2. \end{aligned} \tag{1}$$

2.2.2. Domain Adversarial Loss

Along with an adversarial loss, we introduce domain adversarial loss to lead to appearance changes. We added an auxiliary classifier for each discriminator to derive geometric changes by adversarial learning with the output of that auxiliary classifier.

$$L_{D,CLS} = BCE(D_X) + BCE(D_Y), \tag{2}$$

where

$$BCE(N) \triangleq \mathbb{E}_{y\sim p_{data}(y)}[-\log(N_{CLS}(y))] + \mathbb{E}_{x\sim p_{data}(x)}[-\log(1 - N_{CLS}(x))]. \tag{3}$$

As a two-class classification problem, we set the label of domain X to 0 and the label of domain Y to 1. We train discriminators to classify not only accurate domains of real data, but also real x closer to zero than fake x' generated from F and real y closer to one than fake y' generated from G. On the other hand, G and F try to fool discriminators to classify y' closer to one than y and x' closer to zero than x, respectively. D_X and D_Y try to minimize Eq (2) and Eq (4) and G and F try to maximize Eq (4).

$$\begin{aligned} L_{CLS-ADV} = \mathbb{E}_{x\sim p_{data}(x), y\sim p_{data}(y)}[D_{X,CLS}(x) - D_{X,CLS}(F_{IMG}(y)) + \\ D_{Y,CLS}(G_{IMG}(x)) - D_{Y,CLS}(y)], \end{aligned} \tag{4}$$

2.2.3. Background Consistency Loss

$$L_{G,CLS} = BCE(G) + BCE(F). \tag{5}$$

Unlike traditional image translation tasks, we want to cross-domain via transforming only the instrument while maintaining the background. In our method, as in the case of the discriminator, we added an auxiliary classifier for each generator to find out which region is important to classify the domain via weakly supervised learning technique (Zhou et al., 2016; Singh and Lee, 2017) by minimizing Eq (5). We set the label of domain X to 0 and the label of domain Y to 1 exactly like 2.2.2. By utilizing that region as a format of attention map, we try to find out where the instrument is in the given image without ground truth mask and transform the target instance only, while maintaining the background by minimizing Eq (6).

$$\begin{aligned} L_{BG-CONSIST} = \mathbb{E}_{x \sim p_{data}(x), y \sim p_{data}(y)}[\|(x - G_{IMG}(x)) \otimes (1 - G_{ATT}(x))\|_1 + \\ \|(y - F_{IMG}(y)) \otimes (1 - F_{ATT}(y))\|_1]. \end{aligned} \tag{6}$$

2.2.4. CYCLE CONSISTENCY LOSS

In the task of unpaired image-to-image translation, it is known that adversarial losses are not enough to guarantee the mapping between an input and the desired output because random data of target domain distribution can be generated when only adversarial loss functions are optimized. Since the task we address is to change only the instrument while maintaining the background, an input *i* and $G_{IMG}(i)$ or $F_{IMG}(i)$ must have one-to-one correspondence in theory. Cycle consistency loss is a great help in this context and we train *G* and *F* to minimize this objective.

$$\begin{aligned} L_{CYC-CONSIST} = \mathbb{E}_{x \sim p_{data}(x), y \sim p_{data}(y)}[\|x - F_{IMG}(G_{IMG}(x))\|_1 + \\ \|y - G_{IMG}(F_{IMG}(y))\|_1]. \end{aligned} \tag{7}$$

2.2.5. FULL OBJECTIVE

To sum it up, the full objective functions to optimize discriminators and generators are as follows, respectively.

$$L_D = L_{adv} + \lambda_{CLS} * (L_{D,CLS} + L_{CLS-ADV}). \tag{8}$$

$$\begin{aligned} L_G = &- L_{adv} + \lambda_{CLS}(L_{G,CLS} - L_{CLS-ADV}) \\ &+ \lambda_{BG} L_{BG-CONSIST} + \lambda_{CYC} L_{CYC-CONSIST}. \end{aligned} \tag{9}$$

For hyper-parameter setting, we set λ_{CLS} = 1, λ_{BG} = 10 and λ_{CYC} = 10.

