OfGAN: Realistic Rendition of Synthetic Colonoscopy Videos

Jiabo Xu[1,2(✉)], Saeed Anwar[1,2], Nick Barnes[1], Florian Grimpen[3],
Olivier Salvado[2], Stuart Anderson[2], and Mohammad Ali Armin[1,2]

[1] Research School of Engineering, Australian National University, Canberra,
Australia
jiabo.xu@anu.edu.au
[2] Data61, CSIRO, Canberra, Australia
[3] Department of Gastroenterology and Hepatology, RBWH, Herston, Australia

Abstract. Data-driven methods usually require a large amount of labelled data for training and generalization, especially in medical imaging. Targeting the colonoscopy field, we develop the Optical Flow Generative Adversarial Network (OfGAN) to transform simulated colonoscopy videos into realistic ones while preserving annotation. The advantages of our method are three-fold: the transformed videos are visually much more realistic; the annotation, such as optical flow of the source video is preserved in the transformed video, and it is robust to noise. The model uses a cycle-consistent structure and optical flow for both spatial and temporal consistency via adversarial training. We demonstrate that the performance of our OfGAN overwhelms the baseline method in relative tasks through both qualitative and quantitative evaluation.

Keywords: Colonoscopy · Optical flow · Generative adversarial network · Domain transformation

1 Introduction

Deep learning achieves impressive performance in many machine learning tasks, such as classification [14,26], semantic segmentation [24], and object detection [13]. Those remarkable models rely on large high-quality datasets, such as ImageNet [7], Cityscape [5] and Pascal VOC [10]. However, the amount and quality of labelled data for training is often a limiting factor. In medical imaging, annotating real data is a challenging and tedious task; besides, medical data are usually subject to strict privacy rules that impose a limitation on sharing. A solution to this problem is generating synthetic data in a large quantity within
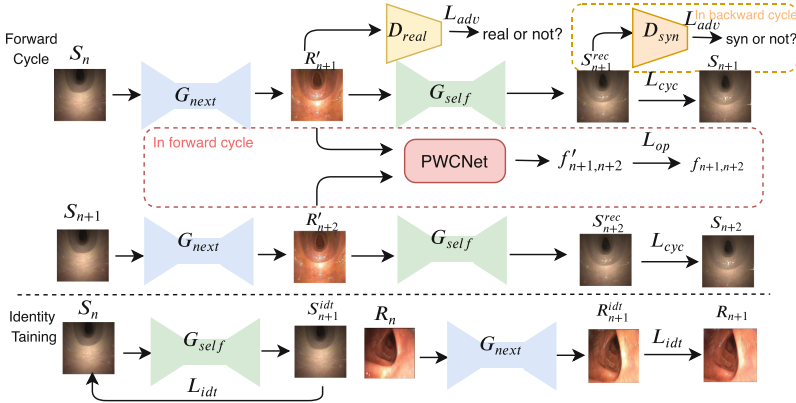
Fig. 1. Forward cycle and identity loss of OfGAN. G_{next} transforms the input to the next frame of the target domain. It is trained by the temporal-consistent loss which forces the reserved generation (G_{self}) to be same as the ground-truth synthetic frame.

controlled conditions. However, the lack of realism of synthetic data might limit the usefulness of the trained models when applied to real data. In this study, we propose a novel, ConvNet based model to increase the realism of synthetic data. We specifically work on simulated colonoscopy videos, but our approach can be expanded to other surgical assistance simulations.

Deep convolutional networks achieve remarkable performance on extracting low-dimensional features from image space [3]. Transforming one image to another requires the model to "understand" both the input and output domain in spatial domain. However, not only involving the spatial domain, our video transformation task overcomes three more challenges: (1) How to transform the input to the target domain while preserving the original annotation. (2) How to capture the temporal information between frames to form a consistent video. (3) How to synthesize near-infinite colonoscopy frames.

Generative Adversarial Networks (GANs) [12,20,23,25,28,29] fill the gap of generating and transforming high-quality images [28]. Generally, GANs consist of a generator and a discriminator. The generator is trained to generate a sample approximating the target distribution while the discriminator learns to judge the realness of the given sample. An elaborate adversarial training makes it possible to fit or transform a complex distribution.

