7 Appendix

In this section, we present more detailed information and experimental results of our OP-GAN.

Results on multicentre colonoscopy adaptation. In this experiment, the extremely small ETIS-Larib dataset (196 frames) is used as the test set, while the relatively larger CVC-Clinic dataset (612 frames) is used for network optimization (80:20—training and validation). For medical image segmentation, the U-shape network may be a more appropriate choice, compared to the PSP-Net. Therefore, we adopt ResUNet-50 [9,25] to perform polyp segmentation for the evaluation of multicentre adaptation. Table 5 lists the F1 scores of polyp segmentation using original and translated ETIS-Larib images. The proposed OP-GAN remarkably boosts the accuracy of polyp segmentation, i.e., +8.28% to the direct transfer. In addition, the performance of statistical approach, i.e., histogram equalization, is also evaluated for comparison, i.e., an F1 score of 63.11% is achieved. Since the approach performs imaging condition alignment only using the statistical information, it avoids the problem of content distortion, i.e., +1.91% higher than direct transfer.

Table 5. F1 score (%) of polyp segmentation on the CVC validation (val.) set and ETIS test set

	CVC (val.)	**ETIS** (test)
Direct transfer	79.22	61.20
UNIT [22]		21.96
DRIT [18]		19.97
CycleGAN [37]		45.25
OP-GAN (ours)		**69.48**

Analysis on the grid size. The source and translated images are respectively separated into a set of patches for the shared-weight encoders of our self-supervised framework to extract features. To analyze the influence generated by grid sizes, we compare the performance of OP-GAN with different grids on the CamVid dataset. The evaluation result is shown in Table 6. It can be observed that the 3×3 grid is more appropriate for our self-supervised framework, which yields the highest mIoU (51.40%).

Architecture of shared-weight encoders. The architecture of shared-weight encoders adopted in our OP-GAN is shown in Table 7. The input is two 171×171 patches (P_A, P_B) and output is four $11 \times 11 \times 512$ features (c_A, c_B, d_A, d_B).

Training procedure with self-supervisions. The detailed process of training OP-GAN with self-supervised signals is presented in Alg. 1.

Table 6. Analysis of grid size on the CamVid dataset.

	Bicyclist	Building	Car	Pole	Fence	Pedestrian	Road	Sidewalk	Sign	Sky	Tree	mIoU
Sunny (validation)												
PSPNet	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)												
2×2	49.87	69.32	66.14	22.78	14.43	35.29	70.62	51.06	**15.68**	67.68	67.49	47.94
3×3	**51.28**	**73.10**	**74.19**	**25.84**	12.42	**42.75**	70.48	51.74	14.71	**81.09**	72.40	**51.40**
4×4	50.88	70.91	61.72	23.07	**16.13**	36.06	**70.66**	49.46	13.95	79.22	**74.28**	49.26
5×5	44.14	68.08	65.77	22.54	15.72	32.11	70.25	**54.33**	14.99	59.69	63.12	46.39

Table 7. The encoder architecture. The Conv and L-ReLU denote the convolutional and Leaky ReLU layers, respectively. The Layer Info contains the parameters of convolutional layers (number of channel, kernel size, padding, stride). The input patch size is 171×171.

Layers	Encoder	Layer Info	Output size
1	Conv, L-ReLU	(64, 3, 1, 2)	86×86
2	Conv, L-ReLU	(128, 3, 1, 2)	43×43
3	Conv, L-ReLU	(256, 3, 1, 2)	22×22
4	Conv, L-ReLU	(512, 3, 1, 2)	11×11

Semantic segmentation results. The semantic segmentation results of original and translated images yielded by different frameworks on three tasks are shown in the Fig. 6 (cloudy-to-sunny adaptation), Fig. 7 (night-to-day adaptation), and Fig. 8 (multicentre colonscopy adaptation), respectively.

