Self-Supervised CycleGAN for Object-Preserving Image-to-Image Domain Adaptation

Xinpeng Xie[1], Jiawei Chen[2⋆], Yuexiang Li[2], Linlin Shen[1], Kai Ma[2], and Yefeng Zheng[2]

[1] Computer Vision Institute, Shenzhen University, Shenzhen, China
xiexinpeng2017@email.szu.edu.cn llshen@szu.edu.cn
[2] Tencent Jarvis Lab, Shenzhen, China
{jiaweichen,vicyxli,kylekma,yefengzheng}@tencent.com

Abstract. Recent generative adversarial network (GAN) based methods (e.g., CycleGAN) are prone to fail at preserving image-objects in image-to-image translation, which reduces their practicality on tasks such as domain adaptation. Some frameworks have been proposed to adopt a segmentation network as the auxiliary regularization to prevent the content distortion. However, all of them require extra pixel-wise annotations, which is difficult to fulfill in practical applications. In this paper, we propose a novel GAN (namely OP-GAN) to address the problem, which involves a self-supervised module to enforce the image content consistency during image-to-image translations without any extra annotations. We evaluate the proposed OP-GAN on three publicly available datasets. The experimental results demonstrate that our OP-GAN can yield visually plausible translated images and significantly improve the semantic segmentation accuracy in different domain adaptation scenarios with off-the-shelf deep learning networks such as PSPNet and U-Net.

Keywords: Image-to-Image translation, domain adaptation, semantic segmentation.

1 Introduction

Deep learning networks have shown impressive successes on various computer vision tasks such as image classification [9,19,28] and semantic segmentation [2,11,33]. However, most of current deep learning based approaches easily suffer from the problem of domain shift—the models trained on a dataset (source) seldom maintain the same performance on other datasets (target) obtained under different conditions. Image-to-image (I2I) translation is one of the potential solutions to address the problem by enforcing the input data distributions of two domains to be similar. Due to the recent success of generative adversarial network (GAN) [8] on generating high-quality synthetic images, many studies

⋆ The first two authors are equal contribution.

adopted GANs for the I2I domain adaptation [12,36], which authentically convert an input image to a corresponding output image by constructing a pixel-to-pixel mapping. As a representative method, Pix2Pix [12] shows a strategy to learn such adaptation mapping with a conditional setting to capture structure information. However, it requires paired cross-domain images as training data, which are often difficult to acquire.

To loose the requirement of pairwise training images, GAN-based unpaired I2I domain adaptation methods, e.g., CycleGAN [36], DiscoGAN [13], and DualGAN [32] were recently proposed, where a cycle consistency constraint was applied to encourage bidirectional image translations with regularized structural output. Although these GANs present realistic visual results on several I2I translation tasks, the corruptions of image content are frequently observed in the translated images, which is unacceptable for the domain adaptation scenarios, requiring rigorous preservation of image content. Some researchers [10,34] spent efforts to address the problem of content distortions. They employed additional segmentation branches to embed the semantic information to the generators, which enforced the CycleGAN to perform an content-aware image translation. Nevertheless, the obvious drawback of these methods is the demand of pixel-wise annotations.

Inspired by the recent study [3], using a self-supervised loss to retain the benefit of conditional GAN, we explore the potential of self-supervised task for improving CycleGAN's capacity of image content preservation without the demand of pixel-wise annotations. In this paper, we propose an object-preserving I2I domain adaptation network, namely OP-GAN, with the specific capability to address the problem of content distortion occurred in the typical CycleGAN. To be more specific, the newly introduced self-supervised task disentangles the features of image content from the disturbance of domain differences, so as to bring additional regularization for maintaining the consistency of image-objects. The proposed OP-GAN is evaluated on three publicly available datasets. The experimental results show that our OP-GAN can produce satisfactory cross-domain images, while impeccably preserving the image content. The quantitative results demonstrate that the proposed OP-GAN can significantly increase the performance of semantic segmentation networks such as PSPNet [35] and U-Net [24], so as to close the performance gap between different domains.

2 Related Work

In this section, we briefly review previous works on self-supervised learning and unpaired I2I translation.

2.1 Self-supervised learning

To deal with the deficiency of annotated data, researchers attempted to exploit useful information from unlabeled data without direct supervision information.

The self-supervised learning, as a new paradigm of unsupervised learning, attracts increasing attentions from the community. The typical self-supervised learning framework defines a proxy task to enforce neural networks to deeply mine useful information from the unlabeled raw data, which can boost the accuracy of the subsequent target task with limited training data. Various proxy tasks have been proposed, which include grayscale image colorization [15], jigsaw puzzles [23] and object motion estimation [16]. More recently, researchers began to adopt the idea of self-supervised learning to address some key issues of deep learning. For example, Chen et al. [3] introduced an auxiliary self-supervised loss to the typical conditional GAN to address the problem of discriminator forgetting during training. Gidaris et al. [7] integrated the self-supervised learning task to the few-shot learning framework for exploiting richer and more transferable visual representations from few annotated samples.