The domain distributions play a vital role in transformation. Hence, directly transforming colonoscopy images to another domain is challenging when the distance in between is significant. Recently, Shrivastava *et al.* [25] refined synthetic small-sized grayscale images to be real-like through their S+U GAN. After that, Mahmood *et al.* [20] applied the idea of S+U GAN to remove patient-specific feature from real colonoscopy images. Both mentioned methods employ the target domain in grayscale, which dramatically reduces the training burden.

Combining adversarial training with paired images [1,15,23] usually fulfills impressive results. However, rare paired datasets compel researchers to tackle unpaired datasets, similar to our case, Zhu *et al.* [29] proposed the powerful Cycle-consistent GAN (CycleGAN), which trained two complemented generators to form the reconstruction loss. Oda *et al.* [21] transformed endoscopy CT images to the real domain by using CycleGAN with deep residual U-Net [24]. To replace surgical instruments from surgical images, DavinicGAN [18] extended CycleGAN with attention maps. Although, these methods achieve limited success and are unable to achieve temporal consistency in the video dataset. To solve video flickering, Engelhardt *et al.* [9] combined CycleGAN with temporal discriminators for realistic surgical training. But it is difficult to ensure high temporal consistency with only image-level discrimination. In terms of unpaired video-to-video translation, existing methods [2,4] on general datasets utilized similar structures as CycleGAN and novel networks for predicting future frames. However, these methods do not restrict the transformed structure to its origin; instead, they encourage novel realistic features. Our OfGAN improves Cycle-GAN to temporal level by forcing the generator to transform the input frame to its next real-alike frame while restricting the optical flow between two continuous output frames to be identical with their input counterparts. This setup achieves remarkable performance at pixel-level in spatial as well as temporal domain transformation.

The contributions of this exposition are:

1. **Optical Flow GAN**: Based on the standard cycle-consistent structure, we create and implement the OfGAN, which is able to transform the domain while keeping the temporal consistency of colonoscopy videos and rarely influencing the original optical flow annotation.
2. **Real-enhanced Colonoscopy Generation**: Our method can be incorporated in a colonoscopy simulator to generate near-infinite real-enhanced videos. The real generated videos possess very similar optical flow annotation with the synthetic input. Frames inside the transformed videos are consistent and smooth.
3. **Qualitative and Quantitative Evaluation**: The model is evaluated on our synthetic and a published CT colonoscopy datasets [23] both qualitatively and quantitatively. The transformation can be applied to annotation and thus create labels associated with the new realistic data.

2   Methodology

Let us consider that we are given a set of synthetic colonoscopy videos $S = \boldsymbol{s}$ and real colonoscopy videos $R = \boldsymbol{r}$, where $\boldsymbol{s} = s_1, s_2, \ldots, s_n$ and $\boldsymbol{r} = r_1, r_2, \ldots, r_m$, then s_n represents the n-th frame in the synthetic video and r_m represents the m-th frame in the real video. It should be noted that there is no real frame corresponding to any synthetic frame. Furthermore, we have ground-truth optical flow $F = \boldsymbol{f}$ for all synthetic data, where $\boldsymbol{f} = f_{1,2}, \ldots, f_{n-1,n}$ and $f_{n-1,n}$ indicates the ground-truth optical flow between frame $n-1$ and n. The goal is to

learn a mapping $G : S \rightarrow R'$ where R' is a set of novel videos whose optical flow is identical to S while keeping the structure of S unchanged. To achieve this, we follow cycle-adversarial [29] training by using two generative models G_{self} and G_{next}, their corresponding discriminators D_{syn} and D_{real} as well as an optical flow estimator Op to form an optical flow cycle-consistent structure.

