Self-Supervised Dense Consistency Regularization for Image-to-Image Translation

Supplementary Document

Minsu Ko[1*] Eunju Cha[1*] Sungjoo Suh[1] Huijin Lee[1] Jae-Joon Han[1]
Jinwoo Shin[2] Bohyung Han[3]
[1]Samsung Advanced Institute of Technology (SAIT), South Korea
[2]Korea Advanced Institute of Science and Technology (KAIST), South Korea
[3]Seoul National University (SNU), South Korea

A. Implementation Details

A.1. Augmentation

DCR uses the almost same set of image augmentation methods as SimCLR [1]. Specifically, it performs color distortion, which consists of a random sequence of brightness, contrast, saturation, and hue adjustments. After that, grayscale conversion is considered. Note that DCR does not employ the horizontal flipping because it is not compatible to the dense consistency regularization. We crop random patches in an image and resize them to the half the target image. We make crop ratios larger than 0.7, which guarantees overlapping regions to be non-empty.

A.2. Pseudo-code of DCR

Algorithm 1 presents the pseudo-code of DCR.

Algorithm 1 DCR Pseudo-code, PyTorch-like

```
# R: DCR module
# D = D1 · D0 : discriminator
# G : generator
# crit : GAN loss
# sim_nc : Negative cosine similarity loss of
    corresponding feature
# weight : hyperparameter, which is set to 1

for x in loader: # load a minibatch x with n samples

   ## Discriminator update
   pred_real = D(x)
   pred_fake = D(G(x).detach())
   loss_real = crit(pred_real, True)
   loss_fake = crit(pred_fake, False)
   L_dist = loss_real/2 + loss_fake/2

   x1, x2 = aug(x), aug(x) # random augmentation
   z1, z2 = D0(x1), D0(x2) # Representation, n*C*H*W
   p1, p2 = R(z1), R(z2) # Predictions, n*C*H*W

   L_DCR = sim_nc(p1, z2)/2 + sim_nc(p2, z1)/2

   L = L_dist + weight * L_DCR

   L.backward() # back-propagate
   update(f, h) # SGD update
```

B. Additional Qualitative Results

We provide additional qualitative image-to-image translation results on various datasets, including the Horse $\rightarrow$ Zebra, Winter $\rightarrow$ Summer, and Cat $\rightarrow$ Dog datasets in Figure 1 to compare the quality of translated images using the baseline methods and those using the methods with DCR. The fifth row in Figure 1 shows that DRIT++ [8] and FSeSim [12] fail to generate the properly translated images from the given cat image, where the overall shapes and the detailed appearance near eyes and mouths do not look natural. However, the integration of DCR into the models helps the baseline models generate images with significantly improved quality in terms of shape and texture. Therefore, we believe that DCR opens up a new generation of regularization techniques for image-to-image translation.

We compare qualitative results with multi-modal image-to-image translation models in Figure 2. The translation results from the wild to the dog domain on the MUNIT [5] model provide clear differences. The proposed method generates the translated image by rendering the geometry of the source domain and the local context of the target domain more successfully than the baseline. For StarGANv2 [3] models, the qualities of generated images are very similar and only differ in small details.

C. Additional Ablation study

Comparison with CR-GAN We present the benefit of consitency regularization using the comparative experiments between CR-GAN [10] and DCR. Since DCR performs dense regularization, it is well-suited for image generation tasks that require pixel-level high-fidelity predictions by fo-

*These authors contributed equally.