2.2 Unpaired image-to-image translation

Witnessing the success of cycle-consistency-based approaches [36,13,32], an increasing number of researchers [4,6,22,25] made their effort to the area of unpaired I2I translation. For example, UNIT [21], a recently proposed model, assumes that there exists a shared-latent space in which a pair of corresponding images from different domains could be mapped to the same latent representation. Through such latent representation, the I2I cross-domain translation can be achieved. To further increase the output diversity, Lee et al. [17] proposed a disentangled representation framework, namely DRIT, with unpaired training data. DRIT embedded images into two spaces—a domain-invariant content space capturing shared information across domains, and a domain-specific attribute space to achieve diversity of the translated results. However, none of those approaches explicitly takes the image content preservation into account during translation, which may result in content distortion of the translated images and limit their practicality for the task requiring rigorous preservation of image-objects such as domain adaptation.

3 Revisiting the Problem 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 the inverse B to A $\{G_{BA}, D_A\}$. 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 original data from the translated ones. Thereby, the generators and discriminators are gradually updated during this adversarial competition.

As shown in Fig. 1, 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 A

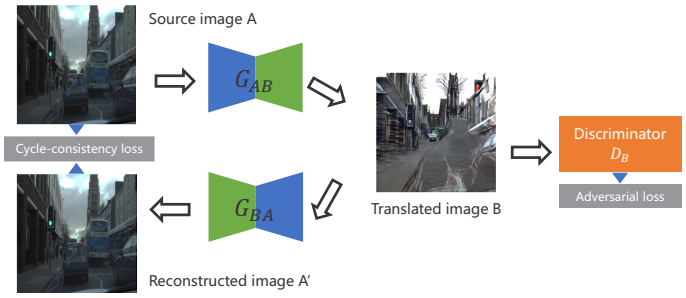


Fig. 1. The overview of CycleGAN [36]. Taking the translation from domain A to domain B as an example, the framework consists of generators (G_{AB} and G_{BA}) and discriminator (D_B), which are supervised by cycle-consistency and adversarial losses, respectively. Due to the bijective geometric transformation (e.g., translation, rotation, scaling, or even nonrigid transformation) and its inverse (i.e., T and T^{-1}), an obvious content distortion is observed in the translated image B, which limits the application of CycleGAN for the tasks required rigorous image-object preservation such as domain adaptation.

to domain B as an example, the adversarial loss can be written as:

$$\begin{aligned}\mathcal{L}_{adv}(G_{AB}, D_B) =& \mathbb{E}_{x_B \sim p_{x_B}}\left[(D_B(x_B) - 1)^2\right] \\ &+ \mathbb{E}_{x_A \sim p_{x_A}}\left[(D_B(G_{AB}(x_A)))^2\right]\end{aligned} \tag{1}$$

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 $\mathcal{L}_{cyc}$ relieves the requirement of 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:

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

With these two losses, CycleGAN can perform I2I translation using unpaired training data. However, recent study [34] found that the cycle-consistency has an intrinsic ambiguity with respect to geometric transformations. Let T be a bijective geometric transformation (e.g., translation, rotation, scaling, or even nonrigid transformation) with inverse transformation T^{-1}, the following generators G'_{AB} and G'_{BA} are also cycle consistent.

$$G'_{AB} = G_{AB}T,\ G'_{BA} = G_{BA}T^{-1}. \tag{3}$$

Consequently, due to lack of penalty in content disparity between source and translated images, the results produced by CycleGAN may suffer from geometrical distortions, as the translated result shown in Fig. 1. To address this problem,

existing studies [34,10] proposed to use a segmentation sub-task with pixel-wise annotation as an auxiliary regularization to assist the training of the generators, which enabled CycleGAN to be applied to tasks such as domain adaptation [34] and data augmentation [10]. However, the expensive and laborious pixel-wise image annotation process limits the practical values of those frameworks.

Motivated by the recent advancements of self-supervised learning, we try to address the content distortion problem of CycleGAN using a novel self-supervised task. The proposed self-supervised task disentangles the content information from the domain variations and accordingly optimizes the generators of CycleGAN without any extra annotations.

4 OP-GAN

In this section, we introduce the proposed OP-GAN in details. Similar to the original CycleGAN, our OP-GAN involves the adversarial and cycle-consistency losses to achieve unpaired I2I translation. In addition, a multi-task self-supervised siamese network (S) is integrated in our OP-GAN, which takes the source and translated images as input, to preserve image content during the I2I image translation.

4.1 Multi-task self-supervised learning

We formulate two self-supervised learning tasks, the content registration and domain classification, to disentangle the features of image content and domain information. We introduce the proposed multi-task self-supervised learning framework in the following section.

