OfGAN: Realistic Rendition of Synthetic Colonoscopy Videos

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

[1] Data61, Commonwealth Scientific Industrial Research Organization of Australia (CSIRO)

[2] Research School of Engineering, Australian National University

[3] Department of Gastroenterology and Hepatology, Royal Brisbane and Women's Hospital, Brisbane, Australia

www.data61.csiro.au

Introduction

Background:

In medical imaging, annotating real data is a challenging and tedious task. However, machine learning methods have to inference on patient-specific and camera-specific data, similarly for clinical training.

Related works:

Previous works [20, 29] achieve good performance on image-to-image translation in general image. However, simply applying the methods on colonoscopy videos leads to bad results.

Challenges:

- Hard to transform domain and keep temporal consistency simultaneously.
- Large domain gap makes the training much harder.

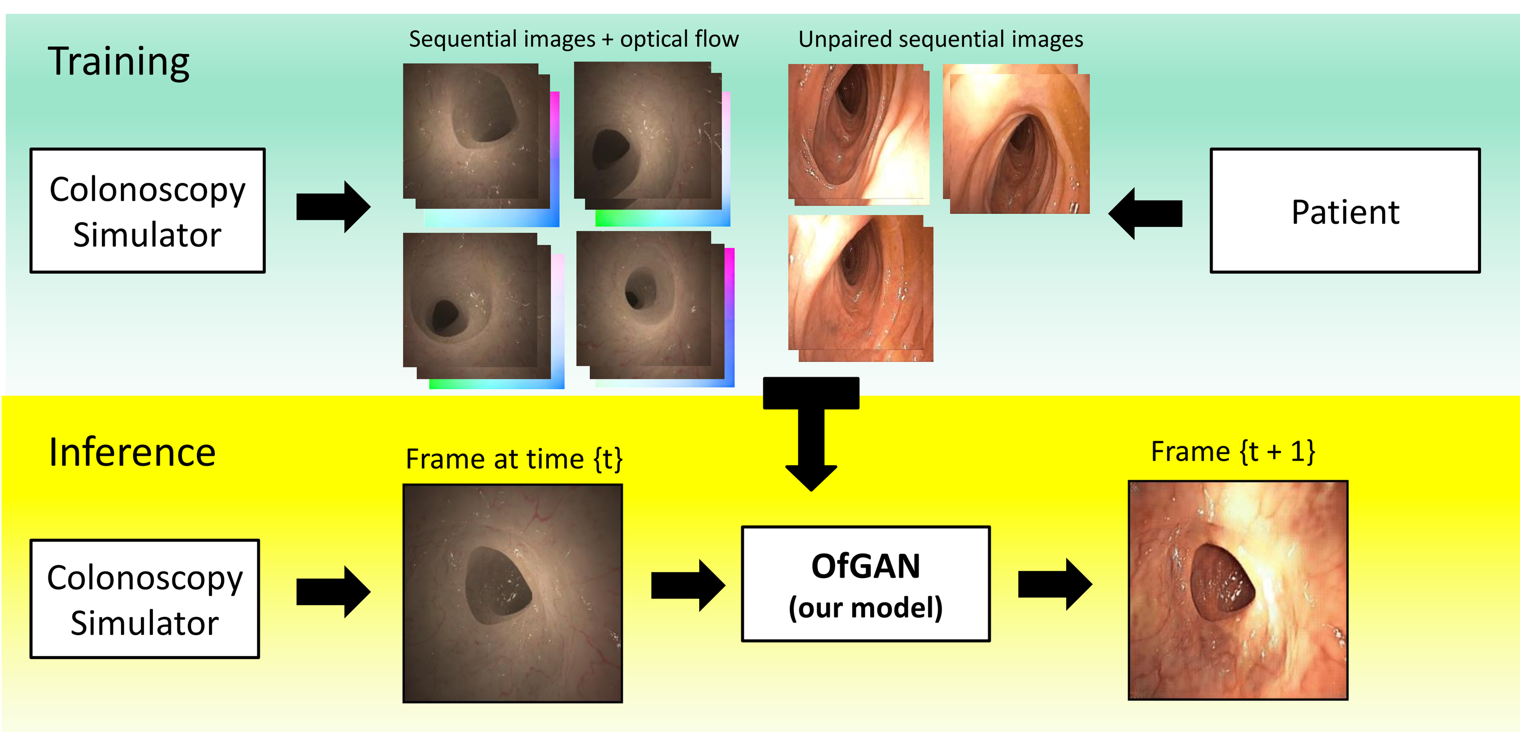

Fig.1: Overall workflow of our proposed method.

