2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-7358-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISBI53787.2023.10230716

UNPAIRED IMAGE-TO-IMAGE TRANSLATION BASED DOMAIN ADAPTATION FOR POLYP SEGMENTATION

Xinyu Xiong[1], *Siying Li*[2], *Guanbin Li*[1,3*]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China
[3]Research Institute, Sun Yat-sen University, Shenzhen, China

ABSTRACT

Deep polyp segmentation methods have made tremendous progress recently. However, due to the domain shift among different imaging modalities, existing methods learned on white-light imaging (WLI) achieve inferior results on other modalities such as narrow-band imaging (NBI), which limits their clinical usage. To tackle this problem, we propose a Polyp Style Translation Network (PST-Net). Specifically, test images from the NBI domain are translated by PST-Net to have the style and features of WLI images. In this way, the already deployed segmentation model can be easily generalized to images from the unseen NBI domain, without the need for tedious re-training and re-labeling. Besides, three additional designs, content consistency, attention map consistency, and adversarial segmentation loss, are proposed to achieve better translation as well as domain adaptation. Extensive experiments demonstrate that PST-Net achieves state-of-the-art performance.

***Index Terms*—** Polyp Segmentation, Generative Adversarial Network, Domain Adaptation, Image-to-Image Translation

1. INTRODUCTION

According to Global Cancer Statistics [1], colorectal cancer ranks third in both incidence and mortality, posing a serious threat to human health. Fortunately, localization and segmentation of polyps can help diagnose colorectal cancer at an early stage. With the development of deep learning, recent automatic polyp segmentation methods [2, 3, 4] have achieved impressive performance, showing their great potential in constructing better computer-aided diagnosis systems.

Despite the remarkable success, there are still some under-discussed problems preventing the broader application of these segmentation methods, one of which is domain shift. As shown in Fig. 1(a) and 1(b), white-light imaging (WLI) and narrow-band imaging (NBI) are two imaging modalities widely used in clinical practice. However, most existing models are trained and evaluated on WLI datasets, such as Kvasir [5] and ClinicDB [6]. Due to the appearance difference and underlying feature distribution mismatch between WLI and NBI, a model well-trained on the source WLI domain has a higher risk of failure on the target NBI domain, as illustrated in Fig. 1(c).

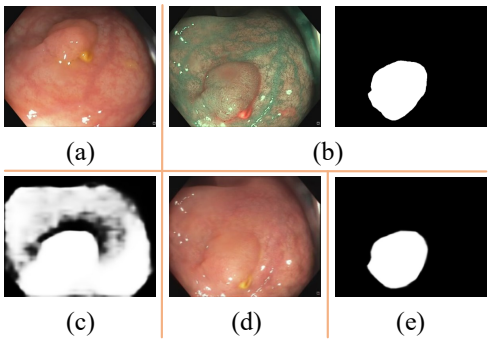

Fig. 1. (a) A sample from the WLI domain. (b) A sample from the NBI domain and its corresponding ground truth mask. (c) Segmentation result of the NBI image before translation. (d) Our translated image. (e) Segmentation result after translation.

Unpaired image-to-image translation (I2I), which aims to translate images from one style to another style, can be employed as a domain adaptation technique to alleviate domain shift. Hoffman *et al.* [7] develop a semantic consistency loss to guide the image translator to preserve the structure and content of the original image. Murez *et al.* [8] leverage a translation classification loss to constrain the target encoder to be trained with supervision on images similar to target domain ones. An adversarial contrastive training strategy is proposed in [9] to jointly analyze both style and content of a sample.

