Generative Adversarial Networks for Video-to-Video Domain Adaptation

Jiawei Chen, Yuexiang Li[*]**, Kai Ma, Yefeng Zheng**
YouTu Lab, Tencent, Shenzhen, China
{jiaweichen, vicyxli, kylekma, yefengzheng}@tencent.com

Abstract

Endoscopic videos from multicentres often have different imaging conditions, e.g., color and illumination, which make the models trained on one domain usually fail to generalize well to another. Domain adaptation is one of the potential solutions to address the problem. However, few of existing works focused on the translation of video-based data. In this work, we propose a novel generative adversarial network (GAN), namely VideoGAN, to transfer the video-based data across different domains. As the frames of a video may have similar content and imaging conditions, the proposed VideoGAN has an X-shape generator to preserve the intra-video consistency during translation. Furthermore, a loss function, namely color histogram loss, is proposed to tune the color distribution of each translated frame. Two colonoscopic datasets from different centres, i.e., CVC-Clinic and ETIS-Larib, are adopted to evaluate the performance of domain adaptation of our VideoGAN. Experimental results demonstrate that the adapted colonoscopic video generated by our VideoGAN can significantly boost the segmentation accuracy, i.e., an improvement of 5%, of colorectal polyps on multicentre datasets. As our VideoGAN is a general network architecture, we also evaluate its performance with the CamVid driving video dataset on the cloudy-to-sunny translation task. Comprehensive experiments show that the domain gap could be substantially narrowed down by our VideoGAN.

Introduction

The colorectal and stomach cancers are the leading causes of worldwide cancer deaths in 2018, accounting for 9.2% and 8.2% of total cancer deaths, respectively (Bray et al. 2018). The endoscopy is the primary imaging modality for screening and diagnosis of these cancers and over 100 million endoscopy exams are annually performed. However, the traditional screening approach requires specialized physicians to visually analyze extensive endoscopic videos, which is extremely laborious and suffers from various problems, e.g. inter-observer variations. With the recent development of deep learning, an increasing number of studies tried to develop computer-aid diagnosis (CAD) system for endoscopic
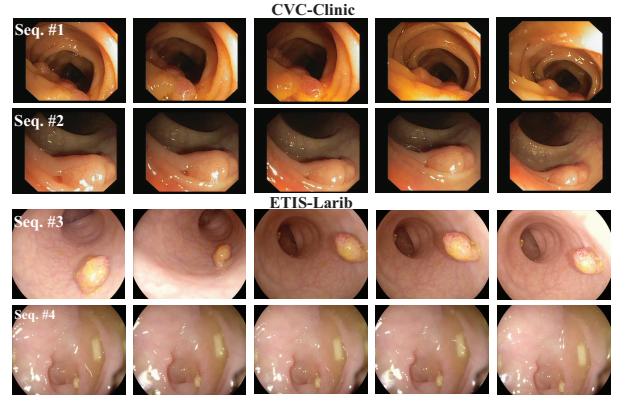
Figure 1: Examples of colonoscopy videos from multicentres. The colonoscopic images from CVC-Clinic and ETIS-Larib databases have different distributions of color and illumination.

images. For example, Zhang et al. (Zhang et al. 2018) proposed a regression-based convolutional neural network (R-CNN) pipeline for the automated detection of polyps during colonoscopy. Yu et al. (Yu et al. 2017) developed an offline and online 3D deep learning integration framework by leveraging the 3D fully convolutional network (3D-FCN) to tackle the problem of colorectal polyp detection. Although the existing studies significantly improved the diagnosis performance of CAD systems, the generality of the established models was merely investigated as data from single source domain was often used. The endoscopic images from multicentres usually have different imaging conditions such as color distribution and illumination effects, because imaging devices and imaging parameter settings are not well standardized. As shown in Fig. 1, the colonoscopic video frames captured by the CVC-Clinic centre (Vázquez et al. 2017) is warm-toned (higher red intensity) compared to the ones from the ETIS-Larib centre (Silva et al. 2014). Such variations would make models trained on one domain fail on another.

Domain adaptation is one of the potential solutions to address the problem of variations of imaging conditions among

multi-domains. Although the area of domain adaptation has been extensively studied, most of the existing works focus on the image-to-image domain adaptation (Isola et al. 2017; Zhu et al. 2017), which may not be the optimal solution for video-based endoscopic data. As shown in Fig. 1, the frames in the same endoscopic video have similar color distributions. To this end, specific constraints on video data, i.e., intra-video color consistency, also need to be applied during the domain adaptation process.

In this paper, we propose a novel generative adversarial network, namely VideoGAN, for domain adaptation of multicentre endoscopic videos. Our VideoGAN addresses two key problems that are overlooked by the current GAN-based translation approaches: 1) video content distortions, and 2) intra-video color inconsistency. To maintain the content and intra-video color consistencies during the domain adaptation process, the proposed VideoGAN model incorporates an X-shape generator that simultaneously translates two frames, i.e., the reference and source frames, to the same mode in the target domain. The mode defined for video-to-video domain adaptation is a video clip containing similar contents, which is totally different from that of the image-to-image translation, i.e., a single image. Meanwhile, we also want to preserve the relative shift in color distributions between the source and reference frames, e.g., if the source frame is brighter than the reference frame, this trend should be preserved after translation. To this end, we propose a novel color histogram loss that calculates the difference between two histograms of source and reference frames.