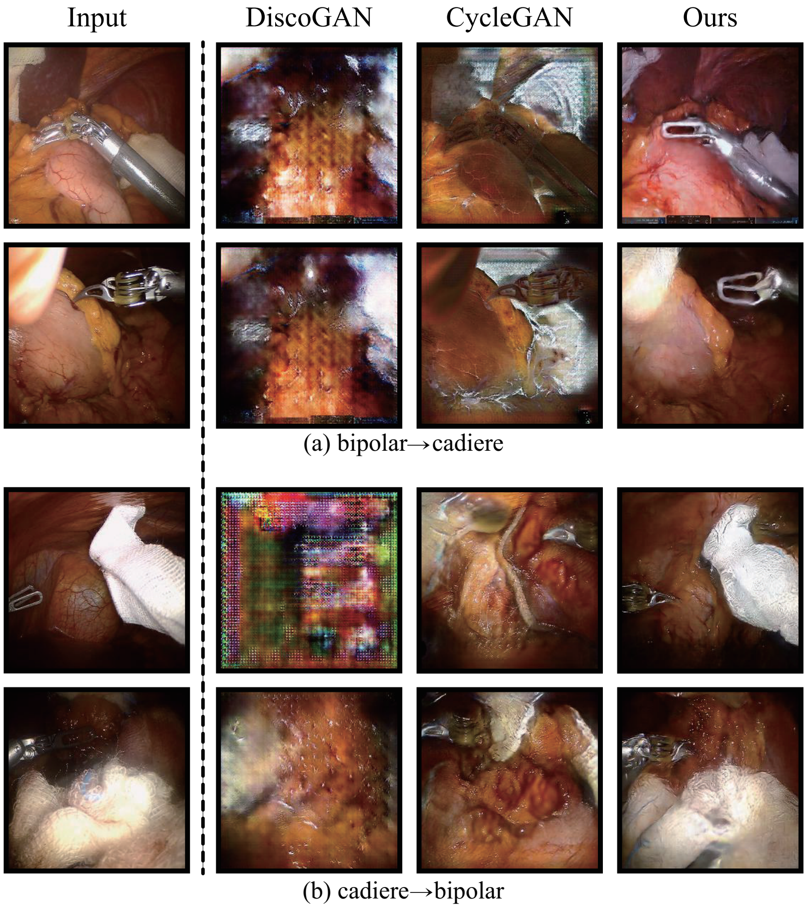
3. Experiments and results

3.1. Dataset

With 8 surgery videos using the daVinci Surgical System, we label frames having only one corresponding instrument. Since we found that bipolar is common and cadiere is rare relatively among all surgical instruments, we chose these two instruments to conduct experiments and finally, we built our surgical instrument dataset, consisting of 29,207 images where 15,344 are bipolar and 13,863 are cadiere.

3.2. Experiemental settings

We compare DavinciGAN with baseline models such as DiscoGAN and CycleGAN. The size of input and output images in our experiment is 256 × 256. For this setting, we added extra parameters to DiscoGAN addressing 64 × 64 size originally to equalize the number of learnable parameters. We adopt an Encoder-Decoder architecture which can be better in realizing appearance changes for our generators. Our discriminator uses 32 × 32 PatchGAN (Isola et al., 2017; Li and Wand, 2016;



(a) bipolar→cadiere

(b) cadiere→bipolar

Figure 3: **Additional translation results** (*a*): bipolar→ cadiere and (*b*): cadiere→ bipolar. From left to right: input, DiscoGAN (Kim et al., 2017), CycleGAN (Zhu et al., 2017), DavinciGAN (ours).

Ledig et al., 2017; Zhu et al., 2017) to classify whether patches are real or fake with the spatial information. Our network also used CBAM bottlenecks (Woo et al., 2018) between layers to get better performance of self attention.

We have chosen a mini-batch size of 8, and only horizontal flip was used as a data augmentation technique. Furthermore, we use the Adam optimizer (Kinga and Adam, 2015) with learning rate of 0.0002, β_1 of 0.5 and β_2 of 0.999. All models are implemented using Tensorflow and trained on a NVIDIA TITAN Xp GPU.

