Jiabo Xu | Video-to-video Translation and Relevant Research on Medical Images

Dec. 2019

www.data61.csiro.au

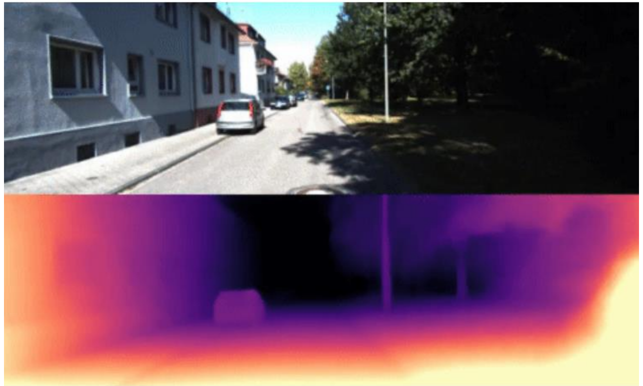
Contents

- Introduction to Image-to-image translation
- Introduction to Video-to-video translation
- Transformation from synthetic colonoscopy videos to real ones.

Image-to-image translation

Special cases:

Monocular depth estimation (Godard, 2018)



Semantic segmentation(Richter, 2016)

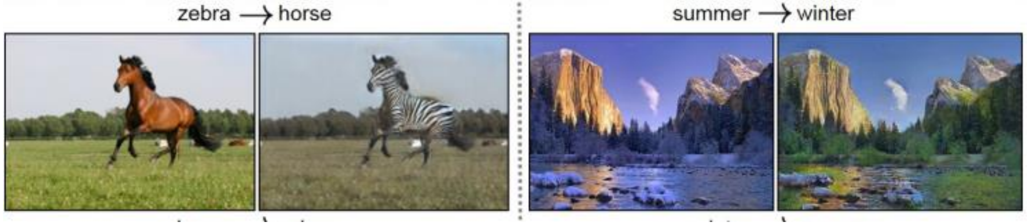
Colorization (Zhang, 2016)

Neural style transfer (Gatys, 2016)

Image-to-image translation

General cases: following images all from pix2pix (Isola, 2016) and cycleGAN (Zhu, 2017)

Domain transformation



Pose transfer

Semantic segmentation

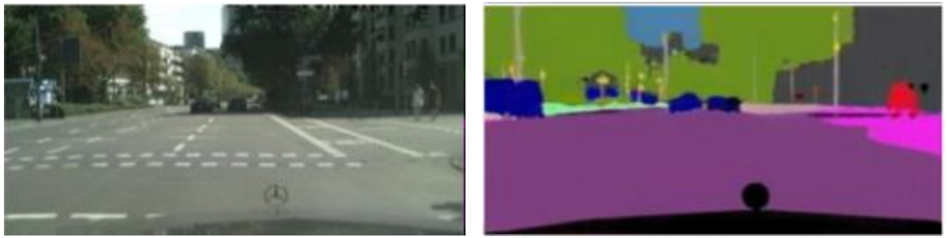


Photo inpainting

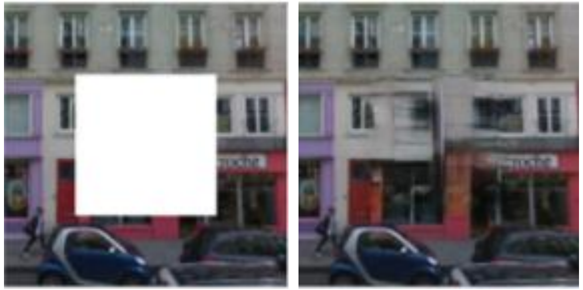


Image-to-image translation

How to achieve it?

Variants of Generative adversarial network (Goodfellow, 2014)

$$L_{adv}(G, D) = E_{y \sim Pdata(y)}[\log D(y)] + E_{x \sim Pdata(x)}[1 - \log D(G(x))]$$