Method

Our model is a GAN-based model, which consists of two key components:

- **Temporal Cycle Structure**

Compared with CycleGAN [29] which forms a circle only on a single time step, we force a temporal mapping chain (forward cycle) as :

$$s_n \xrightarrow{G_{next}} R'_{n+1} \xrightarrow{G_{self}} s^{rec}_{n+1}$$

where the subscript indicates the number of continuous frames, prime indicates "transformed", "rec" indicates "reconstructed". In the novel mapping chain, the generator G_{flow} has to predict the next frame meanwhile doing the domain transformation.

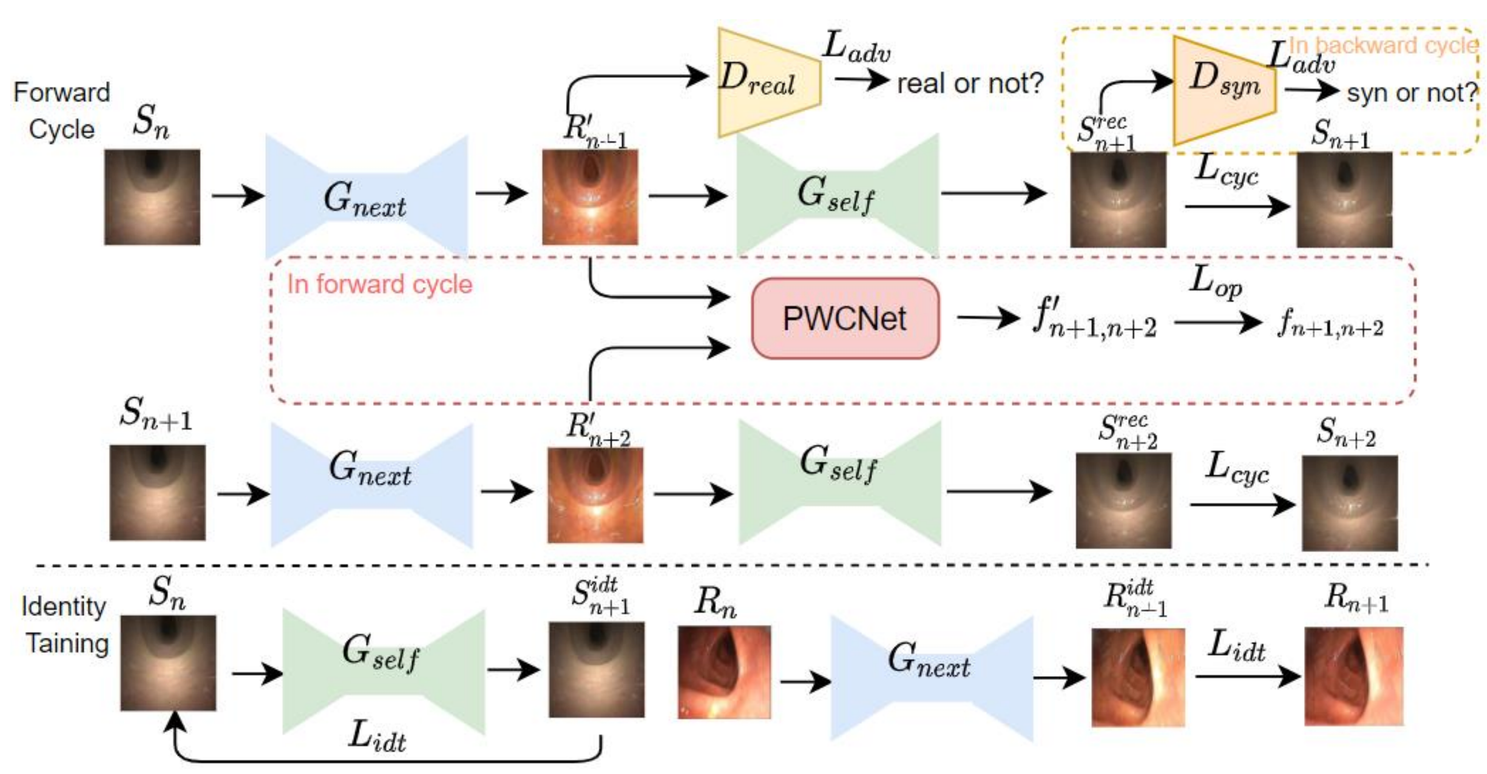

Fig.2: Forward cycle and identity loss of OfGAN. G_{next} transforms the input to the next frame of the target domain. It is trained by the temporal-consistent loss which forces the reserved generation (G_{self}) to be same as the ground-truth synthetic frame.

- **Optical Flow Loss**

we force the G_{next} to learn the underlying optical flow so that we have $f'_{n,n+1} \approx f_{n,n+1}$ where $f_{n,n+1}$ means the ground truth optical flow between the n-th frame and n+1-th frame and f', is the optical flow of the transformed version. We utilized a pretrained PWC-Net [27] for optical flow estimation.