2.1 Temporal Consistent Loss

Different from the reconstruction loss in CycleGAN, our model tries to reconstruct the next frame of the given distribution. More specifically, forward cycle is connected by two mapping functions: $G_{next} : s_n \rightarrow r'_{n+1}$ and $G_{self} : r'_{n+1} \rightarrow s^{rec}_{n+1}$. G_{next} tries to transform a given synthetic frame s_n to be similar to a frame from the real domain, at the same time it predicts the next frame to obtain r'_{n+1}. In the reverse mapping, G_{self} transforms r'_{n+1} to s^{rec}_{n+1}. Further, our temporal consistent loss narrows the gap between s^{rec}_{n+1} and s_{n+1}. The generator G_{next} performs spatial and temporal transformation simultaneously while G_{self} only involves spatial transformation. Besides we have a backward cycle obeying the reverse mapping chain: $r_m \rightarrow G_{self} \rightarrow s'_m \rightarrow G_{next} \rightarrow r^{rec}_{m+1}$. We use ℓ_1 loss to mitigate blurring. The overall temporal consistent loss is given by:

$$
\begin{aligned}
\mathcal{L}_{cyc}(G_{next}, G_{self}) = &\,\mathrm{E}_{s \sim Pdata(S)}[||G_{self}(G_{next}(s_n)) - s_{n+1}||_1] \\
&+ \mathrm{E}_{r \sim Pdata(R)}[||G_{next}(G_{self}(r_m)) - r_{m+1}||_1].
\end{aligned}
\tag{1}
$$

2.2 Adversarial Loss

Adversarial loss [12] is utilized for both mapping functions described in the previous section. For the mapping function $G_{next} : s_n \rightarrow r'_{n+1}$ the formula of adversarial loss is:

$$
\begin{aligned}
\mathcal{L}_{adv}(G_{next}, D_{real}, S, R) = &\,\mathrm{E}_{r \sim Pdata(R)}[\log D_{real}(r)] \\
&+ \mathrm{E}_{s \sim Pdata(S)}[1 - \log D_{real}(G_{next}(s))]
\end{aligned}
\tag{2}
$$

For the reverse pair G_{self} and D_{syn}, the adversarial loss is $\mathcal{L}_{adv}(G_{self}, D_{syn}, R, S)$ where the positions of synthetic and real data are interchanged.

2.3 Perceptual Identity Loss

Nevertheless, the temporal-consistent loss itself is insufficient to force each generator to generate its targets. We use identity loss to force G_{next} to strictly generate the next frame and G_{self} to transform the current frame. Furthermore, we find measuring the distance on the perceptual level achieves better results. Finally, the formula is as follows:

$$
\begin{aligned}
\mathcal{L}_{idt}(G_{next}, G_{self}) = &\,\mathrm{E}_{r \sim Pdata(R)}[\theta(G_{next}(r_m)), \theta(r_{m+1})] \\
&+ \mathrm{E}_{s \sim Pdata(S)}[(\theta(G_{self}(s_n)), \theta(s_n)]
\end{aligned}
\tag{3}
$$

where the $\theta(\cdot)$ indicates the perceptual extractor.

2.4   Optical Flow Loss

In addition to the above operations in the unsupervised situation, the optical flow loss utilizes supervised information to preserve annotation and stabilize the training. We restrict each two continuous real-alike frames to have the same optical flow as their corresponding synthetic frames, as shown in Fig. 1. The optical flow loss is:

$$\mathcal{L}_{op}(G_{next}) = \mathrm{E}_{s \sim Pdata(S), f \sim Pdata(F)}[||Op(r'_n, r'_{n+1}) - f_{n,n+1}||_1], \quad (4)$$

where the $Op(\cdot)$ represents a non-parameteric model for optical flow estimation and $r'_n = G_{next}(s_n)$.

Therefore, the overall loss function can be presented as:

$$\mathcal{L}(G_{next}, G_{self}, D_{syn}, D_{real}) = \mathcal{L}_{adv}(G_{next}, G_{self}) + \lambda \mathcal{L}_{cyc}(G_{next}, G_{self}) \\ + \beta \mathcal{L}_{idt}(G_{next}, G_{self}) + \sigma \mathcal{L}_{op}(G_{next}), \quad (5)$$

where we have λ, β and σ as the importance of each term. The target is to solve the min-max problem of

$$G^*_{next}, G^*_{self} = \arg \min_{G_{next}, G_{self}} \max_{D_{syn}, D_{real}} \mathcal{L}(G_{next}, G_{self}, D_{syn}, D_{real}).$$