Self-supervision formulation. As no preexistent label information is available for the self-supervised siamese network, the supervision is derived from the image data itself. We first divide both the source and translated images into a grid of 3×3.[3] As shown in Fig. 2, letting A and B represent the source and translated images respectively, the generated patches (P) can be written as $P \in \{A_1, ..., A_9\} \cup \{B_1, ..., B_9\}$. There are four scenes if we randomly select two patches from the patch pool, as listed in Table 1. Note that, the $\{A_1, A_5, B_1, B_5\}$ in Fig. 2 are examples for the illustration purpose. During the training stage, the framework randomly selects two patches from the patch pool as the paired input of the siamese network.

Based on the design of the self-supervision, we propose two assumptions to formulate the object-aware domain adaptation: 1) the patches from the same position of source and translated images (C_1) should have consistent content; 2) the patches from the same image (D_1, D_2) should contain similar domain information (e.g., illumination). Accordingly, the relative position of two patches can be used to supervise the proxy task that extracts features with content

[3] The analysis of grid size can be found in *Appendix*.

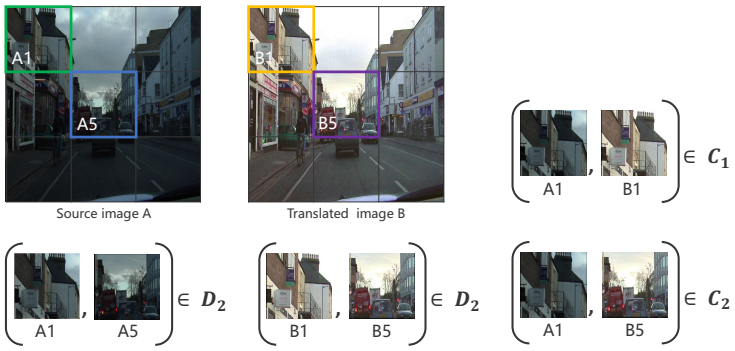


Fig. 2. The supervision signals for the proposed multi-task self-supervised learning framework. The relative position of two patches are used to supervise the content registration, while the domain classification is formulated as a 1-of-K classification task. The A_1, A_5, B_1, B_5 are examples for illustration. The framework randomly selects two patches from the source and translated images as the paired input of the siamese network.

Table 1. Scenes for random patch selection

Scenes
$\mathbf{D_1}$: Two patches are both from the source image.
$\mathbf{D_2}$: Two patches are both from the translated image.
$\mathbf{C_1}$: Two patches are respectively from the source and translated images on the same position of the grid.
$\mathbf{C_2}$: Two patches are respectively from the different positions of source and translated images.

information, while the provenance information of the patches can be used to formulate the proxy task as a domain classification.

Network architecture. The architecture of the proposed siamese network is presented in Fig. 3, which consists of two shared-weight encoders,[4] a content registration branch, and a domain classification branch. The blue, orange, red, and cyan rectangles represent the convolutional, interpolation, global average pooling (GAP), and concatenation layers, respectively.

The shared-weight encoders embed the input patches (P) into a latent feature space (Z) and disentangle the features that contain the content and domain information, respectively. Taking the source (A) and translated (B) images as

[4] The network architecture of the shared-weight encoder is presented in *Appendix*.

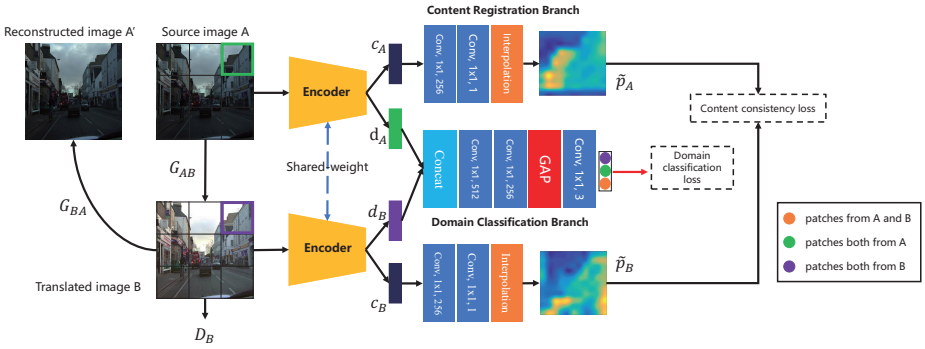


Fig. 3. The architecture of our siamese network. Here, we use the scene with two patches from the same position of two domains as an example. Our siamese network involves three components—two shared-weight encoders, a content registration branch, and a domain classification branch. There are two losses (i.e., content consistency loss and domain classification loss) used for the optimization. The shared-weight encoders embed the two patches into a latent feature space and produce four $11 \times 11 \times 512$ feature maps (c_A, d_A, c_B, d_B) for content registration and domain classification branches.

an example, the feature embedding process can be defined as:

$$E_A : P_A \rightarrow Z(c_A, d_A),\ E_B : P_B \rightarrow Z(c_B, d_B) \tag{4}$$

where c_A and c_B are the disentangled content features; d_A and d_B are the features containing domain information. The size of the four disentangled features is $11 \times 11 \times 512$. Afterwards, c_A and c_B are compacted with 1×1 convolutional layers and interpolated to the original size of the input patch for the computation of content consistency loss, while d_A and d_B are concatenated and fed to the domain classification branch to distill domain information from the features.