3.3. Results

3.3.1. QUALITATIVE RESULTS

Figure 3 shows our qualitative results on our surgical instrument dataset. DavinciGAN generated more convincing results than two baselines, DiscoGAN and CycleGAN. While maintaining the overall structure of background and surgical aids such as gauze, DavinciGAN transformed candidate instruments into target instruments by causing shape changes. In spite of trying various settings such as tuning the learning rate and the learning ratio between generator and discriminator per iteration, we found that DiscoGAN falled into mode collapse. This seems to be due to the fact that there is no constraint on the first translation such as background consistency loss function while handling complex images with organs, blood vessels and surgical aids (e.g., gauze, needle). CycleGAN adopts a fully convolutional ResNet structure (Johnson et al., 2016) which is known to be advantageous in generating high resolution images with little change in input. Although it maintains the structure of an input image well, it has difficulties in making variations to the candidate instrument with no constraint such as domain adversarial loss function.

3.3.2. QUANTITATIVE RESULTS

Table 1 shows the quantitative results of instrument classification utilizing the fixed real data and synthetic data generated by each model as training data. In consideration of the difficulty to collect surgical image data, the experiment was conducted with limited training data. We trained GAN-based models on $3,987$ images where $2,419$ are bipolar and $1,568$ are cadiere from three videos. For instrument classification task, we used 500 real data and 500 synthetic data for each instrument as training data while leaving all images of the remainder five videos as test data. For all cases, ResNet50 (He et al., 2016) trained until convergence. As a result, although DavinciGAN used less parameters than two baselines, it was superior to baselines in test accuracy and even competitive with additional real data.

Table 1: Classification performances and the number of parameters for each method.

Dataset	Method	# of parameters	Accuracy (%)
Real 1000	-	-	58.84
Real 1000 + Synthetic 1000	DiscoGAN	67M	57.91
Real 1000 + Synthetic 1000	CycleGAN	56M	58.61
Real 1000 + Synthetic 1000	DavinciGAN	31M	61.34
Real 2000	-	-	62.31

4. Discussion

4.1. Self attention via weakly supervised learning

Figure 4 (a) visualizes attention maps from generators and reversely attended input for each instrument. Reversely attended inputs which are utilized for the background consistency loss tend to hide instruments' head which is discriminative features to identify which instrument it is. We introduced self attention mechanism via weakly supervised learning to capture the position of instrument to transform without ground truth mask data. However, there are failure cases caused by the failure of attention. When generators classify the instruments, it tends to attend to the surgical aids, such as a gauze or the User Interface of daVinci surgical system, as well as the instrument's shape. As a result, we found that undesired attention maps are generated as shown in Figure 5. Since certain