2.5   Implementation Details

To be fair with competing methods, we adopt many training parameters from CycleGAN. We use an encoder-decoder structure for the generators and Patch-GAN [19] for discriminators. Both generators consist of two down-sample and two up-sample layers with six residual blocks in between. For extracting perceptual features, we use the output of the second convolution block of pre-trained VGG-13 [26] on ImageNet. Similarly, the optical flow is estimated via pre-trained PWC-Net [27]. Furthermore, to optimize the network, we employ Adam [17] optimizer with beta equal to $(0.5, 0.999)$ and a learning rate of $2e^{-4}$. The input frames are resized to 256×256 while corresponding optical flow is re-scaled to the proper value. We set $\lambda = 150$, $\beta = 75$ and $\sigma = 0.1$. The framework is implemented in PyTorch [22] and trained on 4 Nvidia P100 GPUs for 100 epochs.

3   Experiments

The synthetic data we utilized is simulated by a colonoscopy simulator [6]. We extracted 8000 synthetic colonoscopy frames from five videos with ground-truth optical flow and 2741 real frames from 12 videos for training. Similarly, for testing, 2000 unknown synthetic frames are captured from two lengthy videos. The real data is captured from patients by our specialists. We perform fish-eye correction for all the real data and discard the real frames with extreme lighting conditions, wall-only images, and blurred images. Subsequently, we are left with
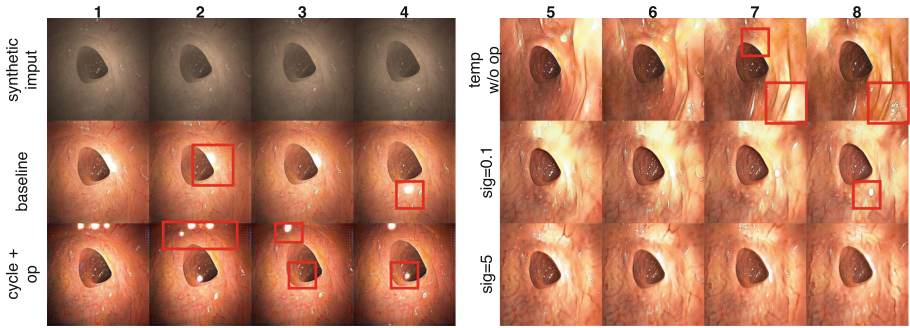
Fig. 2. Qualitative evaluation of four successive frames of each model. Each row, from top to bottom left to right, are input frames, results from the baseline, standard Cycle-GAN plus our optical flow loss, temporal consistent loss only, complete OfGAN with $\sigma = 0.1$ and $\sigma = 5$. Red rectangles highlight unseen features of one front frame. Differences are best viewed on zoom-in screen. (Color figure online)

1472 real images for training. Further, we also test our model on a published CT colonoscopy dataset [23] qualitatively.

We present the qualitative and the quantitative evaluation on our test results. The qualitative evaluation focuses on the single frame quality and temporal consistency in a subjective manner. For quantitative analysis, we use an auxiliary metric, Domain Temporal-Spatial score (DTS), to measure temporal and spatial quality simultaneously.