Content registration. The content registration branch aims to maintain the patch content during the I2I domain adaptation process. As shown in Fig. 3, the content features are separately processed to produce the content attention maps ($\tilde{p}$), which represent the shape and position of image-objects. As minimum content distortion is a mandatory requirement in our domain adaptation task, the image-objects in source and translated images should be geometrically consistent (i.e., maintaining the shape and position of objects). Hence, we formulate the content consistency loss ($\mathcal{L}_{cc}$) using the two content attention maps ($\tilde{p}$) in L2 norm:

$$\mathcal{L}_{cc} = \frac{1}{M \times N} \sum_{x=1}^{M} \sum_{y=1}^{N} (\tilde{p}^A_{x,y} - \tilde{p}^B_{x,y})^2 \tag{5}$$

where M and N are the width and height of the patch under processing, respectively, and (x, y) is the coordinate of pixel in the attention map.[5]

The content consistency loss produces the pixel-wise penalty for any content disparity between the source and translated images, which enables our OP-GAN to synthesize a realistic result without distortions. The relative position of two input patches is also taken into consideration for the calculation of $\mathcal{L}_{cc}$, i.e., the loss is calculated and optimized only for the scene C_1, and set to 0 otherwise.

Domain classification. As aforementioned, the scenes for random patch selection are adopted as the supervision signals for the domain classification task. It is formulated as a 1-of-K classification, which consists of three classes—D_1, D_2 and C ($C = \{C_1, C_2\}$). The domain features (d_A and d_B) are first fused using concatenation, which results in a $11 \times 11 \times 1024$ discriminative feature map. The feature map is then transformed and downsampled to a $1 \times 1 \times 3$ vector by convolutional and global average pooling layers for the following scene prediction. The cross-entropy loss (denoted as $\mathcal{L}_{dc}$) is adopted in this task for optimization, which can be defined as:

$$\mathcal{L}_{dc} = -\sum_{i} log(\frac{e^{p_{g_i}}}{\sum_j e^{p_j}}) \tag{6}$$

where p_j denotes the j^{th} element ($j \in [1, K]$, K is the number of classes) of vector of class scores, and g_i is the label of i^{th} input sample. The domain classification branch mainly distills domain information from the features, which leads to a better disentanglement of content features.

4.2 Generator and discriminator

In consistent with the standard CycleGAN, the proposed OP-GAN has cyclic generators (G_{AB}, G_{BA}) and corresponding discriminators (D_B, D_A), which have the same architectures as described in [36]. The generators employ the instance normalization [29] to produce elegant image translation results, aiming to fool the discriminators, while the discriminators adopt PatchGAN [12,18] to provide patch-wise predictions of given image being real or fake, rather than classifying the whole image.

4.3 Objective function

Given the definitions of content consistency loss and domain classification loss, we define the self-supervised loss ($\mathcal{L}_S$) as: $\mathcal{L}_S = \mathcal{L}_{cc} + \mathcal{L}_{dc}$.[6] Therefore, the full

[5] After excluding the domain specific information, the content features $\tilde{p}$ from different domains are directly comparable. So, we use the simple mean squared error to measure the difference.

[6] The detailed training process with self-supervised loss can be found in *Appendix*.

objective function of our OP-GAN can be written as:

$$\begin{aligned}\mathcal{L}\left(G_{AB}, G_{BA}, D_A, D_B, S\right) =& \mathcal{L}_{adv}\left(G_{BA},\ D_A\right) + \mathcal{L}_{adv}\left(G_{AB},\ D_B\right) \\ &+ \alpha\mathcal{L}_{cyc}\left(G_{AB},\ G_{BA}\right) + \beta\mathcal{L}_{S}\left(G_{AB}, G_{BA}, S\right)\end{aligned} \tag{7}$$

where α and β are loss weights (we heuristically set $\alpha = 10$ and $\beta = 1$ in our experiments).

The optimization of $\mathcal{L}_S$ is performed in the same manner of $\mathcal{L}_{adv}$—fixing the siamese network (S) and D_A/D_B to optimize G_{BA}/G_{AB} first, and then optimize S and D_A/D_B respectively, with G_{BA}/G_{AB} fixed. Therefore, similar to the discriminators, our siamese network can directly pass the knowledge of image-objects to the generators, which helps to improve the quality of their translated results in terms of object preservation.