The proposed VideoGAN is evaluated on publicly available colonoscopic video datasets. The results illustrate that our VideoGAN overcomes the distortion problem of standard CycleGAN and yields elegant adapted videos, which significantly boosts the segmentation accuracy of colorectal polyps. Furthermore, as our VideoGAN is a general network architecture, we also evaluate its performance on natural video datasets. The experiments demonstrate that the proposed VideoGAN is suitable for natural video adaptation as well.

Related Work

Few existing works focus on domain adaptation of videos. Two closely related areas to this work are image-to-image translation and unsupervised domain adaptation. Hence, we briefly summarize those two categories of works in this section.

Image-to-image translation

Since firstly proposed by Goodfellow et al. (Goodfellow et al. 2014), GAN attracts increasing attentions from the community. Many tasks have been successfully tackled by GAN and its variations (Chang et al. 2018; Chen, Lai, and Liu 2018; Ma et al. 2018), such as image synthesis (Wang et al. 2018b), super-resolution (Ledig et al. 2017) and image translation (Isola et al. 2017; Zhu et al. 2017; Chen, Lai, and Liu 2018; Ma et al. 2018). Image-to-image translation aims at constructing a pixel-to-pixel mapping between two domains. A representative method is the conditional GAN (Isola et al. 2017), which shows a strategy

to learn such translation mapping with a conditional setting to capture structure information. However, it requires paired cross-domain images as training data, which are often difficult to aquire. To perform unpaired image-to-image translation, several works (Zhu et al. 2017; Kim et al. 2017; Yi et al. 2017) have been recently proposed. Those works introduce a cycle consistency loss to loose the confinement of paired training images. The proposed VideoGAN adopts the idea of cycle consistency and extends the framework for unpaired video translation. We notice that there are some studies on the topic of video synthesis (Chen et al. 2017; Wang et al. 2018c). However, those works still require paired training data to perform video-to-video translation.

Unsupervised domain adaptation

Apart from image-to-image translation, another area related to our work is the unsupervised domain adaptation (UDA). The UDA aims to close the gap, such as the color distribution, between the source and target domains without target sample annotations. One common choice for domain adaptation is to establish a mapping in the feature subspace between two different domains (Sun, Feng, and Saenko 2016; Ganin et al. 2017). For example, Sun et al. (Sun, Feng, and Saenko 2016) proposed the CORrelation Alignment (CORAL) method for unsupervised domain adaptation, which minimizes domain shift by aligning the second-order statistics of source and target distributions. In more recent studies, researchers tried to adopt GANs to seek an optimal feature space to build the mapping between two domains (Hoffman et al. 2018; Sankaranarayanan et al. 2018; Huang et al. 2018; Chen et al. 2019). Hoffman et al. (Hoffman et al. 2018) developed an approach, called cycle-consistent adversarial domain adaptation (CyCADA), to guide transfer between domains according to a discriminatively trained network. The approach alleviated the divergence problem by enforcing consistency of the relevant semantics before and after adaptation. Recently, the GAN-based domain adaptation is also widely used in the area of person re-identification (re-ID) (Deng et al. 2018; Wei et al. 2018) and medical image processing, e.g., the color normalization of histopathological slices (Zanjani et al. 2018), intensity standardization of magnetic resonance images (Gao et al. in press 2018) and cross-modality adaptation (Zhang, Yang, and Zheng 2018). However, most of the existing GAN-based domain adaptation approaches require strong prior-knowledge, e.g., pixel-wise annotations (Zhang, Yang, and Zheng 2018; Huang et al. 2018), which limits their applications for general tasks. *Our VideoGAN looses the requirements of prior-knowledge and achieves domain adaptation without any supervision signal.*

Proposed Method

In this section, we introduce VideoGAN in details. We first revisit the principle of CycleGAN and then introduce the proposed X-shape generator, color validator and color histogram loss.
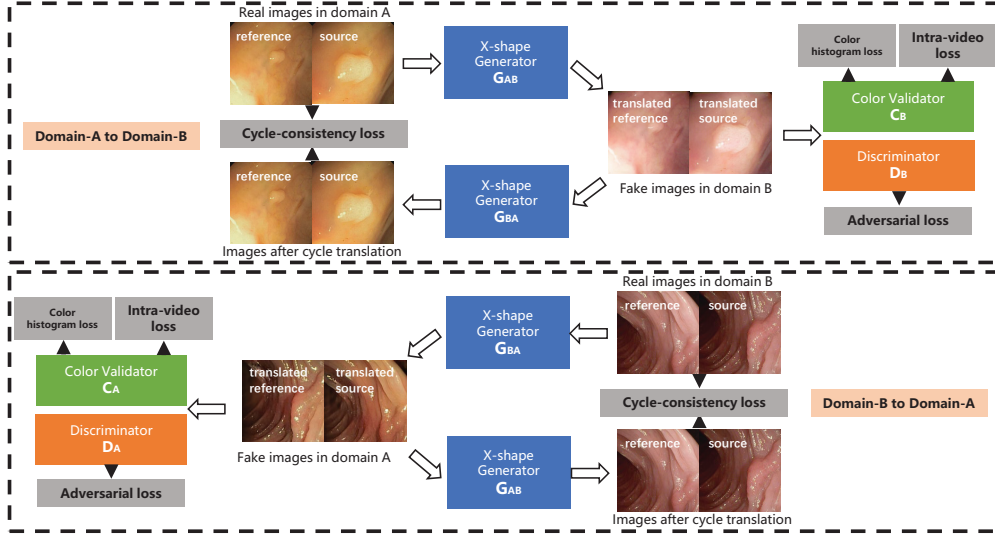
Figure 2: The overview of our VideoGAN. The proposed VideoGAN translates endoscopic videos between the domain-A (CVC-Clinic) and domain-B (ETIS-Larib). The generators (G_{AB} and G_{BA}) are supervised by a cycle-consistency loss, an adversarial loss, an intra-video loss and a color histogram loss. To preserve the intra-video consistency, a reference frame is sent with the source frame together as input to the X-shape generator.