Datasets

State of the art CSIRO simulator:

- Synthetic videos, consists of 8000 video frames with ground truth optical flow from 5 different simulated colons which were generated by a variety of possible camera motion. 2000 frames from two unknown synthetic colonoscopy videos were used for test.
- A Published CT colonoscopy dataset [23]

Real video:

- Real videos, consists of 1472 video frames (after cleaning) captured from patients by our specialists.

Experiments and Results

Qualitative evaluation:

It measures if the transformed frame looks much more like real ones while, on the other hand, it evaluating if it contains less noise. This evaluation is on the baseline model (CycleGAN) and our model testing on both our synthetic dataset and the public CT dataset.

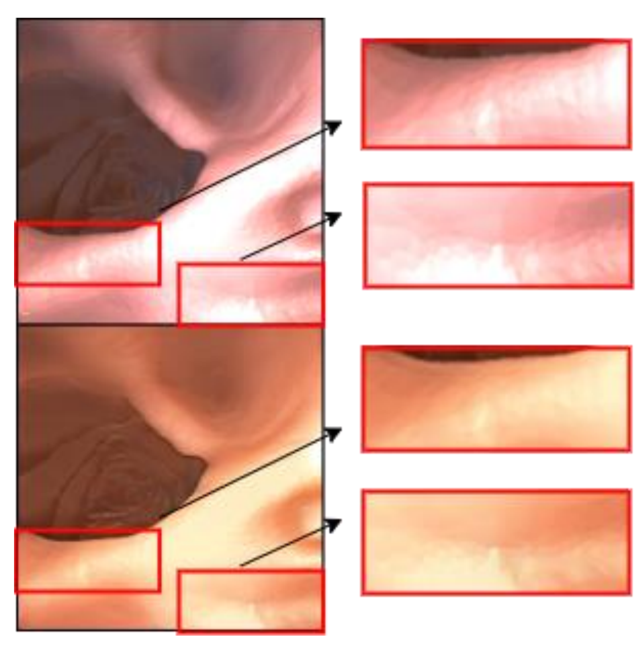

Fig.3: Qualitative evaluation of CT colonoscopy outputs, CT input (top), and our transformed results (bottom).