5 Experiments

Given two domains (A, B), our goal is to narrow down their gap not only in terms of visual perception i.e., plausible adaptation results, but also feature representations, i.e., improved model robustness. We visualize the I2I domain adaptation results to qualitatively evaluate the former factor. For the latter one, we evaluate the OP-GAN in a similar transfer learning scenario as [27]. Let domain A be in good image condition (e.g., daylight scene with proper exposure), while domain B is unsatisfactory (e.g., image is dark, losing detailed information). In this case, the models trained on domain A usually fail to generalize well to the data from domain B, due to the cross-domain variations. To alleviate the problem, we try different I2I translation frameworks to adapt the domain B data to domain A for testing.

5.1 Experiment settings

Datasets. Experiments are conducted on three publicly available datasets to demonstrate the effectiveness of our OP-GAN.

CamVid [1]: It contains driving videos under different weathers, e.g., cloudy and sunny. The task adapting cloudy videos to sunny ones in terms of illumination and color distribution is very challenging, as the cloudy videos are often very dark, which lose much detailed information. We conduct experiments on the cloudy-to-sunny adaptation to evaluate our OP-GAN.

SYNTHIA [38]: It consists of photo-realistic frames rendered from a virtual city. The night-to-day adaptation is a more difficult task than the cloudy-to-sunny, since the night domain suffers from severe loss of context information. We examine how the proposed OP-GAN performs on the night-to-day adaptation task using two sub-sequences (i.e., winter-day and winter-night) from the Old European Town, which is a subset of SYNTHIA.

Colonoscopic datasets: Medical images from multicentres often have different imaging conditions, e.g., color and illumination, which make the model

trained on one centre difficult to generalize to another. Our OP-GAN tries to address the problem. Two publicly available colonoscopic datasets (i.e., CVC-Clinic [30] and ETIS-Larib [26]) are used as two domains for the multicentre adaptation.

Evaluation criterion. To evaluate the performance of domain adaptation, the mean of class-wise intersection over union (mIoU) [35] is used to evaluate the improvement achieved by our OP-GAN on the semantic segmentation tasks for CamVid and SYNTHIA datasets. For medical image segmentation, the widely-used F1 score [20], which measures the spatial overlap index between the segmentation result and ground truth, is adopted as the metric to assess the accuracy of colorectal ploy segmentation on the colonoscopic datasets.

Baseline overview. Several unpaired I2I domain adaptation frameworks, including CycleGAN [36], UNIT [21], and DRIT [17] are taken as baselines for the performance evaluation. The direct transfer approach, which takes the target domain data for testing without any adaptation, is also involved for comparison. Note that the recent proposed GANs for image-based adaptation, e.g., SPGAN [5], PTGAN [31], and AugGAN [10], are not involved for comparison, due to the strong prior-knowledge used in those approaches. SPGAN, which was proposed for person re-identification task, used the prior-knowledge of the personal ID sets of different domains. PTGAN required coarse segmentation results to distinguish foreground and background areas. AugGAN added a segmentation subtask to the CycleGAN-based framework, which requires 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.

Training details. The proposed OP-GAN is implemented using PyTorch. The generator, discriminator, and siamese network are iteratively trained for 200 epochs with the Adam solver [14]. The baselines involved in this study adopt the same training protocol.

5.2 Visualization of adaptation results

The adaptation results for the three tasks generated by different I2I domain adaptation frameworks are presented in Fig. 4, which illustrate the main problem of existing approaches (UNIT [21], DRIT [17], and CycleGAN [36])—image content corruptions. Due to the lack of penalty in content disparity between the source and translated images, the existing I2I adaptation frameworks intend to excessively edit the image content such as changing the shape and colors of image-objects, referring to the distorted road and building in the CamVid and SYNTHIA translated images. Furthermore, the polyps in colonoscopy images are essential clue for the screening of colorectal cancer. However, few of the existing frameworks successfully maintain the shape and texture of polyps during