Revisit of CycleGAN

CycleGAN has two paired generator-discriminator modules, which are capable of learning two mappings, i.e., from domain A to domain B $\{G_{AB}, D_B\}$ and vice versa $\{G_{BA}, D_A\}$. The generators (G_{AB}, G_{BA}) translate images between the source and target domains, while the discriminators (D_A, D_B) aim to distinguish the real and translated data. Thereby, the generators and discriminators are gradually updated during this adversarial competition.

The original CycleGAN is supervised by two losses, i.e., adversarial loss ($\mathcal{L}_{adv}$) and cycle-consistency loss ($\mathcal{L}_{cyc}$). The adversarial loss encourages local realism of the translated data. Taking the translation from domain B to domain A as an example, the adversarial loss can be written as:

$$\mathcal{L}_{adv}(G_{BA}, D_A) = \mathbb{E}_{x_A \sim p_{x_A}} \left[(D_A(x_A) - 1)^2 \right] \\ + \mathbb{E}_{x_B \sim p_{x_B}} \left[(D_A(G_{BA}(x_B)))^2 \right] \quad (1)$$

where p_{x_A} and p_{x_B} denote the sample distributions of domain A and B, respectively; x_A and x_B are samples from domain A and B, respectively.

The cycle-consistency loss ($\mathcal{L}_{cyc}$) tackles the problem of deficient paired training data. The idea behind the cycle-consistency loss is that the translated data from the target domain can be exactly converted back to the source domain, which can be expressed as:

$$\mathcal{L}_{cyc}(G_{BA}, G_{AB}) = \mathbb{E}_{x_A \sim p_{x_A}} \left[\|G_{BA}(G_{AB}(x_A)) - x_A\|_1 \right] \\ + \mathbb{E}_{x_B \sim p_{x_B}} \left[\|G_{AB}(G_{BA}(x_B)) - x_B\|_1 \right] \quad (2)$$

where the $\mathcal{L}_1$ loss is adopted in $\mathcal{L}_{cyc}$ in our VideoGAN.

Problems of using CycleGAN in video-to-video domain adaptation: First, due to the intrinsic ambiguity with respect to geometric transformations of CycleGAN (Zhang, Yang, and Zheng 2018), content distortion may exist in the translated results produced by CycleGAN. Second, as the CycleGAN separately transfers each frame of the videos, the frames may be mapped to different modes of the target domain. Such problems violate the intuition of domain adaptation of videos, but they have not been addressed in existing methods.

VideoGAN

We propose the VideoGAN to solve the aforementioned problems of using CycleGAN in the video-to-video domain adaptation. Fig. 2 illustrates the workflow of our VideoGAN. The proposed VideoGAN shares a similar cyclic processing procedure as CycleGAN. To preserve the intra-video consistency and prevent the content distortion, however, we set a reference frame with the source frame as a dual-input to the generator, which also generates two corresponding frames in the target domain. We name the generator as X-shape generator because the dual-in-dual-out and the encode-decode properties formulate the network structure in an X-shape. Beyond the regular discriminator that distinguishes real/fake images, we develop another color validation module (referred as color validator for convenience) that tunes the color distribution of the source frame according to the paired reference frame. Supervised by the proposed losses (adversarial loss, cycle-consistency loss, color histogram loss and intra-video loss), VideoGAN is trained in an end-to-end fashion that follows the protocol proposed in (Zhu et al. 2017).