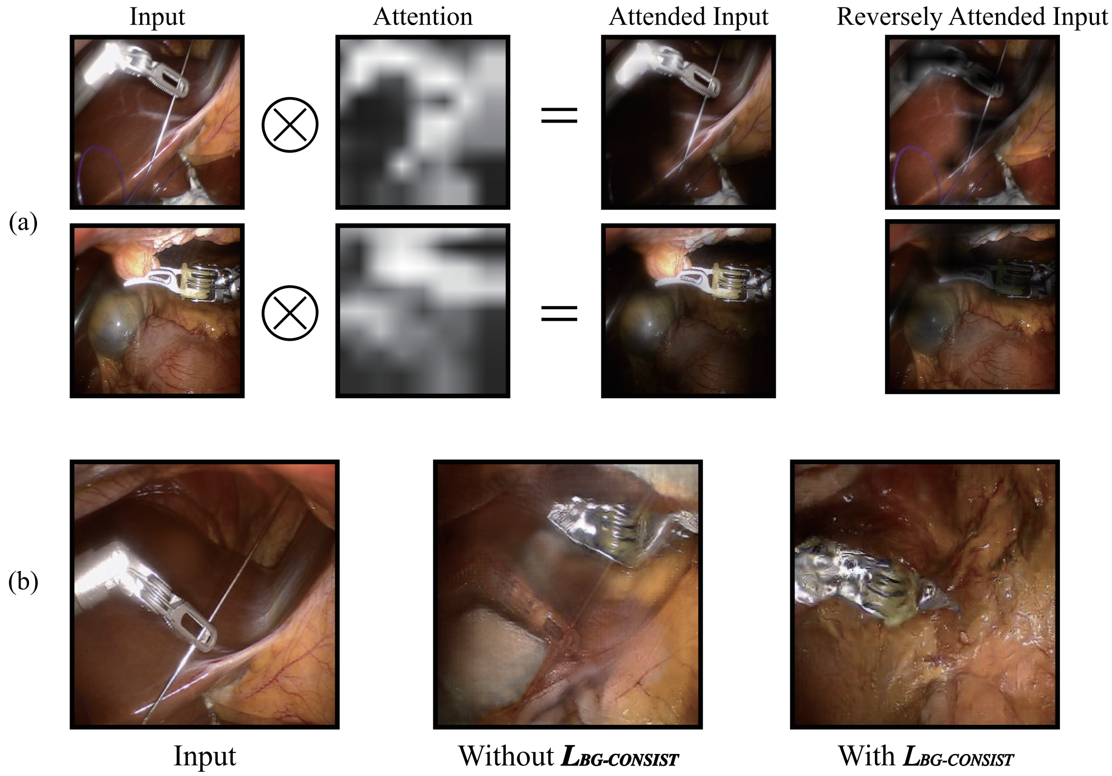


Figure 4: **How the generated attention is utilized in DavinciGAN and the visual analysis of the background consistency loss.** *(a)*: Attention maps produced by generators. Attended input shows discriminative features such as the head of instruments. On the other hand, reversely attended input which is utilized for the background consistency loss tends to hide discriminative features of the instrument. *(b)*: Without the background consistency loss, the bipolar is generated at an arbitrary position, but with background consistency loss, it is generated at the position of cadiere.

instruments tend to appear only in certain situations, classifier also tends to predict an output using other discriminative features rather than the instrument itself. As mentioned in section 3.3.2, limited data was used as training data due to the consideration of challenges to collect rich surgical video data. Surgical videos of more diverse surgeons will increase the appearance of instruments in orthogonal contexts, which can improve the performance of attention.

4.2. The effectiveness of background consistency loss

Interestingly, as shown in Figure 4 (b), the background consistency loss shows an intuitive result. Without the background consistency loss, our network generated a synthetic bipolar in the upper right side while hiding the cadiere by changing its color. This is because generators can fool discriminators by generating target instruments at any location, which is not a desired output for us.

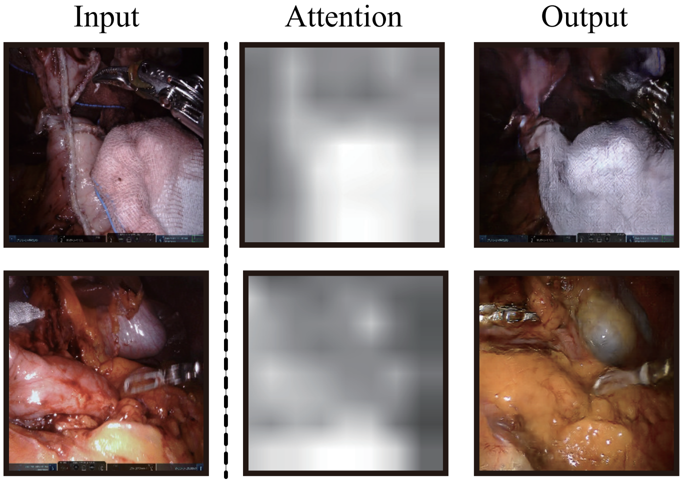


Figure 5: **Failure cases of DavinciGAN with undesired attention maps.** Attention maps from generators focus on gauze and User Interface of daVinci Surgical System, not on the shapes of instruments. With these undesired attention maps, generators produced undesired outputs.

However, with the background consistency loss, our network shows its capability to generate the target instrument at the desired position.

5. Conclusion

In this paper, we propose a novel generative adversarial network, DavinciGAN which transforms only surgical instrument while maintaining the background. This is achieved by the domain adversarial loss function and the background consistency loss function. We have showed qualitative and quantitative results on how useful generated data can be as training data compared to two baselines such as DiscoGAN and CycleGAN. One of advantages with our method is that DavinciGAN utilizes the self attention mechanism through weakly supervised learning approach so that we do not require any other annotation data like segmentation masks.

References

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

Donggyu Joo, Doyeon Kim, and Junmo Kim. Generating a fusion image: One's identity and another's shape. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

D Kinga and J Ba Adam. A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.

Hao Tang, Wei Wang, Dan Xu, Yan Yan, and Nicu Sebe. Gesturegan for hand gesture-to-gesture translation in the wild. *arXiv preprint arXiv:1808.04859*, 2018.

Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 2017.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2018.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.