Recently, some attempts have emerged to deal with the multi-centre problem [9] in polyp segmentation, while how to address the domain shift caused by imaging modality is still under-explored. To tackle this problem, we propose a Polyp Style Translation Network (PST-Net) to enhance the generalizability of existing segmentation models over differ-

* Corresponding Author.

ent imaging modalities. Specifically, NBI images for testing are directly translated to have the style and features of WLI images (Fig. 1(d)). In such a manner, the performance of the already deployed segmentation model is boosted (Fig. 1(e)), which is meaningful in clinical practice since it is laborious to re-label target domain images and re-training a new segmentation model. In addition, we propose three additional designs, where content consistency is used to encourage content retention during translation, attention map consistency facilitates the network to learn the multi-level discrepancies between NBI and WLI, and adversarial segmentation loss makes the translated images better adapted to the segmentation network. In summary, our contributions are three-fold:

- By constructing an unpaired image-to-image translation network, the generalizability of the existing polyp segmentation model is boosted in a simple and effective way.
- Three additional designs are proposed to achieve better translation as well as domain adaptation.
- Compared to existing state-of-the-art methods, extensive experiments demonstrate the superiority of PST-Net.

2. METHOD

2.1. Problem Definition

Define WLI as the source domain and NBI as the target domain. f_s is a polyp segmentation network pre-trained on the source domain. Given the source domain images $x^s \in \mathcal{X}^s$ and target domain images $x^t \in \mathcal{X}^t$. p_s and p_t are the distributions of $\mathcal{X}^s$ and $\mathcal{X}^t$. Our goal is to develop a generative adversarial network that translates images from the target domain to the source domain. Then, the translated images $x^{t\to s}$ are input into f_s to obtain better segmentation results.

2.2. Framework Overview

The architecture of our method is shown in Fig. 2, which is CycleGAN-like [10]. Specifically, PST-Net consists of two generators, $\mathcal{G}_s$, $\mathcal{G}_t$, and three discriminators, $\mathcal{D}_s$, $\mathcal{D}_t$, $\mathcal{D}_a$. Next, we will describe how to integrate network modules with loss functions to build a basic translation network.

Adversarial Translation Loss. From the target domain to the source domain, the generator $\mathcal{G}_s$ needs to make the generated images $x^{t\to s} = \mathcal{G}_s(x^t)$ realistic to deceive the discriminator, while the discriminator $\mathcal{D}_s$ should learn to distinguish the real source images x^s from the fake ones $x^{t\to s}$. Such process of adversarial learning can be formalized as follows:

$$\begin{aligned}\mathcal{L}_{\text{ADT}}(\mathcal{G}_s, \mathcal{D}_s) = \mathbb{E}_{x^s\sim p_s}[\log \mathcal{D}_s(x^s)] \\ + \mathbb{E}_{x^t\sim p_t}[\log(1 - \mathcal{D}_s(\mathcal{G}_s(x^t)))],\end{aligned} \tag{1}$$

where $\mathcal{D}_s$ is expected to maximize the full objective and $\mathcal{G}_s$ tries to minimize $\log(1 - \mathcal{D}_s(\mathcal{G}_s(x^t)))$. Similarly, a symmetric loss function $\mathcal{L}_{\text{ADT}}(\mathcal{G}_t, \mathcal{D}_t)$ can be obtained for translation from the source domain to the target domain. In summary, adversarial translation loss is defined as follows:

$$\mathcal{L}_{\text{ADT}} = \mathcal{L}_{\text{ADT}}(\mathcal{G}_s, \mathcal{D}_s) + \mathcal{L}_{\text{ADT}}(\mathcal{G}_t, \mathcal{D}_t). \tag{2}$$

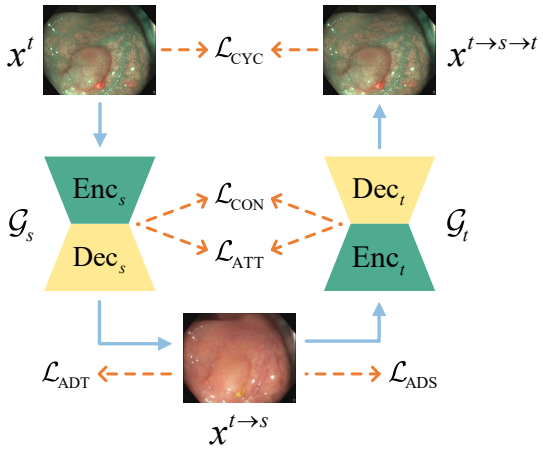

Fig. 2. The architecture of our proposed PST-Net. For ease of illustration, the reverse translation process and discriminators are omitted in the figure. $\mathcal{G}_s$/$\mathcal{G}_t$ should generate plausible source/target domain images, while $\mathcal{D}_s$/$\mathcal{D}_t$ should distinguish whether an image is really from the source/target domain. $\mathcal{D}_a$ is an auxiliary discriminator for adversarial segmentation loss.