X-shape generator. In order to maintain the consistency, i.e., content and color, of generated frames belonging to the

same video, the proposed X-shape generator simultaneously translates two images, i.e., source and reference frames, from the source domain to the target domain. As there are multiple modes in the target domain, to bring the generated frames to a certain mode, the reference frame acts as an anchor to pull other source frames to be into the same mode. Hence, the reference frame (anchor) should be invariant during the translation of the whole video. In our experiments, the first frame of each video[1] is chosen as the reference frame, which is paired with each remaining frame of the corresponding video and sent to the generator as the input. Consequently, the reference frame appears multiple times during training and gains more weights than other source frames, which can help determine the marching direction of the rest video frames. Furthermore, the extracted features of the reference and source frames are embedded into a new feature space by a dense fusion block, where the fusion process can be seen as adding mutual watermarks to both frames. The content distortions caused by the bijective geometric transformation and its inverse operation (Zhang, Yang, and Zheng 2018) can be prevented, since they may scramble the latent watermarks and do not maintain the cycle consistency any more.
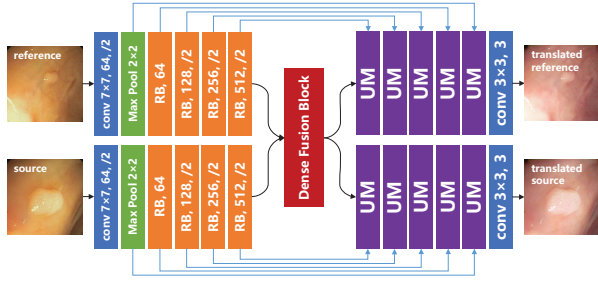

Figure 3: The proposed X-shape generator. The source and reference frames have separate encoder-decoder modules. The size of source and reference frames is $256 \times 256 \times 3$. The extracted features from the two frames are fused in the feature space by a dense fusion block. The blue, green, orange, red and purple rectangles represent the convolution layer, max-pooling layer, residual block, dense fusion block and upsampling layer, respectively.

Fig. 3 presents the network architecture of our X-shape generator. The blue, green, orange, red and purple rectangles represent the convolution layer, max-pooling layer, residual block (RB) (He et al. 2016), dense fusion block and upsampling module (UM, consisting of a 3×3 convolution layer and an upsampling layer), respectively. As shown in Fig. 3, the source and reference frames are processed by independent encoders and decoders. The extracted high-level features are fused by the dense fusion block to preserve the intra-video consistency and alleviate the problem of content distortion. The short-cut connection (Ronneberger, Fischer, and Brox 2015) is employed to facilitate the feature infor-

mation flowing between the encoders and decoders.

Dense fusion block. The dense fusion block ensures that the source frames are mapped to the same mode in the target domain, and prevents content distortions by embedding content information of the reference frame to the source frames. Fig. 4 shows the architecture of our dense fusion block. The green, gray and yellow rectangles represent the average pooling, concatenation and resizing layers, respectively. Naive fusion approaches, e.g., stacking two images as different input channels or fusing the feature maps generated at the middle stages of encoders, suffer from the spatial information interference between the paired frames. Hence, the 8×8 feature maps from the source and reference frames are average-pooled to 1×1 size before the fusion. The symbol B in Fig. 4 stands for the batchsize and 512 is the number of feature maps. Inspired by the dense upsampling convolution (Wang et al. 2018a), a 1×1 convolution with $512 \times 8 \times 8$ channels is used to fuse the information of source and reference frames. The fused $32768 \times 1 \times 1$ feature maps are then resized back to size of $512 \times 8 \times 8$ and sent to the decoders.
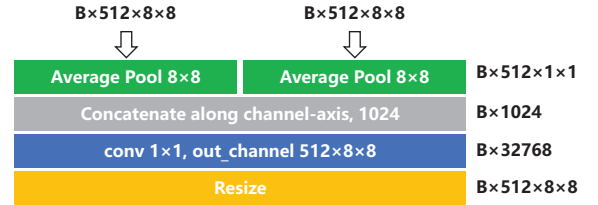

Figure 4: The proposed dense fusion block. It uses the average pooling layer (green rectangles) and 1×1 convolution to fuse the high-level features of the source and reference frames.

Color validator. Color validators[2] (C_A, C_B) in our VideoGAN perform two tasks. First, they regulate the color variation trend of the translated frames with the color histogram loss. For example, suppose the source frame is brighter than the reference frame, we want to maintain the relationship after mapping the source frame to the target domain. Second, they act as secondary discriminators assessing the real/fake identities of the paried frames with an intra-video loss. In this setting, the actual samples input to the color validators are any two frames of the same video from target domain. Therefore, to trick such color validators, the generators latently maintain the long-term intra-video consistency. The color validator of our VideoGAN takes the two translated frames as the input and yields a 1×15 (C^{hist}) and a 1×1 ($C^{i.v.}$) vector as predictions to calculate the color histogram loss and intra-video loss, respectively.

Color histogram loss. The reference frame in our VideoGAN provides the identity information as well as the relative color distribution, which can be used to tune the color of source frames. Assuming the histogram of each of RGB channels can be expressed as $hist_c$ ($c \in R, G, B$), the

---

[1]For a long video, we truncate it into a set of clips based on scene changes. And, each video clip has a consistent scene setting.

[2]Detailed information of color validator can be found in the full version at arXiv.

relative color distribution ($hist_{rcd}$) can be defined as:

$$hist_{rcd} = cat(hist_R^{ref}, hist_G^{ref}, hist_B^{ref}) \\ - cat(hist_R^{src}, hist_G^{src}, hist_B^{src}) \quad (3)$$

where the $hist^{src}$ and $hist^{ref}$ refer to the histograms of source and reference frames, respectively; $cat(.)$ is the concatenation operation.