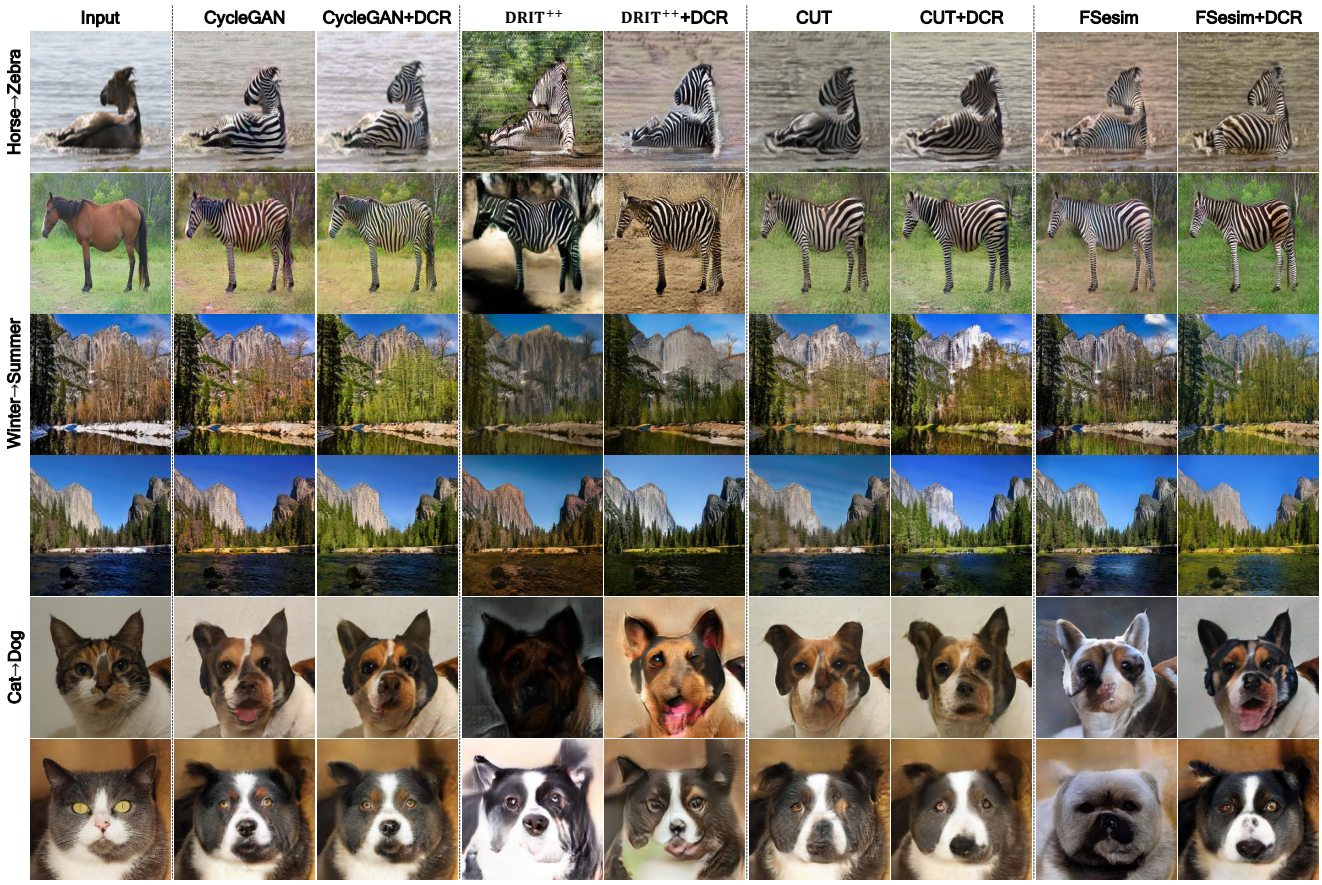

Figure 1. Qualitative comparison of image-to-image translation results on the Horse → Zebra, Winter → Summer, and Cat → Dog datasets. CycleGAN [13], CUT [9], DRIT++ [8], and FSeSim [12] are employed as the baseline models and the proposed DCR is integrated into those models.

cusing on specific features in images. Therefore, as shown in the fourth column and the eighth column in Table 1, CR-GAN [10] enhances performance compared to the baseline but DCR achieves a further improvement.

Benefit of stop-gradient Stop-gradient [2] and momentum update [4] are the standard methods to prevent the collapse and degeneracy of trained models in self-supervised contrastive learning. Refer to the results of DCR with and without the stop-gradient in Table 1 to see its benefit. Without the stop-gradient, the performance is consistently dropped compared to the baseline models. Since the DCR utilizes no negative pairs, the discriminator can not be free from the mode collapse, which leads to the degraded performance.

What to regularize Fake images may have unnatural contents and style, and the application of CR to fake images may lead to unexpected artifacts in generated images. Table 1 shows that DCR with both real and fake images degrades performance. Note that, although [11] adopts CR also to generated images, it is also done in latent space, which is not feasible in DCR involving image cropping. Table 1 also shows that applying DCR to whole images enhances the performance but still worse than DCR, which is partly because more diversity in contrastive learning is helpful. Therefore, we apply DCR to the cropped regions of the real images to enable the discriminator to learn better representations for the image-to-image translation.

D. DCR with other regularization

With the purpose of semantic consistency and visual harmony in the spatial domain, DCR is applied to the intermediate layer of the discriminator. Thanks to the flexibility of DCR, there is no constraint with the use of DCR on existing GAN models. Furthermore, DCR is one of the regularization techniques to provide better performance of image-to-image translation and image generation. This implies that DCR can provide more performance gains to existing GAN models when used with various regularization techniques such as data augmentation and contraD [6].