Cycle Consistency Loss. Adversarial translation loss only encourages the translated images to be realistic at the style level. To realize basic content level retention, following [10], we obtain $x^{t\to s\to t} = \mathcal{G}_t(x^{t\to s})$ and $x^{s\to t\to s} = \mathcal{G}_s(x^{s\to t})$ by cycle-translate $x^{t\to s}$ and $x^{s\to t}$ back to their original domain. Then, cycle consistency loss is given by:

$$\begin{aligned}\mathcal{L}_{\text{CYC}} = \mathbb{E}_{x^t\sim p_t}[\|\mathcal{G}_t(\mathcal{G}_s(x^t)) - x^t\|_1] \\ + \mathbb{E}_{x^s\sim p_s}[\|\mathcal{G}_s(\mathcal{G}_t(x^s)) - x^s\|_1].\end{aligned} \tag{3}$$

Identity Loss. By promoting generators to perform identity mapping for input images from the same domain, identity loss can be defined as:

$$\mathcal{L}_{\text{IDT}} = \mathbb{E}_{x^t\sim p_t}[\|\mathcal{G}_t(x^t) - x^t\|_1] + \mathbb{E}_{x^s\sim p_s}[\|\mathcal{G}_s(x^s) - x^s\|_1]. \tag{4}$$

As denoted in [10], introducing identity loss can improve the quality of translated images and stabilize the training process.

Summary. The above design constitutes a basic network that can perform image translation. Next, we will discuss how it can be refined to better work with domain adaptation.

2.3. Content Consistency

With the introduction of identity loss, generators are required to reconstruct the input images themselves. From this perspective, the architecture of generators can be considered the

auto-encoder, consisting of an encoder and a decoder. The encoder encodes images into latent content codes, and the decoder is responsible for decoding content codes back into the correct domain.

Simply equipping cycle consistency does not guarantee the translated images retain their original semantics since the decoder may be able to reconstruct meaningless noisy content codes back to the original images when overfitting. To tackle this problem, we enable better semantic information retention by encouraging latent content codes of translated images $x^{t\to s} = \mathcal{G}_s(x^t)$ and $x^{s\to t} = \mathcal{G}_t(x^s)$ to be consistent with ones from input images x^t and x^s, which can be solved by:

$$\begin{aligned}\mathcal{L}_{\text{CON}} = \mathbb{E}_{x^t\sim p_t}[\|Enc_s(x^t) - Enc_t(\mathcal{G}_s(x^t))\|_1] \\ + \mathbb{E}_{x^s\sim p_s}[\|Enc_t(x^s) - Enc_s(\mathcal{G}_t(x^s))\|_1],\end{aligned} \tag{5}$$

where Enc_s denotes the encoder of $\mathcal{G}_s$, and the rest can be obtained similarly.

2.4. Attention Map Consistency

Aside from the most evident color inconsistencies, the differences between NBI and WLI are multifaceted. For example, blood vessels are more easily identified in NBI than in WLI. Therefore, the network should deal with discrepancies at different levels to achieve reasonable translation. Unfortunately, generators can take a shortcut by simply modifying the color to deceive discriminators, leading to sub-optimal translation results.