$$G^* = \arg\min_G \max_D L_{adv}(G, D)$$

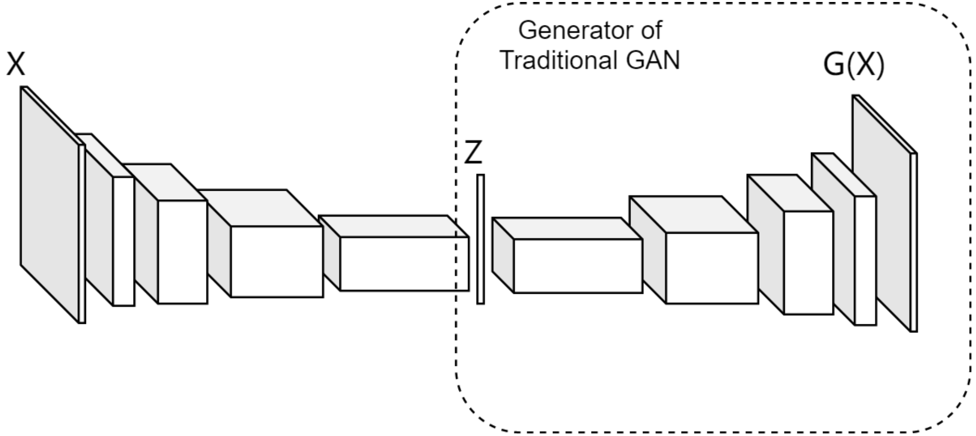
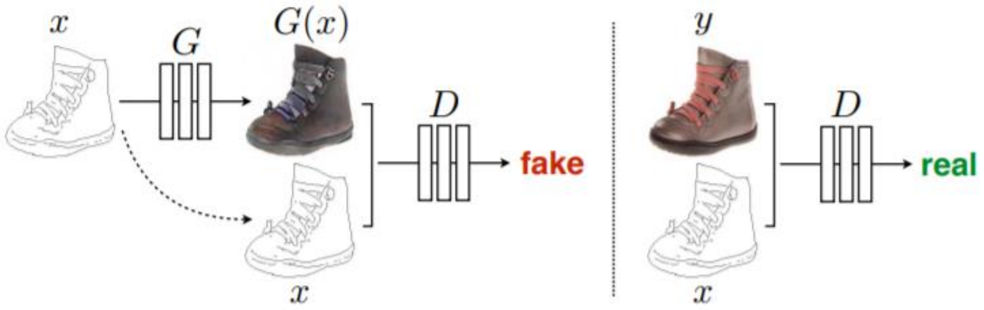


Image-to-image translation

Pix2Pix
- Paired dataset
- Conditional GAN



CycleGAN
- Unpaired dataset
- Cycle-consistent loss

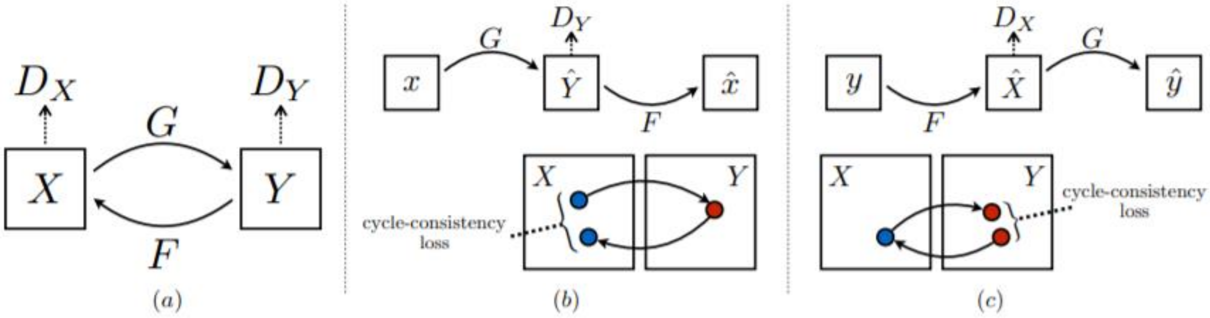
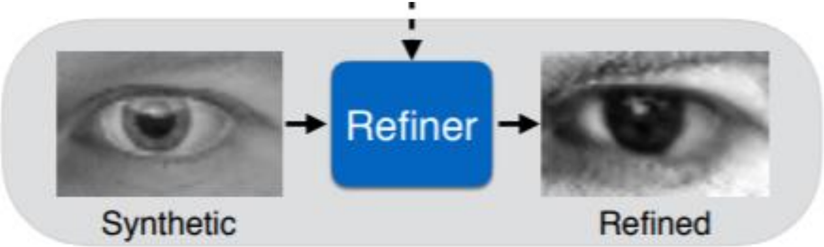


Image-to-image translation

Why need it?