The translated frames are expected to maintain the information of relative color distribution after domain adaptation. Hence, we calculate a 5-bin histogram for each of RGB channels and form a 15-bin $hist_{rcd}$ as ground truth to supervise the VideoGAN. The color histogram loss ($\mathcal{L}_{hist}$) can thereby be written as:

$$\mathcal{L}_{hist}(G_{BA}, C_A^{hist}) = \| C_A^{hist}(G_{BA}(x_B^{src}), G_{BA}(x_B^{ref})) \\ - hist_{rcd}(x_B^{src}, x_B^{ref}) \|_1 \quad (4)$$

where the $\mathcal{L}_1$ loss is adopted in $\mathcal{L}_{hist}$ in our VideoGAN; x_B^{src} and x_B^{ref} are the source and reference frames from domain B, respectively.

Intra-video loss. The intra-video loss ($\mathcal{L}_{i.v.}$) enables our VideoGAN to maintain the intra-video consistency for lengthy videos, such as the ones in CamVid dataset.[3] The color validator distinguishes the translated paired frames from the actual frames selected from the same video of the target domain, which enforces the generator to consider long-term intra-video consistency. The intra-video loss has the same form as the adversarial loss ($\mathcal{L}_{adv}$), which can be denoted as $\mathcal{L}_{i.v.}(G_{BA}, C_A^{i.v.})$.

Discriminator. In consistent with the standard Cycle-GAN, the proposed VideoGAN has discriminators (D_A, D_B) to validate the quality of each translated frame. The SegNet (Badrinarayanan, Kendall, and Cipolla 2017) is adopted to provide pixel-wise prediction of real or fake, rather than to classify the whole frame or sub-image. Such an approach encourages the discriminators to take more information into account while making real/fake decisions.

Objective. Apart from the cycle-consistency, adversarial, color histogram and intra-video losses defined above, we also involve the identity constraint (Deng et al. 2018), i.e., $\mathcal{L}_{idt}$, as an auxiliary loss. The target domain identity constraint regularizes the generator to be the identity matrix on samples from target domain. Finally, our full objective function is summarized as:

$$\mathcal{L}(G_{BA}, G_{AB}, D_A, D_B, C_A, C_B) = \mathcal{L}_{adv}(G_{BA}, D_A) \\ + \mathcal{L}_{adv}(G_{AB}, D_B) + \mathcal{L}_{hist}(G_{BA}, C_A^{hist}) \\ + \mathcal{L}_{hist}(G_{AB}, C_B^{hist}) + \mathcal{L}_{i.v.}(G_{BA}, C_A^{i.v.}) \quad (5) \\ + \mathcal{L}_{i.v.}(G_{AB}, C_B^{i.v.}) + \mathcal{L}_{cyc}(G_{BA}, G_{AB}) \\ + \mathcal{L}_{idt}(G_{BA}, G_{AB})$$

The optimization of $\mathcal{L}_{cyc}$ and $\mathcal{L}_{adv}$ is in an alternative manner following the protocol proposed in (Zhu et al. 2017). The $\mathcal{L}_{idt}$ is updated at the same time of $\mathcal{L}_{cyc}$, while the $\mathcal{L}_{hist}$ and $\mathcal{L}_{i.v.}$ are updated with the $\mathcal{L}_{adv}$ optimization.

[3]http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid

Experiments

This section evaluates and discusses the performance of our VideoGAN. Suffering from the problem of multicentres, the conditions of endoscopic videos often have wide variations, which dramatically degrade the generalization of a computer-aid diagnosis system trained on a single domain. In this section, we begin with tackling the multicentre problem for endoscopic videos using our VideoGAN. As the proposed VideoGAN is a general network architecture, we then evaluate its domain adaptation performance on the CamVid video dataset.

Datasets. The publicly available colonoscopic video datasets, i.e., CVC-Clinic[4] (Vázquez et al. 2017) and ETIS-Larib[5] (Silva et al. 2014), are selected for our experiments. The CVC-Clinic dataset is composed of 29 sequences with a total of 612 images. The ETIS-Larib consists of 196 images, which can be manually separated to 29 seuqences as well. Those short videos are extracted from the colonoscopy videos captured by different centres using different endoscopic devices. All the frames of the short videos contain polyps. The dataset providers annotate the pixel-wise ground truth covering the polyps for data users. Sample frames of the two datasets are shown in Fig. 1.

Training details. The proposed VideoGAN is implemented using PyTorch. The Adam solver (Kingma and Ba 2014) with betas = (0.5, 0.999) is adopted for the optimization of VideoGAN. The network is trained with a mini-batch size of 1 on one GPU (Tesla P40 with 24 GB memory). The initial learning rate is set to 0.0002. The proposed VideoGAN yields visually satisfactory translated frames after 200 training epochs.

Experiment settings. Given two datasets (A, B), our goal is to narrow down the gap between them not only in terms of visual perception i.e., elegant translated results, but also the representation in feature space, i.e., improvement of the robustness of models. We present the video translation results to evaluate the former factor. For the latter one, the learning via translation framework (Liu, Breuel, and Kautz 2017; Deng et al. 2018) is adopted for the evaluation. The framework consists of two steps, i.e., cross-domain image translation for training data creation, and supervised feature learning for a specific task, e.g., polyp segmentation. Accordingly, we evaluate VideoGAN in two scenarios: **transfer learning** (CVC-Clinic → ETIS-Larib) and **data augmentation** (ETIS-Larib → CVC-Clinic), respectively. In transfer learning, the target domain data has no labels. We use the target domain data to train CycleGAN and VideoGAN to translate the source domain data to the target domain. The translated source domain data are then used to train a segmentation network for the transfer learning on target domain. In data augmentation, the translated source domain data are added to the labeled target domain data to train a segmentation network.