Algorithm 1 Self-supervised training strategy

Input:
- P_i^t and P_j^t: two patches randomly selected from the patch pool $(\{A_1, ..., A_9\} \cup \{B_1, ..., B_9\})$ at iteration t

Supervision signal:
- Content registration: four scenes for random patch selection $\{D_1,\, D_2,\, C_1,\, C_2\}$
- Domain classification: $\{D_1,\, D_2,\, C\}$, where $\{C_1, C_2\}$ are categoried to a single class (C)

Output:
- $\tilde{p}_i$ and $\tilde{p}_j$: attention maps generated by the content registration branch
- p_{dc}: prediction for domain classification
- $\mathcal{L}_S$: total loss of S

Functions:
- $F(P_i^t, P_j^t)$ {forward function of S}
- $CL(.)$ {loss calculation (i.e., $\mathcal{L} \in \{\mathcal{L}_{cc}, \mathcal{L}_{dc}\}$)}
- $B(.)$ {backward function for the calculated loss}

Initialize:

$t \leftarrow 0$

$\mathcal{L}_S \leftarrow 0$

repeat

    $\{\tilde{p}_i, \tilde{p}_j, p_{dc}\} \leftarrow F(P_i^t, P_j^t)$

    $\mathcal{L}_{dc} \leftarrow CL(p_{dc}, \{D_1 : 0, D_2 : 1, C : 2\})$

    if C_1 **then**

        $\mathcal{L}_{cc} \leftarrow CL(\tilde{p}_i, \tilde{p}_j)$

        $\mathcal{L}_S \mathrel{+}= (\mathcal{L}_{dc} + \mathcal{L}_{cc})$

    else

        $\mathcal{L}_S \mathrel{+}= \mathcal{L}_{dc}$

    end if

    $B(\mathcal{L}_S)$

    $t \leftarrow t + 1$

until iteration (t) meets the pre-set number

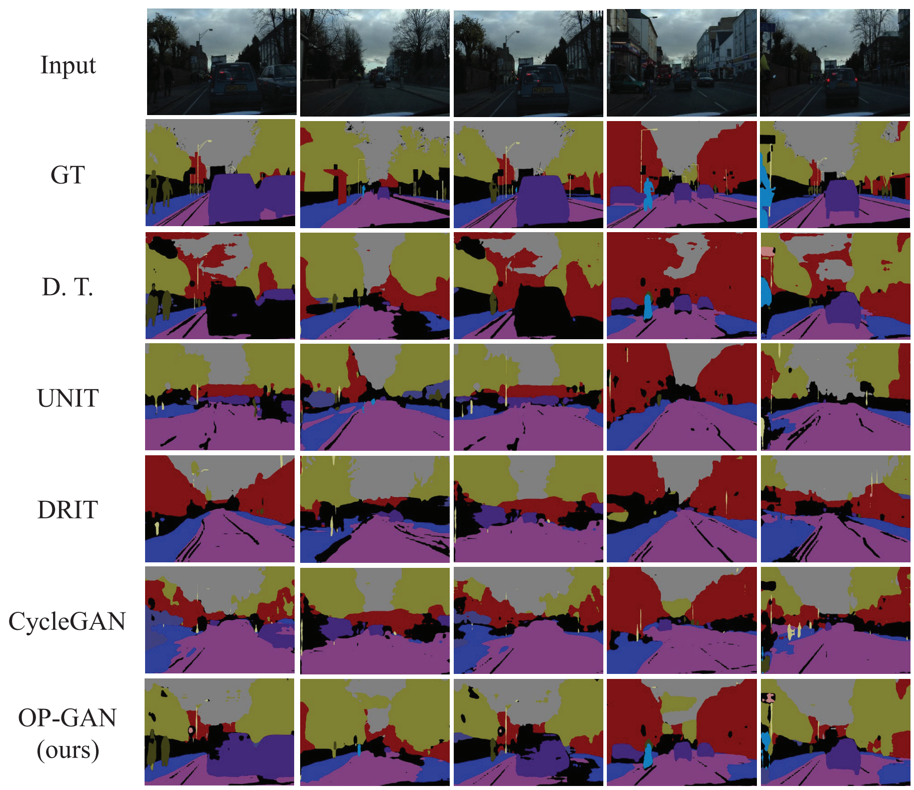

Fig. 6. Semantic segmentation results produced by the sunny-image-trained PSPNet for the original CamVid cloudy images and the ones translated by UNIT [22], DRIT [18], CycleGAN [37], and our OP-GAN. D. T. refers to direct transfer.