Attention maps are widely used in computer vision tasks to explain the decision process of deep networks. Inspired by this, we leverage GradCAM [11] to extract attention maps from generators. Specifically, attention maps should be consistent throughout the translation process, which can be formalized as follows:

$$\begin{aligned}\mathcal{L}_{\text{ATT}} = \mathbb{E}_{x^t\sim p_t}[\|\mathbf{A}^{\mathcal{G}_s}(x^t) - \mathbf{A}^{\mathcal{G}_t}(\mathcal{G}_s(x^t))\|_1] \\ + \mathbb{E}_{x^s\sim p_s}[\|\mathbf{A}^{\mathcal{G}_t}(x^s) - \mathbf{A}^{\mathcal{G}_s}(G_t(x^s))\|_1],\end{aligned} \tag{6}$$

where $\mathbf{A}^{\mathcal{G}_s}(x^t)$ denotes attention maps drawn from $\mathcal{G}_s$ when x^t are input into $\mathcal{G}_s$, and the rest can be obtained similarly. The intuition behind attention map consistency is that if generators simply modify the global color information to achieve translation, attention maps will become more uncontrollable and difficult to be consistent. Instead, if adaptations other than color exist, attentions are focused on the modified areas and are more likely to remain consistent.

2.5. Adversarial Segmentation Loss

For input images x^s that are from the source domain, the segmentation network f_s should output reliable segmentation results $r^s = f_s(x^s)$. Recall that the ultimate goal of PST-Net is to generate realistic images $x^{t\to s} = \mathcal{G}^s(x^t)$, which f_s can also easily deal with and output satisfactory results $r^{t\to s} = f_s(\mathcal{G}^s(x^t))$. Therefore, a critical problem is whether $r^{t\to s}$ are helpful for the segmentation task.

Assessing the quality of $r^{t\to s}$ without ground truth masks is challenging; one advisable way is to evaluate whether they have features similar to r^s, such as low uncertainty and reasonable lesion area. To this end, we propose an auxiliary discriminator $\mathcal{D}_a$ to distinguish whether segmentation results are obtained from real images x^s or generated ones $x^{t\to s}$. In this way, $\mathcal{G}_s$ is encouraged to generate samples performing well in the downstream segmentation task. Such process of adversarial learning can be solved by adversarial segmentation loss:

$$\begin{aligned}\mathcal{L}_{\text{ADS}} = \mathbb{E}_{x^s\sim p_s}[\log\mathcal{D}_a(f_s(x^s))] \\ + \mathbb{E}_{x^t\sim p_t}[\log(1 - \mathcal{D}_a(f_s(\mathcal{G}_s(x^t))))].\end{aligned} \tag{7}$$

2.6. Full Objective

The total loss of PST-Net is as follows:

$$\begin{aligned}\mathcal{L}_{\text{TOTAL}} = \lambda_{\text{ADT}}\mathcal{L}_{\text{ADT}} + \lambda_{\text{CYC}}\mathcal{L}_{\text{CYC}} + \lambda_{\text{IDT}}\mathcal{L}_{\text{IDT}} \\ + \lambda_{\text{CON}}\mathcal{L}_{\text{CON}} + \lambda_{\text{ATT}}\mathcal{L}_{\text{ATT}} + \lambda_{\text{ADS}}\mathcal{L}_{\text{ADS}}.\end{aligned} \tag{8}$$

In our experiments, we set $\lambda_{\text{ADT}} = 1$, $\lambda_{\text{CYC}} = 10$, $\lambda_{\text{IDT}} = 5$, $\lambda_{\text{CON}} = 1$, $\lambda_{\text{ATT}} = 1$, $\lambda_{\text{ADS}} = 1$.

3. EXPERIMENTS

3.1. Dataset

In this paper, we adopt a recently proposed large-scale multi-modal polyp segmentation dataset named PICCOLO [12], which contains 2,131 WLI images and 1,302 NBI images. For dataset partitioning, 80% of the samples in each modality are randomly selected for training, and the remaining 20% serve as the test set.

3.2. Implementation Details

Our model is implemented with Pytorch and trained on a single NVIDIA RTX 2080Ti. We employ the Adam optimizer with a learning rate of 0.0002 for 100 epochs. The batch size is set to 1. All the images are resized to 256 × 256 for translation. For the polyp segmentation network f_s, we adopt PraNet [3] trained on the WLI domain. The generators refer to the design in [13], and discriminators use the architecture of PatchGAN [14].