- SO COOL!
- Integrate miscellaneous tasks
- Domain transformation



S+U GAN (Shrivastava, 2017)

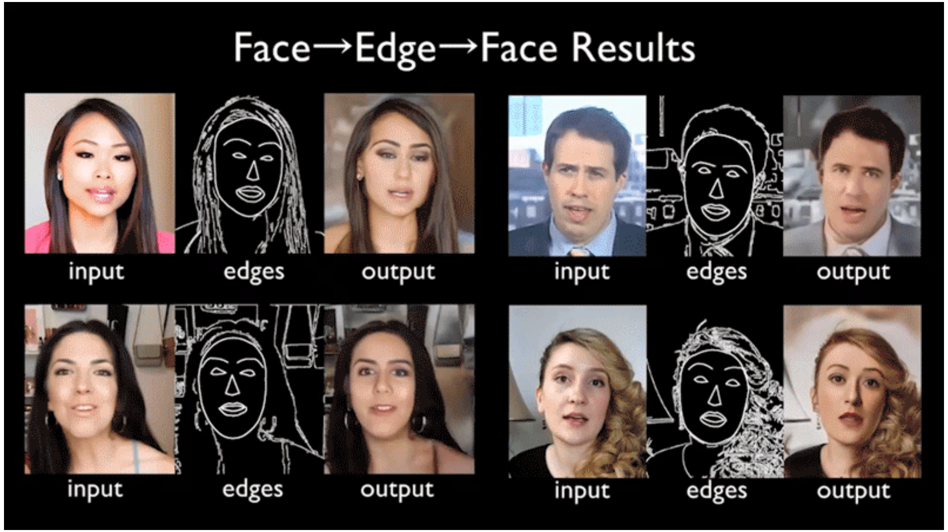
source image (GTA5)

adapted to Cityscape

CyCADA (Hoffman, 2017)

Video-to-video translation

Vid2Vid(Wang, 2018) : requires paired videos



Video-to-video translation

RecycleGAN (Bansal, 2018): video retargeting

Video-to-video translation

Difficulties compared with image-to-image translation

1. How to use temporal information to improve single frame quality?
2. How to make generated frames as consistent as the input?

results from cycleGAN

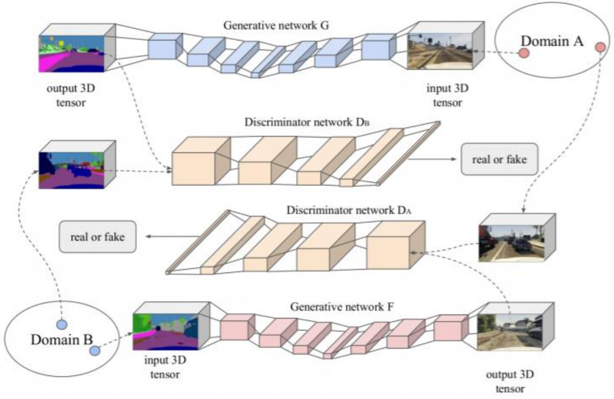
Video-to-video translation

results from vid2vid

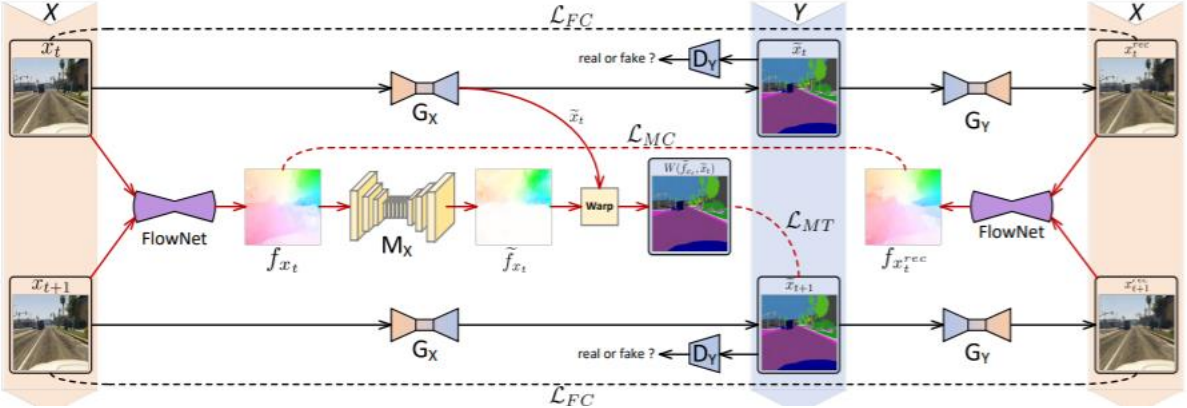
Video-to-video translation

How to tackle it in unspervised cases?