3.1   Qualitative Evaluation

The single-frame quality measures are two-fold. On the one hand, it measures if the transformed frame looks much more like real ones while, on the other hand, it evaluates if it contains less noise. For temporal consistency, we select four continuous frames and mainly concentrate on inconsistency among them. We regard the famous CycleGAN as our baseline model and furnish four models for ablation study. Results show that merely adding optical flow loss to the model does not improve rather results in worse performance on both spatial and temporal quality. The standard cycle structure does not involve in any temporal information, besides no spatial and temporal information can be learned at the same time. As a result, the black corner turns to be more obvious, and more inconsistent white spots emerge. Furthermore, only applying temporal-consistent loss (Fig. 2 *row 1, column 5–8*) intervenes in the converging of original training, which produces large scale mask-like noises. The combination of both optical flow loss and temporal-consistent loss gives much more realistic and consistent results (Fig. 2 *row 2, column 5–8*). Almost no white spots appear on any frames where the colon wall looks more bloody. A pale mask-like noise arises on the right. In terms of single frame quality (Fig. 3b), our method achieves better realness than the baseline method. By comparison, it is obvious that our method successfully
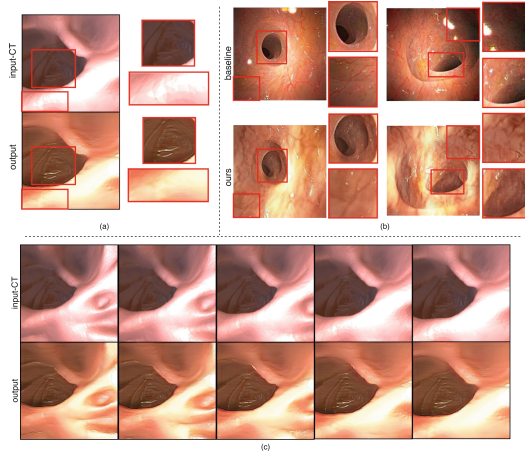
Fig. 3. (a) Qualitative evaluation of CT colonoscopy outputs, CT input (top), and our transformed results (bottom). (b) Qualitative assessment of the selected frame between two selected pairs from the baseline (top) and our method (bottom). Zoomed zones come from the detail inside the nearby red rectangles. (c) Qualitative evaluation on five continuous CT frame pairs, CT input (top), and our transformed results (bottom). (The images are best viewed on screen using zoom functionality). (Color figure online)

removes black corners and complements the detail in the deeper lumen. Besides, the white spots are rare in our results. The surface of the baseline method is so glossy that it looks far different from a human organ. On the contrary, our method has more vivid light effect.

The choice of parameter σ is a trade-off between consistency and realness. From values 0.1 to 5, results vary from the best realistic to the best consistent. Hence, we can adjust it depending on specific application scenarios.

We also test our method on CT colonoscopy videos (Fig. 3a *top row*) whose surface is coarse and texture-less compared with our synthetic data. In this case, we have no ground-truth optical flow of the input; instead, we use the estimated optical flow as the ground-truth for training. Our method successfully colors the surface to be realistic. Besides, it also removes the coarse surface and adds more realistic reflections of light inside the dark lumen. The lack of blood vessels is due to our non-blood-vessel-rich real data. Sequential frames (Fig. 3c) show that the innovative light reflection is consistent throughout these frames. In addition, no apparent noise nor inconsistent features appear.

3.2   Quantitative Evaluation

The quantitative evaluation should combine temporal, spatial, and domain distance measurements to overcome the trade-off problem. Hence, we utilize the DTS weighted by four normalized metrics, Average End point Error (AEPE) [8], average perceptual loss ($\mathcal{L}_{perc}$) [16], and average style loss ($\mathcal{L}_{style}$) [11] as the

auxiliary metric for combining spatial and temporal dimensions. AEPE is used to measure how well two continuous outputs possess the same optical flow as the corresponding inputs, which also indicates the consistency of temporal output. We use AEPE-GT (E_{gt}) and AEPE-Pred (E_{pred}), which are AEPEs between the result and ground-truth, and estimated optical flow of the input. $\mathcal{L}_{perc}$ and $\mathcal{L}_{style}$ is for spatial quality and domain distance. The weight selection is depended on the prior of each term. The coefficients are set up empirically based on the importance of each term. To calculate the mean, we randomly select ten samples from the entire real dataset for each test data. Finally, these metrics are normalized on 36 test cases with different hyper-parameters. The smaller the DTS, the better the performance. The overall formula of DTS is:

$$DTS = \frac{3}{8}\mathcal{N}(E_{gt}) + \frac{1}{8}\mathcal{N}(E_{pred}) + \frac{1}{4}\mathcal{N}(\mathcal{L}_{perc}) + \frac{1}{4}\mathcal{N}(\mathcal{L}_{style}) + 0.5, \qquad (6)$$

where $\mathcal{N}(\cdot)$ means normalization, adding 0.5 to make every value positive.

Table 1. Quantitative evaluation on test cases, referring to Fig. 2.