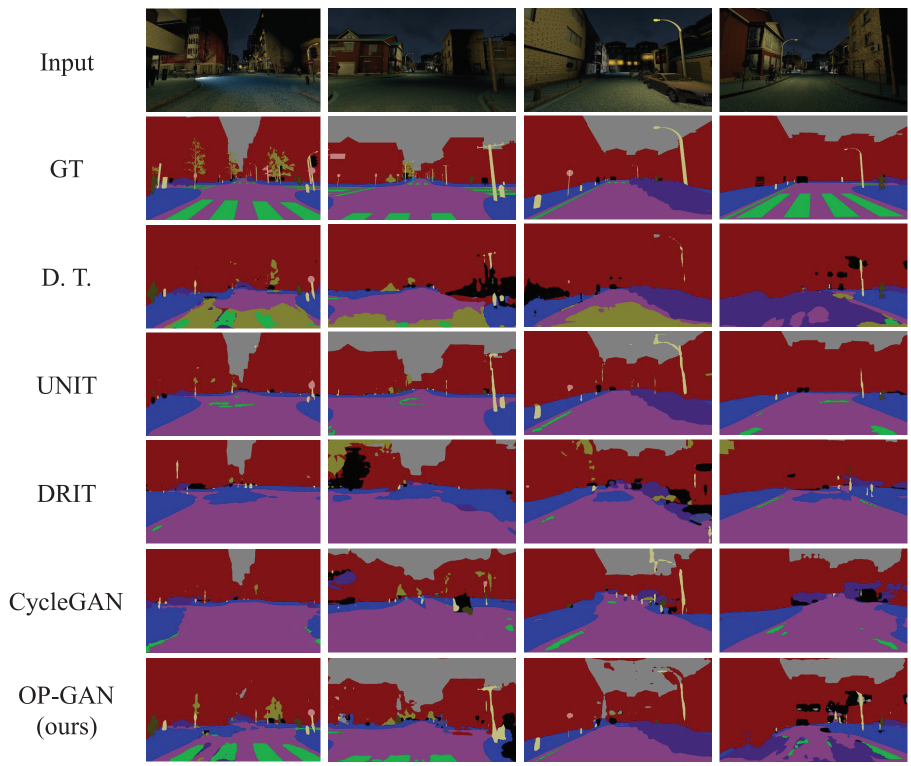

Fig. 7. Semantic segmentation results produced by the day-image-trained PSPNet for the original SYNTHIA night images and the ones translated by UNIT [22], DRIT [18], CycleGAN [37], and our OP-GAN. D. T. refers to direct transfer.

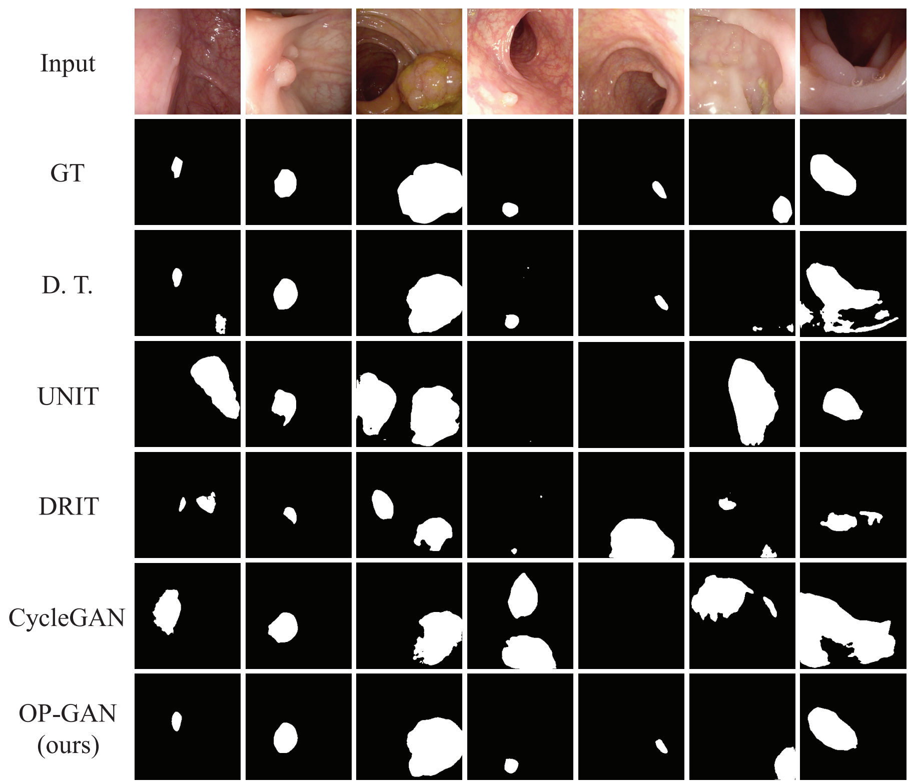

Fig. 8. Polyp segmentation results produced by the CVC-trained ResUNet for the original ETIS images and the ones translated by UNIT [22], DRIT [18], CycleGAN [37], and our OP-GAN. D. T. refers to direct transfer.