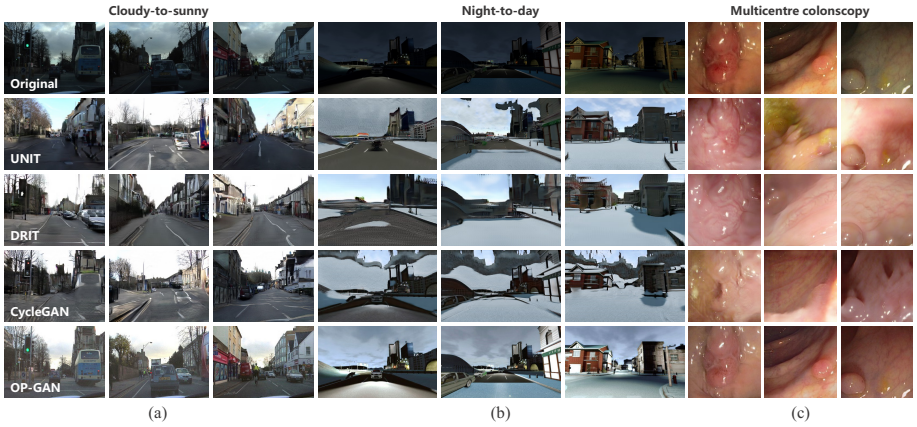


Fig. 4. Comparison of translated images with different approaches. The image-based adaptation performance is evaluated on three tasks: (a) cloudy-to-sunny, (b) night-to-day, and (c) multicentres. The original images, adaptation results produced by UNIT [21], DRIT [17], CycleGAN [36], and our OP-GAN are presented.

Table 2. Semantic segmentation IoU (%) of the cloudy images from CamVid with different I2I domain adaptation frameworks. (val.—validation)

	Bicyclist	Building	Car	Pole	Fence	Pedestrian	Road	Sidewalk	Sign	Sky	Tree	mIoU
Sunny (val.)	84.03	86.30	90.91	18.36	74.91	63.09	94.07	89.75	7.49	94.00	91.48	70.38
Cloudy (test)												
Direct transfer	48.70	40.29	51.15	21.59	4.95	**43.84**	**71.64**	**60.69**	**21.29**	40.59	63.90	41.86
UNIT [21]	0.67	49.28	6.99	6.10	1.54	0.79	45.05	15.16	0.00	62.61	39.34	19.94
DRIT [17]	0.34	40.00	0.31	0.33	0.83	0.23	48.27	26.81	0.00	64.85	23.73	18.20
CycleGAN [36]	6.16	56.64	10.76	8.61	0.01	4.06	50.49	30.21	8.67	75.97	45.62	26.26
OP-GAN (Ours)	**51.28**	**73.10**	**74.19**	**25.84**	**12.42**	42.75	70.48	51.74	14.71	**81.09**	**72.40**	**51.40**

multicentre adaptation, which is unacceptable and limits their practical values in medical-related applications. On the contrary, the proposed OP-GAN can excellently perform cross-domain adaptation, while preserving the image-objects.

5.3 Cloudy-to-sunny adaptation on CamVid

The CamVid dataset contains four sunny videos (577 frames in total) and one cloudy video (124 frames). Each frame of the videos is manually annotated, which associate each pixel with one of the 32 semantic classes. Based on the widely-accepted protocol [37], we focus on 11 classes including bicyclist, building, car, pole, fence, pedestrian, road, sidewalk, sign, sky, and tree. To evaluate the domain adaptation performance yielded by our OP-GAN, a semantic segmentation

Table 3. Semantic segmentation IoU (%) of the night images from SYNTHIA with different I2I domain adaptation frameworks. (Veg.—Vegetation, Ped.—Pedestrian, l.-m.—lane-marking, val.—validation)

	Sky	Building	Road	Sidewalk	Fence	Veg.	Pole	Car	Sign	Ped.	Bicycle	l.-m.	mIoU
Day (val.)	94.84	93.71	96.49	92.09	27.99	50.06	54.04	91.35	46.31	69.21	54.46	77.18	70.02
Night (test)													
Direct transfer	0.04	61.45	70.41	78.22	**11.35**	37.76	34.67	80.88	26.62	53.51	54.33	38.42	44.49
UNIT [21]	**63.52**	**72.69**	65.80	47.44	0.93	48.76	28.11	37.54	9.71	28.10	26.76	9.87	34.97
DRIT [17]	33.63	43.96	50.35	13.10	0.03	17.92	0.41	2.69	0.06	0.38	0.03	0.62	12.66
CycleGAN [36]	37.73	51.79	55.48	36.37	0.71	44.23	14.60	17.94	1.57	8.19	10.67	11.89	22.88
OP-GAN (Ours)	21.90	66.22	**86.78**	**79.05**	7.70	**54.86**	**39.11**	**85.09**	**31.40**	**55.77**	**54.54**	**47.61**	**50.86**

network (PSPNet[7] [35]) is trained with the sunny frames and tested on the original cloudy frames and the translated ones. In the experiment, the sunny frames are separated into training (three videos) and validation (one video) sets. The evaluation results are shown in Table 2.