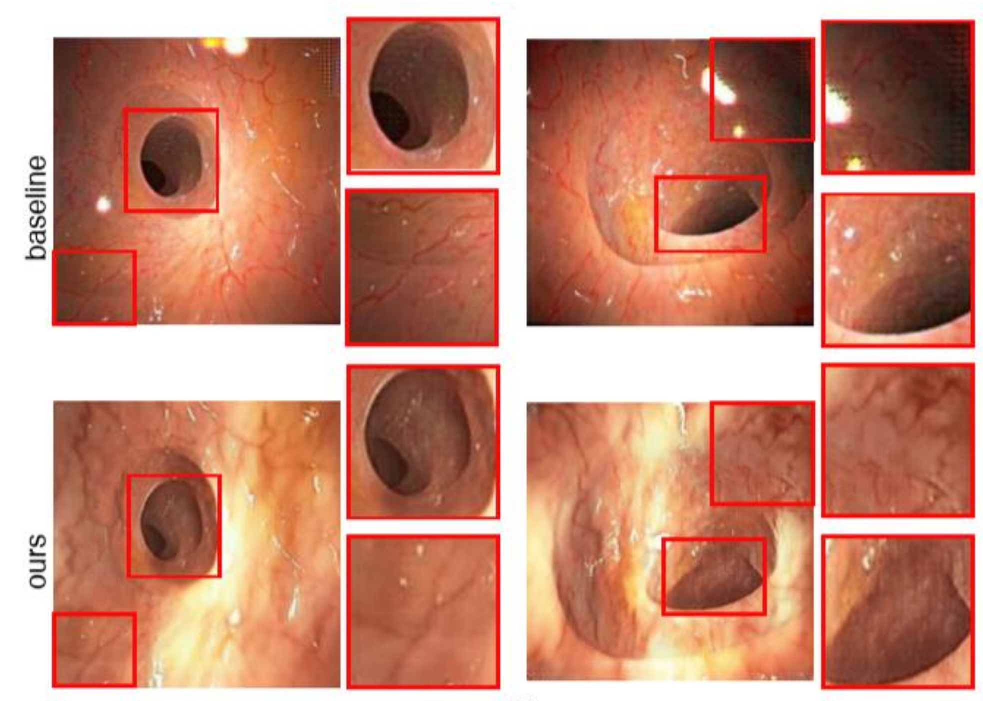

Fig.4: Qualitative assessment of the selected frame between two selected pairs from the baseline (top) and our method (bottom). Zoomed zones come from the detail inside the nearby red rectangles.

Quantitative evaluation:

It quantifies the performance by using a joint metric (DTS) of the spatial and the temporal dimension. We weighted four well-known losses, E_{gt} (AEPE to ground truth), E_{pred} (AEPE to estimated), L_{perc} (Perceptual loss), L_{style} (Style loss).

These metrics are normalized on 36 test cases with different hyper-parameters. The smaller the DTS, the better the performance.

$$DTS = \frac{3}{8}\mathrm{N}(E_{gt}) + \frac{1}{8}\mathrm{N}(E_{pred}) + \frac{1}{4}\mathrm{N}(L_{perc}) + \frac{1}{4}\mathrm{N}(L_{style}) + 0.5$$

Approach	E_{gt}	E_{pred}	L_{perc}	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**

Tab.1: "cycle + op" and "temp w/o op" are two setting for ablation study. Our full model is the last two with different sigmas (weight of optical flow loss).

Summary

Our OfGAN achieves high spatial and temporal quality on video domain transformation; It can be easily applied on a given simulator for generating realistic data; The transformed data can be patient-specific.

Limitations: The performance might reduce if it fails to transform a frame correctly in a sequence. This can cause a dramatic effect on generating long videos. We will verify our model on other styles of datasets in the future.

FOR FURTHER INFORMATION

Jiabo Xu
jiabo.xu@foxmail.com

Mohammad Ali Armin
Ali.Armin@data61.csiro.au

REFERENCES

[20] Mahmood, F., Chen, R., Durr, N.J.: Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. IEEE transactions on medical imaging 37(12), 2572–2581 (2018)

[23] Rau, A., Edwards, P.E., Ahmad, O.F., Riordan, P., Janatka, M., Lovat, L.B., Stoyanov, D.: Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. International journal of computer assisted radiology and surgery 14(7), 1167–1176 (2019)

[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 (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 (2017)

MICCAI 2020 Lima PERU VIRTUAL