3.3. Compared Methods

Our method is compared with the following state-of-the-art methods: CycleGAN [10], MUNIT [15], CUT [16], NEGCUT [17], and GP-UNIT [18]. Eight metrics are used to evaluate the segmentation performance quantitatively, including "Recall", "Specificity", "Precision", "Dice Score", "IoU

Table 1. Quantitative comparison on polyp segmentation. The best results are shown in bold.

Methods	Rec	Spec	Prec	Dice	IoUp	IoUb	mIoU	Acc
w/o translation	82.82	88.32	58.66	61.34	49.10	84.29	66.69	86.95
CycleGAN	82.20	92.19	67.83	68.98	58.84	88.74	73.79	90.66
MUNIT	81.30	91.68	66.85	66.49	56.60	87.00	71.80	88.81
CUT	77.49	94.88	71.71	68.41	58.92	90.12	74.52	91.59
NEGCUT	79.23	93.66	68.07	68.35	59.06	89.59	74.32	91.21
GP-UNIT	78.72	84.22	54.64	55.47	44.16	78.83	61.49	81.39
Ours	**83.12**	**96.76**	**78.19**	**76.85**	**68.67**	**92.60**	**80.64**	**93.79**

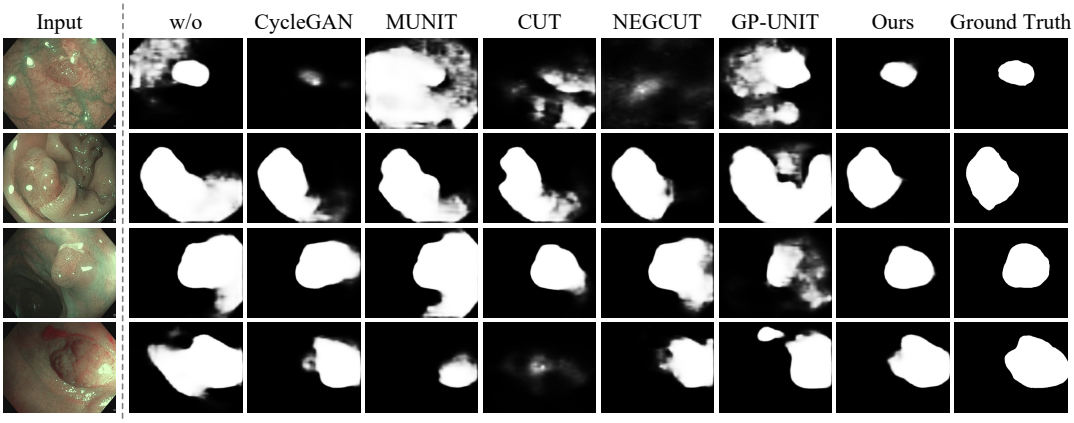

Fig. 3. Qualitative comparison on polyp segmentation. "w/o" is the abbreviation for "without translation".

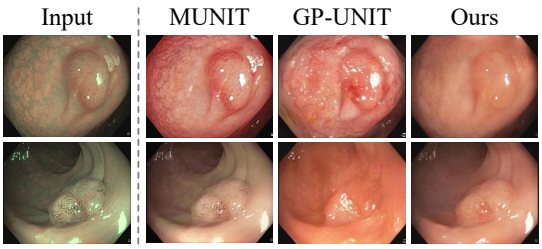

Fig. 4. Qualitative comparison on image-to-image translation.

for Polyp (IoUp)", "IoU for Background (IoUb)", "Mean IoU (mIoU)" and "Accuracy".

3.4. Result Analysis

As shown in Table 1, our proposed PST-Net achieves superior performance over other competitive methods in all metrics. Take IoUp as an example. When no image translation is performed, the resulting IoUp is only 49.10%, verifying the domain shift between WLI and NBI. Therefore, it is suboptimal to directly apply the model trained on the WLI domain to NBI images. After applying our proposed PST-Net, the IoUp is boosted to 68.67%, surpassing the second-best method (NEGCUT) by a large margin. In addition, it can be observed that not all methods result in performance gains. The results of GP-UNIT are even worse than the case without translation, denoting the importance of maintaining semantic information in the translation process.