Due to the loss of detailed information, it can be observed from Table 2 that the performance of PSPNet trained with sunny images dramatically drops to 41.86% while tested on the original cloudy images. As the existing I2I domain adaptation approaches encounter the content distortion problem, the segmentation mIoU of PSPNet further degrades to 26.26%, 19.94% and 18.20% using the CycleGAN, UNIT and DRIT, respectively. In contrast, the proposed OP-GAN achieves a significant improvement (+9.54%) compared to the direct transfer, which demonstrates that our OP-GAN can narrow down the gap between the cloudy and sunny domains while excellently preserving the image-objects. The proposed OP-GAN significantly boosts the IoU of some object-related classes such as building, car, and fence (i.e., +32.81%, +23.04%, and +7.47%, respectively). Specifically, AugGAN [10], using pixel-wise annotations for image-object preservation, achieves a mIoU of 55.31% in our experiment, which can be regarded as the upper bound for our approach.

5.4 Night-to-day adaptation on SYNTHIA

We adopt two sub-sequences (winter-day and winter-night) from SYNTHIA to perform night-to-day adaptation. The winter-day and winter-night sequences contain 947 and 785 frames, respectively. SYNTHIA dataset provides pixel-wise semantic annotations for each frame, which can be categoried to 13 classes (12 semantic classes and background). The partition of training, validation, and test sets complies with the same protocol to that of the CamVid dataset—the day images are separated into training and validation sets according to the ratio of 70:30, while all the night images are used as the test set. The fully convolutional network (PSPNet [35]) is also adopted in this experiment to perform semantic segmentation.

[7] The top-1 solution (without extra training data) on the leaderboard of semantic segmentation on CamVid: https://paperswithcode.com/sota/semantic-segmentation-on-camvid.

Table 4. Ablation study of OP-GAN for the semantic segmentation task on CamVid (%). (D. T.—Direct Transfer)

	-	A	B	C	D
Setup	D. T.	Original CycleGAN	A + L_{cc}	A + L_{dc}	A + L_{cc} + L_{dc}
mIoU	41.86	26.26	45.63	45.86	51.40

The segmentation mIoUs yielded by different testing strategies are shown in Table 3. Similar to the cloudy images, the PSPNet trained with day images fails to propoerly process the night images—an mIoU of 44.49%, due to the loss of information. The night-to-day adaptation on SYNTHIA is a more challenging task compared to the cloudy-to-sunny adaptation on CamVid, since a large portion of the night images is dark, where the image-objects (e.g., buildings) are difficult to recognize. Refer to Fig. 4, existing I2I domain adaptation frameworks used to create extra contents to fill the extremely dark areas, which consequently corrupts the original image-objects. Due to these distortions, the images translated by UNIT, CycleGAN and DRIT further decrease the mIoU of PSPNet to 34.97%, 22.88% and 12.66%, respectively. Our OP-GAN can excellently prevent image-object corruptions during night-to-day adaptation (as shown in Fig. 4) and achieves the best mIoU (50.86%) for the night images, which is +6.37% higher than the direct transfer.

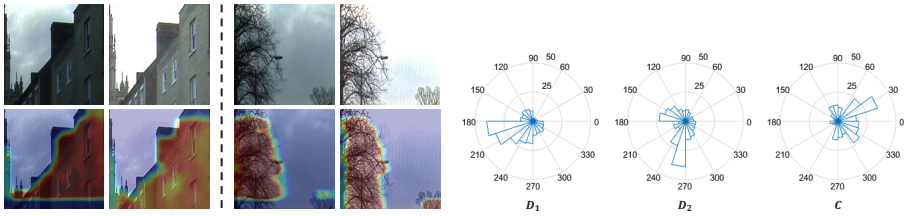
5.5 Multicentre colonoscopy adaptation

Due to the limitation of paper length, the experimental results on multicentre colonoscopy adaptation are presented in *Appendix*.

5.6 Ablation study

An ablation study is conducted on the cloudy-to-sunny adaptation task on CamVid to evaluate the contribution produced by each component of our OP-GAN. The result of ablation study is presented in Table 4. Due to the capacity of feature distillation, the content registration and domain classification branches can respectively improve the mIoU of original CycleGAN with +19.37% and +19.6%. The combination of these two branches lead to a better disentanglement of content and domain information, which results in the highest improvement (+25.14%). To validate the effectiveness of the proposed self-supervised learning tasks, we visualize the knowledge learned by different branches and analyze their contributions for image-object preservation.

Content registration. The content registration branch aims to maintain the shape and texture of image-objects before and after domain adaptation. We visualize two pairs of attention maps ($\tilde{p}$) generated by content registration branches to validate whether they have the 'object'-related concept. The attention maps are presented in Fig. 5 (a), which shows that the content registration branch



(a) (b)

Fig. 5. Validation of self-supervised proxy tasks. (a) Visualization of attention maps produced by content registration branch. For an object-preserving I2I domain adaptation, the patches on the same position of source and translated images should have similar content attention maps ($\tilde{p}$). (b) The activation patterns of different classes (D_1, D_2, and C) in domain classification.

prefers to activate the areas containing image-objects (e.g., buildings and trees) and ignore those containing more domain information such as sky. As a result, this branch penalizes the generator if the translated image-objects have large distortions, which encourages the OP-GAN to perform object-aware translation.