[4]https://polyp.grand-challenge.org/CVCClinicDB/
[5]https://polyp.grand-challenge.org/EtisLarib/

Evaluation criterion. The Dice coefficient, which measures the spatial overlap index between the segmentation result and ground truth, is adopted as the metric to assess the accuracy of colorectal ployp segmentation.

Baselines overview. Based on the experiment settings, several unpaired image-to-image translation frameworks, i.e., G(.), are taken as baselines for the performance evaluation, including CycleGAN (Zhu et al. 2017), UNIT (Liu, Breuel, and Kautz 2017) and DRIT (Lee et al. 2018). We also compare VideoGAN with a direct transfer approach, which directly takes the source domain data for training without any adaptation. Note that the recent proposed GANs for image-based domain adaptation, e.g., SPGAN (Deng et al. 2018), PTGAN (Wei et al. 2018) and AugGAN (Huang et al. 2018), are not involved for comparison, due to the strong prior-knowledge used in those approaches. SPGAN used the prior-knowledge that the ID sets of different re-ID domains used in their experiments are totally different from each others. PTGAN required coarse segmentation results to distinguish foreground and background areas. AugGAN added a segmentation subtask to the CycleGAN-based framework, which required pixel-wise annotations. The use of prior-knowledge degrades the generalization of those GANs, which are only suitable for the domains fulfilling the specific requirements.

Visualization of domain adaptation results

The domain adaptation results between CVC-Clinic and ETIS-Larib domains are shown in Fig. 5, which illustrates two main problems of existing image-to-image translation approaches (UNIT (Liu, Breuel, and Kautz 2017), DRIT (Lee et al. 2018) and CycleGAN (Zhu et al. 2017)), i.e., the content distortion and random color variation. The CycleGAN taking the source and reference frames as two input channels is also adopted for comparison. Due to the interference of spatial information of source and reference frames, the CycleGAN with stacked input fails to generate meaningful translation results. As Fig. 5 shows, our VideoGAN can simutaneously maintain the video contents and the trend of color variation throughout the video.

Quantitative analysis of content distortions. To quantitatively evaluate the content distortion degrees, we invite an experienced senior physician to annotate the polyps in the original ETIS-Larib videos and the ones translated by Cycle-GAN and our VideoGAN, respectively. To evaluate the gap between the physician and the ETIS-Larib experts, the Dice coefficient between new annotations on the original ETIS-Larib videos and the ground truths is measured.

As Table 1 lists, due to inter-observer variability, the physician achieves a Dice coefficient of 82.73% on the original ETIS-Larib dataset. The annotation accuracy of physician on translated videos should thereby be close to 82.73%, if the contents are preserved through the translation. However, the Dice coefficient drops to 41.39% while the physician annotating the video translated by CycleGAN. Due to content distortion, a polyp may totally disappear or move to a different location in some translated images of Cycle-GAN, resulting in zero Dice coefficient, which deteriorates the overall Dice coefficient significantly. On the contrary,
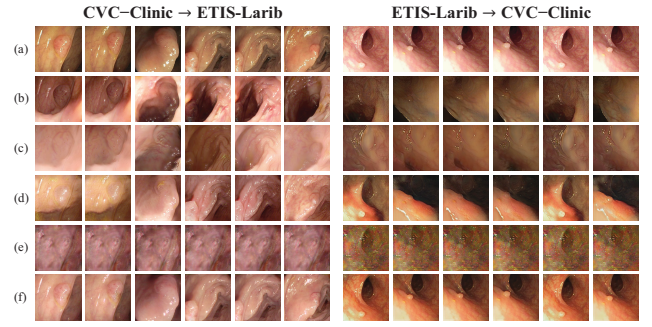

Figure 5: The domain adaptation videos cross CVC-Clinic and ETIS-Larib domains. (a) The original videos in source domains. (b)-(e) are the domain adaptation videos produced by UNIT, DRIT, CycleGAN, and CycleGAN with stacked input, respectively. (f) The domain adaptation videos produced by VideoGAN.

since the proposed VideoGAN well maintains video contents before and after the video translation, the physician achieves a 80.37% Dice coefficient on the videos translated by our VideoGAN.

Table 1: Dice coefficients between new annotations on different ETIS-Larib videos and the ground truth.

	Original	CycleGAN	VideoGAN
Dice (%)	82.73	41.39	80.37

Evaluation of learning via translation

As aforementoined, we conduct experiments to evaluate the proposed VideoGAN from two aspects, i.e., transfer learning and data augmentation.