- Cycle-consistent loss (CycleGAN)
- 3-D convolution (BashKirova, 2018)
- Optical flow - Mocycle-gan (Chen, 2019)
- Other priors (CyCADA)



3D-conv + cyclegan



Mocycle-GAN

Transformation from synthetic colonoscopy video into real ones via optical temporal consistent GAN

Jiabo Xu, Ali Armin, Saeed Anwar, Nick Barnes

Introduction

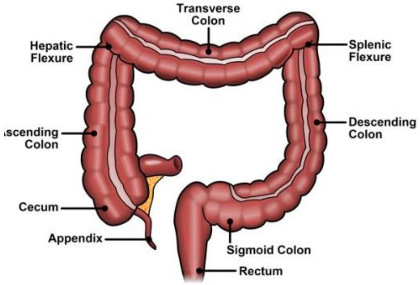
Medical images different from general ones:

- No general objects at all $\rightarrow$ hard to use pretrained model of general image. e.g. ImageNet
- Patient-specific $\rightarrow$ hard to create robust models
- Confidential $\rightarrow$ less amount of dataset
- Hard to label $\rightarrow$ less supervised tasks

Domain transformation:

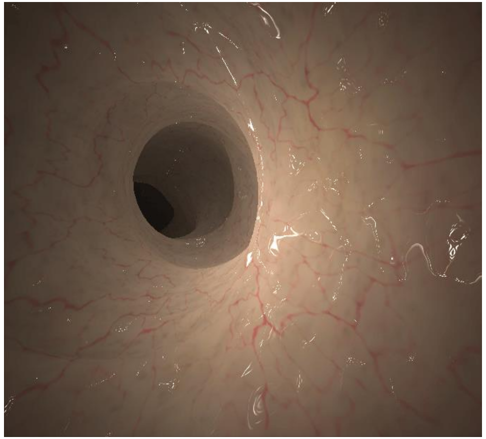
Transform labelled synthetic data into real-alike ones $\rightarrow$ labelled real dataset

Introduction



Colonoscopy data we have:

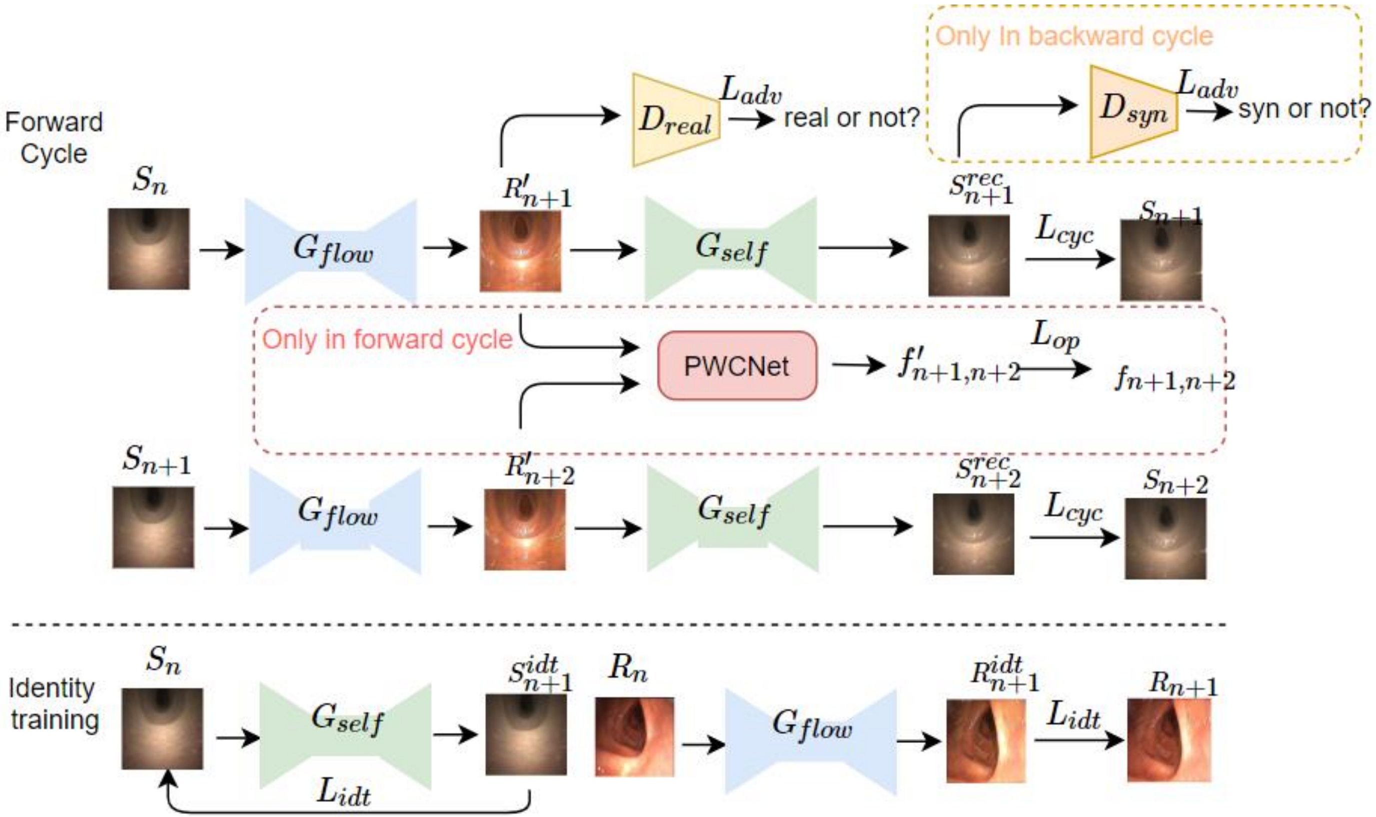
- Infinite synthetic data and corresponding annotations
- 2700 real data without labels


synthetic


optical flow


real

Methodology



Methodology

- Cycle structure like cycleGAN

$$L_{cyc} = E_{S_n \sim P_{data}(S)} ||G_{self}\left(G_{next}(S_n)\right) - S_{n+1}||_1 + E_{R_n \sim P_{data}(R)} ||G_{next}\left(G_{self}(R_n)\right) - R_{n+1}||_1$$

- Learn to predict the future frame

$$L_{idt} = E_{R_n \sim P_{data}(R)} \varphi(G_{next}(R_n) - R_{n+1}) + E_{S_n \sim P_{data}(S)} \varphi(G_{self}(S_n) - S_n)$$

- Optical flow loss

$$L_{op} = E_{S_n \sim P_{data}(S)} ||P\left(G_{next}(S_n), G_{next}(S_{n+1})\right) - f_{S_{n+1}, S_{n+2}}||_1$$

- Overall loss

$$L = L_{adv} + \lambda L_{cyc} + \beta L_{idt} + \sigma L_{op}$$

Experiments

Training details:

- 1400 cleaned real data, sample same size from 8000 synthetic data
- 200 epoch; batch-size 4; resize to (256, 256)
- Adam Learning rate = 2e-4, betas=(0.5, 0.999)
- $\lambda = 150, \beta = 75, \sigma = 0.1$
- Generator: resnet-6blocks; Discriminator: Patch-GAN (Li, 2016)
- Identity loss function: perceptual loss (Johnson, 2016); output layer: 2nd

- 4 Nvidia P100 on Bracewell for 24 hours

Experiments

Evaluations on:

- Image quality (Qualitative)
- Frame consistency / Annotation preservation (Quantitative)

Quantitative Metric:

- End point error to predicted(EPE-pred) -- only for comparison

$$EPE_{pred} = ||f'_{R'_n,R'_{n+1}} - f'_{S_n,S_{n+1}}||_1$$

- Conditioning on PWCnet to avoid optical flow estimation error.