Some qualitative segmentation results are shown in Fig. 3. Our method alleviates domain shift and helps the segmentation model better deal with various lesions. Some translation results are shown in Fig. 4. Compared with other methods, PST-Net can realistically translate NBI images into WLI images, thus enabling better domain adaptation.

We also conduct additional ablation experiments. Specifically, removing "Content Consistency", "Attention Map Consistency", and "Adversarial Segmentation Loss" causes IoUp to drop by 2.75%, 2.53%, and 3.39%, respectively, demonstrating the contribution of proposed strategies.

4. CONCLUSION

In this paper, we propose an image-to-image translation network for domain adaptation. Our method can be regarded as a means of data preprocessing to improve the generalizability of existing polyp segmentation models effectively. In future work, we will verify whether the same gains can be obtained in more imaging modalities.

5. REFERENCES

[1] Rebecca L. Siegel, Kimberly D. Miller, Hannah E. Fuchs, and Ahmedin Jemal, "Cancer statistics, 2022," *CA: A Cancer Journal for Clinicians*, vol. 72, no. 1, pp. 7–33, 2022.

[2] Ruifei Zhang, Guanbin Li, Zhen Li, Shuguang Cui, Dahong Qian, and Yizhou Yu, "Adaptive context selection for polyp segmentation," in *MICCAI*, 2020, pp. 253–262.

[3] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in *MICCAI*, 2020, pp. 263–273.

[4] Jun Wei, Yiwen Hu, Ruimao Zhang, Zhen Li, S Kevin Zhou, and Shuguang Cui, "Shallow attention network for polyp segmentation," in *MICCAI*, 2021, pp. 699–708.

[5] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen, "Kvasir-seg: A segmented polyp dataset," in *MMM*, 2020, pp. 451–462.

[6] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized medical imaging and graphics*, vol. 43, pp. 99–111, 2015.

[7] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell, "Cycada: Cycle consistent adversarial domain adaptation," in *ICML*, 2018.

[8] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim, "Image to image translation for domain adaptation," in *CVPR*, 2018, pp. 4500–4509.

[9] Ta Duc Huy, Hoang Cao Huyen, Chanh DT Nguyen, Soan TM Duong, Trung Bui, and Steven QH Truong, "Adversarial contrastive fourier domain adaptation for polyp segmentation," in *ISBI*, 2022, pp. 1–5.

[10] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2223–2232.

[11] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.

[12] Luisa F Sánchez-Peralta, J Blas Pagador, Artzai Picón, Ángel José Calderón, Francisco Polo, Nagore Andraka, Roberto Bilbao, Ben Glover, Cristina L Saratxaga, and Francisco M Sánchez-Margallo, "Piccolo white-light and narrow-band imaging colonoscopic dataset: a performance comparative of models and datasets," *Applied Sciences*, vol. 10, no. 23, pp. 8501, 2020.

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016, pp. 694–711.

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.

[15] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz, "Multimodal unsupervised image-to-image translation," in *ECCV*, 2018, pp. 172–189.

[16] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu, "Contrastive learning for unpaired image-to-image translation," in *ECCV*, 2020, pp. 319–345.

[17] Weilun Wang, Wengang Zhou, Jianmin Bao, Dong Chen, and Houqiang Li, "Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation," in *ICCV*, 2021, pp. 14020–14029.

[18] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy, "Unsupervised image-to-image translation with generative prior," in *CVPR*, 2022, pp. 18332–18341.

Acknowledgments

This work was supported in part by the Shenzhen Science and Technology Program (NO. JCYJ20220530141211024), in part by the Guangdong Basic and Applied Basic Research Foundation (NO. 2020B1515020048), in part by the National Natural Science Foundation of China (NO. 61976250), in part by the Fundamental Research Funds for the Central Universities under Grant 22lgqb25.

Compliance with Ethical Standards

This is a numerical simulation study for which no ethical approval was required.