Transfer learning. The size of ETIS-Larib dataset is extremely small (196 frames), which is insufficient to train a deep learning network. A common solution to this problem is the transfer learning, i.e., using additional related dataset, e.g. CVC-Clinic, as the training set for model learning. We evaluate the segmentation performance of FCNs, trained with CVC-Clinic videos translated by different GANs, on the ETIS-Larib dataset. ResUNet-50 (He et al. 2016; Ronneberger, Fischer, and Brox 2015) is chosen as the backbone for FCNs because it shows excellent performance on many segmentation tasks. Table 2 lists the polyp segmentation accuracies of FCNs trained with different training sets.

The training sets translated by UNIT, DRIT and Cycle-GAN decrease the segmentation accuracy of polyps, i.e., -44.96%, -31.50% and -2.26%, respectively, compared to that of direct transfer approach. Since no target domain data are used for training, the content distortion in the translated source domain data deteriorates the accuracy. Oppositely, our VideoGAN can achieve high-quality domain adaptation while maintaining the video contents,

Table 2: Comparison of Dice coefficients of polyp segmentation yielded by different domain adaptation methods on ETIS-Larib. The best result is in **bold**.

	Dice (%)
Direct Transfer	71.57
UNIT (Liu, Breuel, and Kautz 2017)	26.61
DRIT (Lee et al. 2018)	40.07
CycleGAN (Zhu et al. 2017)	69.31
VideoGAN	**76.69**

which leads to a significant improvement for the Dice coefficient, i.e., $+5.12\%$.

Data augmentation. In this experiment, we translate the ETIS-Larib videos to CVC-Clinic domain for data augmentation. ResUnet-50 (He et al. 2016; Ronneberger, Fischer, and Brox 2015) is adopted as the FCN as well. The FCNs trained with different training sets are evaluated on the additional CVC-Clinic-2018 dataset, consisting of 300 colonoscopic frames[6] (Vázquez et al. 2017). The polyp segmentation accuracies of FCNs with different domain adaptation methods are listed in Table 3.

Table 3: Comparison of Dice coefficients of polyp segmentation yielded by different domain adaptation methods on CVC-Clinic-2018. The best result is in **bold**.

	Dice (%)
No Augmentation	79.22
Mixing with Original ETIS-Larib	79.98
UNIT (Liu, Breuel, and Kautz 2017)	76.07
DRIT (Lee et al. 2018)	76.05
CycleGAN (Zhu et al. 2017)	81.14
VideoGAN	**83.50**

The FCN trained with the set directly mixing the CVC-Clinic training set and ETIS-Larib videos generates a marginal improvement, i.e., $+0.76\%$, compared to that of baseline, i.e., no augmentation. UNIT and DRIT deteriorate the accuracy due to severe content distortions. Surprisingly, CycleGAN achieves a marginal improvement of $+1.92\%$, probably, because in this data augmentation setting, target domain data outnumbers the translated source domain data and a small amount of distorted source data is ignored by the segmentation network. The proposed VideoGAN yields the highest improvement of Dice coefficient, i.e., $+4.28\%$, compared to the one without augmentations, which remarkably boosts the accuracy of polyp segmentation.

It is worthwhile to mention that the polyp segmentation is a new task announced in GIANA 2018 challenge, which uses the CVC-Clinic-2018 and CVC-Clinic datasets as the training and test set, respectively. As the challenge organizers only announced the team ranking rather than the top segmentation accuracies, we can not directly compare the

---

[6]https://giana.grand-challenge.org/PolypSegmentation/

Table 4: Ablation study of VideoGAN for the transfer learning (TF) and data augmentation (DA) tasks (**Dice (%)**).

	Setup	TF	DA
A	Original CycleGAN	69.31	81.14
B	A + X-G	74.30	82.27
C	B + L_{hist}	75.86	82.91
D	C + $L_{i.v.}$	76.69	83.50
E	A + weight-sharing X-G	68.24	79.20

performance of our approach with state-of-the-art on this dataset.

Ablation study. To evaluate the improvement yielded by each component of VideoGAN, an ablation study is conducted on the transfer learning (TF) and data augmentation (DA) tasks. The results of ablation study are presented in Table. 4. The result shows that all of our components can improve the segmentation accuracy – X-shape generator (X-G) ($+4.99\%$ and $+1.13\%$), color histogram loss (L_{hist}) ($+1.56\%$ and $+0.64\%$) and intra-video loss ($L_{i.v.}$) ($+0.83\%$ and $+0.59\%$) – for the TF and DA tasks, respectively.

Furthermore, a variant of X-G, i.e., weight-sharing X-shape generator (X-G), is involved for comparison. The weight-sharing X-G shares weights of the last resiudal block and the first upsampling module – as same as (Liu, Breuel, and Kautz 2017) – for the source and reference frames instead of using dense fusion block. As the weight-sharing X-G easily suffers from the spatial information interference between the paired frames, which may result in the content distortions, degradations of segmentation accuracy (-1.07% and -1.94%) are respectively observed for the TF and DA tasks, compared to the original CycleGAN.

Comparison of translated results on CamVid. We also evaluate domain adaptation performance of the proposed VideoGAN on the CamVid dataset. The experimental results are presented in **Appendix**.[7]

Conclusion and Future Work

In this paper, we present a framework, namely VideoGAN, for the domain adaptation of video-based data. To our best knowledge, this is the first work to address the problem of video-to-video domain adaptation. We evaluate the domain adaptation performance of our VideoGAN on the endoscopic and natural videos. The experimental results demonstrate that our VideoGAN can significantly narrow down the gap between different domains.