Experiments -- Ablation Study

Model structure

- CycleGAN (baseline)
- CycleGAN + op
- TCGAN
- TCGAN + op (proposed)

Loss function

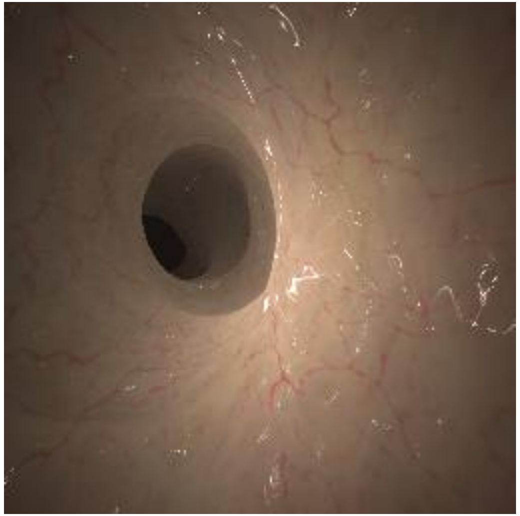
- Perceptual Loss
- L1

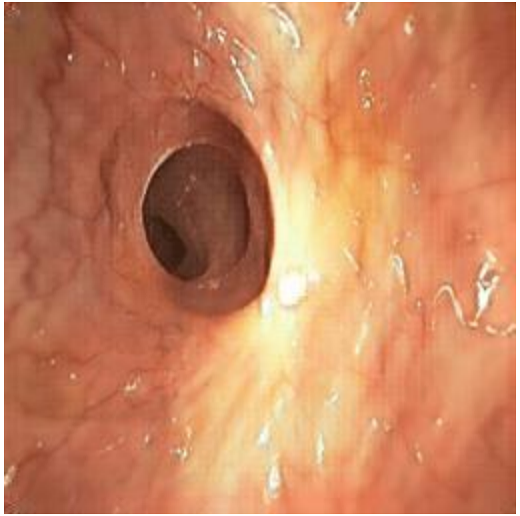
σ = 0.1 or 5 or 10

(weight of L_{op})

Generator: res6, res9, unet

TCGAN + op: other parameters following the training detail

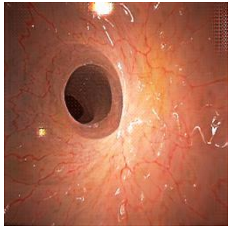




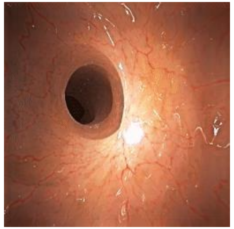
Experiments -- Ablation Study

Model structures

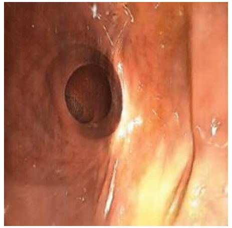
baseline (cycleGAN)



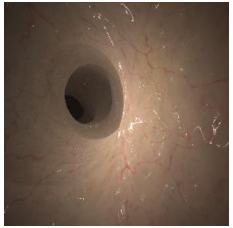
cycle+op



temp w/o op



input



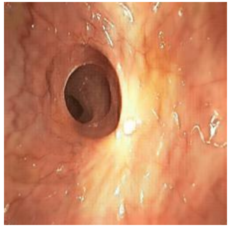
Base parameters:
resnet6, temp + op, x2, sig0.1

Network

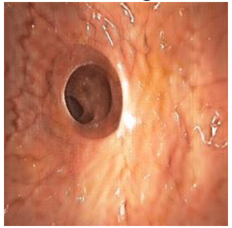
unet

Sigma & lossfunc

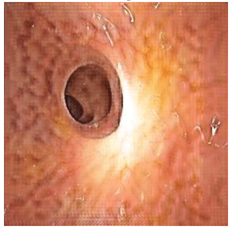
x2, sig0.1



x2, sig5



x2, sig10



x0, sig0.1

x2 means output of second conv block of pretrained vgg

x0 means the image itself.

Experiments -- Ablation Study

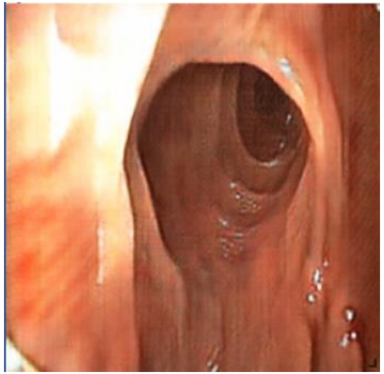
Approach	EPE-Pred	real-alike	no mask-noise	no bright-spot
synthetic	**0**	F	-	-
l1	1.38	T	F	T
unet	1.30	T	F	T
res9	0.59	T	F	F
sig10	0.40	F	-	-
sig5	0.38	T	T	F
cyclegan	**0.27**	F	-	-
cyclegan+op	0.61	F	-	-
TCGAN	2.41	T	F	T
TCGAN+op	0.35	T	T	F