Domain classification. To ensure the scene classification is a learnable proxy task, we plot the $1 \times 1 \times 256$ averaged activation patterns for different classes (D_1, D_2, C) produced by the global average pooling layer of domain classification branch in Fig. 5 (b). It can be observed that different neurons are activated when processing paired patches from different scenes, which demonstrates that the scenes defined in Table 1 indeed contain specific domain information for the classifier to distinguish each other.

6 Conclusion

In this paper, we proposed a novel GAN (namely OP-GAN) to perform object-preserving image-to-image domain adaptation without supervision from manual labels. Extensive experiments have been conducted on three publicly available datasets. The experimental results demonstrated the effectiveness of our OP-GAN—performing excellent cross-domain translation while preserving image-objects.

Acknowledge

This work is supported by the Natural Science Foundation of China (No. 91959108 and 61702339), the Key Area Research and Development Program of Guangdong Province, China (No. 2018B010111001), National Key Research and Development Project (2018YFC2000702) and Science and Technology Program of Shenzhen, China (No. ZDSYS201802021814180).

References

1. Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R.: Segmentation and recognition using structure from motion point clouds. In: ECCV (2008)
2. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2019)
3. Chen, T., Zhai, X., Ritter, M., Lucic, M., Houlsby, N.: Self-supervised GANs via auxiliary rotation loss. In: CVPR (2019)
4. Chen, Y., Lai, Y.K., Liu, Y.J.: Cartoongan: Generative adversarial networks for photo cartoonization. In: CVPR (2018)
5. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: CVPR (2018)
6. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Zhang, K., Tao, D.: Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In: CVPR (2019)
7. Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P., Cord, M.: Boosting few-shot visual learning with self-supervision. In: ICCV (2019)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
9. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
10. Huang, S., Lin, C., Chen, S., Wu, Y., Hsu, P., Lai, S.: AugGAN: Cross domain adaptation with GAN-based data augmentation. In: ECCV (2018)
11. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: CCNet: Criss-cross attention for semantic segmentation. In: ICCV (2019)
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
13. Kim, T., Cha, M., Kim, H., Lee, J., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: ICML (2017)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Larsson, G., Maire, M., Shakhnarovich, G.: Colorization as a proxy task for visual understanding. In: CVPR (2017)
16. Lee, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Unsupervised representation learning by sorting sequences. In: ICCV (2017)
17. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M.K., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: ECCV (2018)
18. Li, C., Wand, M.: Precomputed real-time texture synthesis with Markovian generative adversarial networks. In: ECCV (2016)
19. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: CVPR (2019)
20. Li, Y., Xie, X., Liu, S., Li, X., Shen, L.: GT-Net: A deep learning network for gastric tumor diagnosis. In: ICTAI (2018)
21. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: NeurIPS (2017)
22. Ma, S., Fu, J., Wen Chen, C., Mei, T.: DA-GAN: Instance-level image translation by deep attention generative adversarial networks. In: CVPR (2018)
23. Noroozi, M., Vinjimoor, A., Favaro, P., Pirsiavash, H.: Boosting self-supervised learning via knowledge transfer. In: CVPR (2018)
24. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing & Computer Assisted Intervention (2015)

25. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: CVPR (2017)
26. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: 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 (2014)
27. Sun, L., Wang, K., Yang, K., Xiang, K.: See clearer at night: Towards robust nighttime semantic segmentation through day-night image conversion. arXiv preprint arXiv:1908.05868 (2019)
28. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, Inception-ResNet and the impact of residual connections on learning. In: AAAI (2017)
29. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
30. Vázquez, D., Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., López, A.M., Romero, A., Drozdzal, M., Courville, A.: A benchmark for endoluminal scene segmentation of colonoscopy images. Journal of healthcare engineering **2017** (2017)
31. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer GAN to bridge domain gap for person re-identification. In: CVPR (2018)
32. Yi, Z., Zhang, H., Tan, P., Gong, M.: DualGAN: Unsupervised dual learning for image-to-image translation. In: ICCV (2017)
33. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Learning a discriminative feature network for semantic segmentation. In: CVPR (2018)
34. Zhang, Z., Yang, L., Zheng, Y.: Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network. In: CVPR (2018)
35. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)
36. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)
37. Zhu, Y., Sapra, K., Reda, F.A., Shih, K.J., Newsam, S., Tao, A., Catanzaro, B.: Improving semantic segmentation via video propagation and label relaxation. In: CVPR (2019)
38. Zolfaghari Bengar, J., Gonzalez-Garcia, A., Villalonga, G., Raducanu, B., Aghdam, H.H., Mozerov, M., Lopez, A.M., van de Weijer, J.: Temporal coherence for active learning in videos. arXiv preprint arXiv:1908.11757 (2019)