We plan to improve the performance of our VideoGAN from two aspects. First, the RGB color space is currently used to calculate the color histogram loss. However, we notice that there are some other choices, e.g. CIELab color space, that can provide more accurate color representations compared to RGB color space. Furthermore, the 3D color histogram is also a potential choice for the color histogram loss, replacing the 15-bin histogram. Second, the recurrent

---

[7]Appendix can be found in the full version at arXiv.

neural network (RNN) may be a better solution for the domain adaptation of lengthy videos. However, the whole network may be difficult to train while integrating RNN to GAN, due to its complicated architecture. We will make our best effort to address this problem in the future work.

Acknowledgements

References

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(12):2481–2495.

Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegeland, R. L.; Torreand, L. A.; and Jemal, A. 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* 68(6):394–424.

Chang, H.; Lu, J.; Yu, F.; and Finkelstein, A. 2018. PairedCycleGAN: Asymmetric style transfer for applying and removing makeup. In *CVPR*.

Chen, D.; Liao, J.; Yuan, L.; Yu, N.; and Hua, G. 2017. Coherent online video style transfer. In *ICCV*.

Chen, M. H.; Kira, Z.; AlRegib, G.; Woo, J.; Chen, R.; and Zheng, J. 2019. Temporal attentive alignment for large-scale video domain adaptation. In *ICCV*.

Chen, Y.; Lai, Y.; and Liu, Y. 2018. CartoonGAN: Generative adversarial networks for photo cartoonization. In *CVPR*.

Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; and Jiao, J. 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. S. 2017. Domain adversarial training of neural networks. *Journal of Machine Learning Research* 17(1).

Gao, Y.; Liu, Y.; Wang, Y.; Shi, Z.; and Yu, J. in press, 2018. A universal intensity standardization method based on a many-to-one weak-paired cycle generative adversarial network for magnetic resonance images. *IEEE Transactions on Medical Imaging*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NeurIPS*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hoffman, J.; Tzeng, E.; Park, T.; and Zhu., J.-Y. 2018. CyCADA: Cycle-consistent adversarial domain adaptation. In *ICML*.

Huang, S.; Lin, C.; Chen, S.; Wu, Y.; Hsu, P.; and Lai, S. 2018. AugGAN: Cross domain adaptation with GAN-based data augmentation. In *ECCV*.

Isola, P.; Zhu, J. Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*.

Kim, T.; Cha, M.; Kim, H.; Lee, J.; and Kim, J. 2017. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*.

Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; and Shi, W. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*.

Lee, H. Y.; Tseng, H. Y.; andM. K. Singh, J. B. H.; and Yang, M. H. 2018. Diverse image-to-image translation via disentangled representations. In *ECCV*.

Liu, M. Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *NeurIPS*.

Ma, S.; Fu, J.; Chen, C. W.; and Mei, T. 2018. DA-GAN: Instance-level image translation by deep attention generative adversarial networks. In *CVPR*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

Sankaranarayanan, S.; Balaji, Y.; Castillo, C. D.; and Chellappa, R. 2018. Generate to adapt: Aligning domains using generative adversarial networks. In *CVPR*.

Silva, J.; Histace, A.; Romain, O.; Dray, X.; and Granado, B. 2014. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery* 9(2):283–293.

Sun, B.; Feng, J.; and Saenko, K. 2016. Return of frustratingly easy domain adaptation. In *AAAI*.

Vázquez, D.; Bernal, J.; Sánchez, F. J.; Fernández-Esparrach, G.; López, A. M.; Romero, A.; Drozdzal, M.; and Courville, A. 2017. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering* 2017.

Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; and Cottrell, G. 2018a. Understanding convolution for semantic segmentation. In *WACV*.

Wang, T.; Liu, M.; Zhu, J.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018b. High-resolution image synthesis and semantic manipulation with conditional GANs. In *CVPR*.

Wang, T. C.; Liu, M. Y.; Zhu, J. Y.; Liu, G.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018c. Video-to-video synthesis. In *NeurIPS*.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person transfer GAN to bridge domain gap for person re-identification. In *CVPR*.

Yi, Z.; Zhang, H.; Tan, P.; and Gong, M. 2017. DualGAN: Unsupervised dual learning for image-to-image translation. In *ICCV*.

Yu, L.; Chen, H.; Dou, Q.; Qin, J.; and Heng, P. A. 2017. Integrating online and offline 3D deep learning for automated polyp detection in colonoscopy videos. *IEEE Journal of Biomedical and Health Informatics* 21(1):65–75.

Zanjani, F. G.; Zinger, S.; Bejnordi, B. E.; van der Laak, J. A. W. M.; and de With, P. H. N. 2018. Stain normalization of histopathology images using generative adversarial networks. In *ISBI*.

Zhang, R.; Zheng, Y.; Poon, C. C. Y.; Shen, D.; and Lau, J. Y. W. 2018. Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. *Pattern Recognition* 83:209–219.

Zhang, Z.; Yang, L.; and Zheng, Y. 2018. Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network. In *CVPR*.

Zhu, J.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.