Experiments -- Other models

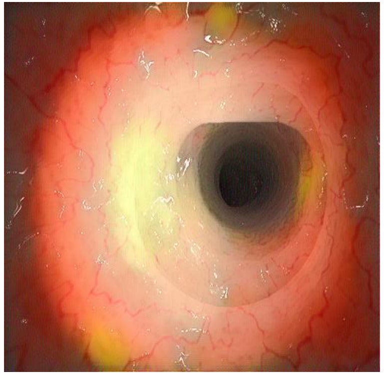
All those models fail to achieve the goal:

Deep Residual U-net CycleGAN(Oda, 2019)
- Irrational mask noise

RecycleGAN (Bansal, 2018)
- Cannot preserve structures

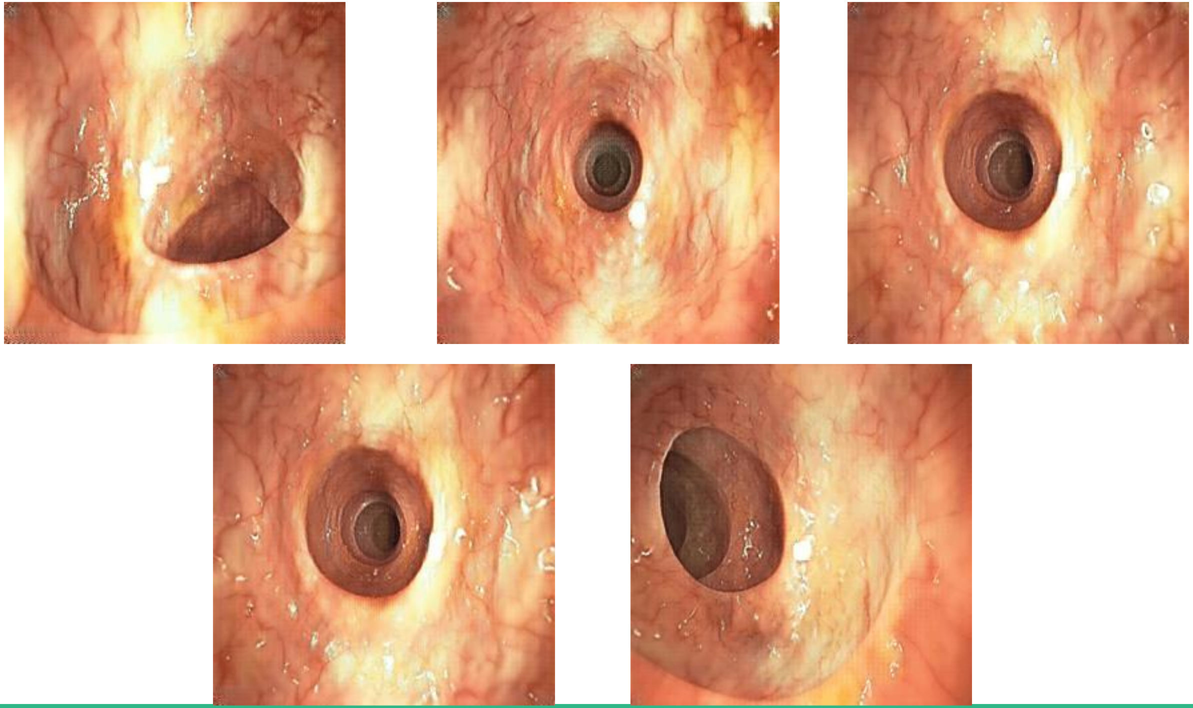


Style transfer (Gatys, 2016)
- Only for art!



Experiments

Our model is very robust on the whole dataset rather than single videos



Conclusion & Future Work

Contribution:

1. Proposed an innovative model that successfully transforms our synthetic colonoscopy video into realistic ones while preserve the structure.
2. Create a labelled synthetic-real colonoscopy dataset which can be use for supervised tasks directly.

Future work:

Generalize the performance on CT colonoscopy videos.

Improve supervised-task performance by using the generated dataset