StyleGAN2-ADA [7] proposes an adaptive data augmentation to stabilize training in limited data regimes. To

	Metric	Baseline	CR-GAN	DCR w/o stop-grad.	DCR with fake/real img.	DCR to whole img.	Full DCR
CUT [9]	FID ↓	43.2±2.3	42.3±2.9	44.3±0.9	37.9±1.3	36.6±1.9	**34.0**±0.4
	D&C (D) ↑	0.73±0.06	0.66±0.02	0.62±0.03	0.72±0.16	0.79±0.05	**0.96**±0.10
	D&C (C) ↑	0.87±0.02	0.65±0.02	0.79± 0.01	0.85±0.07	0.89±0.01	**0.90**±0.00
FSeSim [12]	FID ↓	45.2±4.8	40.9±1.0	58.2±1.8	45.4±2.1	43.6±1.9	**36.7**±1.4
	D&C (D) ↑	0.75±0.14	0.52±0.10	0.52±0.13	0.70±0.05	0.75±0.08	**0.89**±0.02
	D&C (C) ↑	0.83±0.04	0.71±0.03	0.66± 0.11	0.86±0.01	0.89±0.02	**0.91**±0.02

Table 1. Ablation studies to further analyze the propoesd DCR. Quantitative comparison of the best FID scores and D&C on CUT [9] and FSeSim [12] of Horse → Zebra dataset. Standard deviations are calculated from two runs.

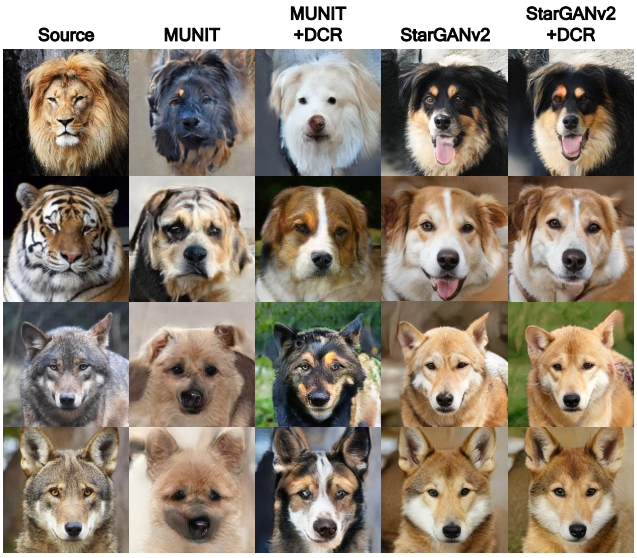

Figure 2. Qualitative comparison of image-to-image translation results on the AFHQ dataset. Since AFHQ dataset is introduced for the multi-domain image-to-image translation, we employ MUNIT [5] and StarGANv2 [3] as the baseline models. The structure of the original image is well preserved with the desired styles, when the proposed DCR is integrated into the base algorithms.

investigate the benefits of applying DCR to StyleGAN2-ADA [7], experiments are performed on the AFHQ Dog dataset. We train the model with $1M$ images in total for computational efficiency. Table 2 shows that DCR can improve the performance of StyleGAN2 when the fusion with the adaptive data augmentation.

Method	StyleGAN2-ADA [7]	StyleGAN2-ADA + DCR
FID ↓	31.55 ± 0.15	29.69 ± 0.12

Table 2. Quantitative results on recent image generation model using the AFHQ Dog dataset. The results are from 2 trials.

E. Discussion

E.1. Potential negative societal impacts

Self-supervised learning often require a large number of computations due to the constastive learing methods (e.g., SimCLR [2] requires 128TPUs). Such enormous computational cost may cause environmental problem such as global warming and air pollution. Hence various supervision (e.g., image generation) and data efficient training would be required which we expect that DCR could contribute in this field. Thanks to the low computational cost and data efficiency of the proposed DCR.

E.2. Limitation

We observed that dense and instance level regularization technique can boost the performance each others, but the role of each method is not clear. As reported in Table 4 in main paper, the representation close to the pixel tends to preserve low-level semantics such as texture and the higher representation is effective to preserve semantic information. We believe that the relative characteristics of representation can be further investigated, which remains as a future work.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[2] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2, 3

[3] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. 1, 3

[4] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 9729–9738, 2020. 2

[5] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 1, 3

[6] Jongheon Jeong and Jinwoo Shin. Training GANs with stronger augmentations via contrastive discriminator. In *International Conference on Learning Representations*, 2021. 2

[7] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020. 2, 3

[8] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 128(10):2402–2417, 2020. 1, 2

[9] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020. 2, 3

[10] Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. In *International Conference on Learning Representations*, 2020. 1, 2

[11] Zhengli Zhao, Sameer Singh, Honglak Lee, Zizhao Zhang, Augustus Odena, and Han Zhang. Improved consistency regularization for gans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11033–11041, 2021. 2

[12] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. The spatially-correlative loss for various image translation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16407–16417, 2021. 1, 2, 3

[13] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2