Approach	E_{gt}	E_{pred}	$\mathcal{L}_{perc}$	$\mathcal{L}_{style}$(1e−3)	DTS
Synthetic	0	0.15	3.31	11.19	1.47
Baseline	1.27	**0.24**	2.53	3.22	0.37
Cycle + op	1.53	0.81	2.43	2.82	0.40
Temp w/o op	2.85	2.35	**2.39**	**2.37**	0.77
sig=5	**1.19**	0.36	2.47	2.96	0.31
sig=0.1	1.22	0.31	2.49	2.64	**0.27**

The baseline method sacrifices the realness to achieve good consistency while only using temporal consistent loss is contrary, and both cases obtain a worse DTS (Table 1). Our method takes both advantages, even though not the best, and beats the baseline on E_{gt}, $\mathcal{L}_{perc}$, $\mathcal{L}_{style}$, and DTS. Notice that E_{gt} relies on the accuracy of the optical flow estimator, PWC-Net, as it has achieved state-of-the-art $E_{gt} = 2.31$ on MPI Sintel [27]. Even though we use different dataset, we think our $E_{gt} = 1.22$ (Table 1 *last row*) indicates the optical flow sufficiently identical to the ground-truth.

4   Conclusion

Our proposed OfGAN extends labeled synthetic colonoscopy video to real-alike ones. We have shown the performance of our OfGAN on our synthetic dataset and published CT datasets. The transformed dataset has outstanding temporal and spatial quality, which can be used for data augmentation, domain adaptation, and other machine learning tasks to enhance the performance. In term of the limitation, the performance of the proposed method might reduce if it fails to transform a frame correctly in a sequence. This can cause a dramatic effect on generating long videos, which needs to be dealt with in the future.

References

1. Armanious, K., et al.: Medgan: medical image translation using gans. Comput. Med. Imaging Graph. **79**, 101684 (2020)
2. Bansal, A., Ma, S., Ramanan, D., Sheikh, Y.: Recycle-gan: unsupervised video retargeting. In: Proceedings of the European conference on computer vision (ECCV), pp. 119–135 (2018)
3. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
4. Chen, Y., Pan, Y., Yao, T., Tian, X., Mei, T.: Mocycle-gan: unpaired video-to-video translation. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 647–655 (2019)
5. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3213–3223 (2016)
6. De Visser, H., et al.: Developing a next generation colonoscopy simulator. Int. J. Image Graph. **10**(02), 203–217 (2010)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255. IEEE (2009)
8. Dosovitskiy, A., et al.: Flownet: learning optical flow with convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp. 2758–2766 (2015)
9. Engelhardt, S., De Simone, R., Full, P.M., Karck, M., Wolf, I.: Improving surgical training phantoms by hyperrealism: deep unpaired image-to-image translation from real surgeries. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 747–755. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_84
10. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vision **88**(2), 303–338 (2010)
11. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2414–2423. IEEE (2016)
12. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in neural information processing systems, pp. 2672–2680 (2014)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969. IEEE (2017)
14. 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, pp. 770–778. IEEE (2016)
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134. IEEE (2017)
16. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
17. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

18. Lee, K., Jung, H.: Davincigan: unpaired surgical instrument translation for data augmentation (2018)

19. Li, C., Wand, M.: Precomputed real-time texture synthesis with markovian generative adversarial networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 702–716. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_43

20. Mahmood, F., Chen, R., Durr, N.J.: Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. IEEE Trans. Med. Imaging **37**(12), 2572–2581 (2018)

21. Oda, M., Tanaka, K., Takabatake, H., Mori, M., Natori, H., Mori, K.: Realistic endoscopic image generation method using virtual-to-real image-domain translation. Healthcare Technology Letters (2019)

22. Paszke, A., et al.: Automatic differentiation in pytorch, (2017)

23. Rau, A., et al.: Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. Int. J. Comput. Assist. Radiol. Surg. **14**(7), 1167–1176 (2019). https://doi.org/10.1007/s11548-019-01962-w

24. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

25. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2107–2116 (2017)

26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

27. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8934–8943. IEEE (2018)

28. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8798–8807. IEEE (2018)

29. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232. IEEE (2017)