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

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

Abstract

Unsupervised image-to-image translation has gained considerable attention due to recent impressive advances in generative adversarial networks (GANs). This paper presents a simple but effective regularization technique for improving GAN-based image-to-image translation. To generate images with realistic local semantics and structures, we propose an auxiliary self-supervision loss that enforces point-wise consistency of the overlapping region between a pair of patches cropped from a single real image during training the discriminator of a GAN. Our experiment shows that the proposed dense consistency regularization improves performance substantially on various image-to-image translation scenarios. It also leads to extra performance gains through the combination with instance-level regularization methods. Furthermore, we verify that the proposed model captures domain-specific characteristics more effectively with only a small fraction of training data.

1. Introduction

Generative adversarial network (GAN) [8] is an innovative framework for generative modeling, *i.e.*, generating images that follow the same distribution as training data. The performance of the state-of-the-art GAN models depends highly on the quality of discriminators, which distinguish real images from fake ones while maintaining the balance with matching generators for the joint optimization. Since discriminators are prone to overfit the training dataset and often lead to the mode collapse of generated outputs, learning robust discriminators is critical to accomplish high-performance generators.

To this end, self-supervised learning methods have been actively used for regularizing discriminators in the GAN framework [15, 16, 32, 34]. The goal of the regularization is to obtain robust representations of images for better discrimination of real and fake images [17]. The existing methods often rely on contrastive learning in an instance-level [3, 15, 16, 32], where a pair of augmented instances from an image are encouraged to have consistent features with respect to predefined global transforms while negative images are optionally considered to achieve better representation learning in discriminators. However, the regularization based only on such global representations may be limited to imposing loose constraints on discriminators and may allow generators to deceive the discriminator despite local structural or semantic inconsistency in output images.

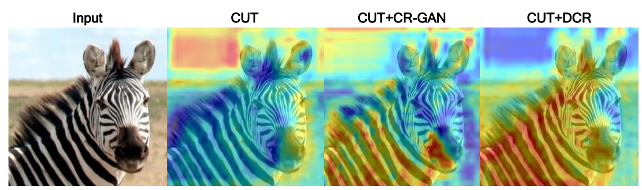

Figure 1. Qualitative comparison of output feature maps given by three different variation of the baseline mode, CUT [26], for an example in the Horse → Zebra dataset. The activations from DCR highlight foreground accurately while suppressing background effectively, which helps the image-to-image translation task focused more on target objects.

To alleviate the drawback, we propose a dense consistency regularization (DCR) approach applicable to the discriminator of a GAN. DCR provides stronger constraints to the learned representations given by discriminators through their point-wise consistency between a pair of patches cropped from the same image. Our work is motivated by the hypothesis that image generation requires pixel-level prediction [14] and a dense regularization of representations is an effective way to improve the supervision quality of a discriminator. The goal of the proposed dense consistency regularization is to generate images with both semantic con-

*These authors contributed equally.

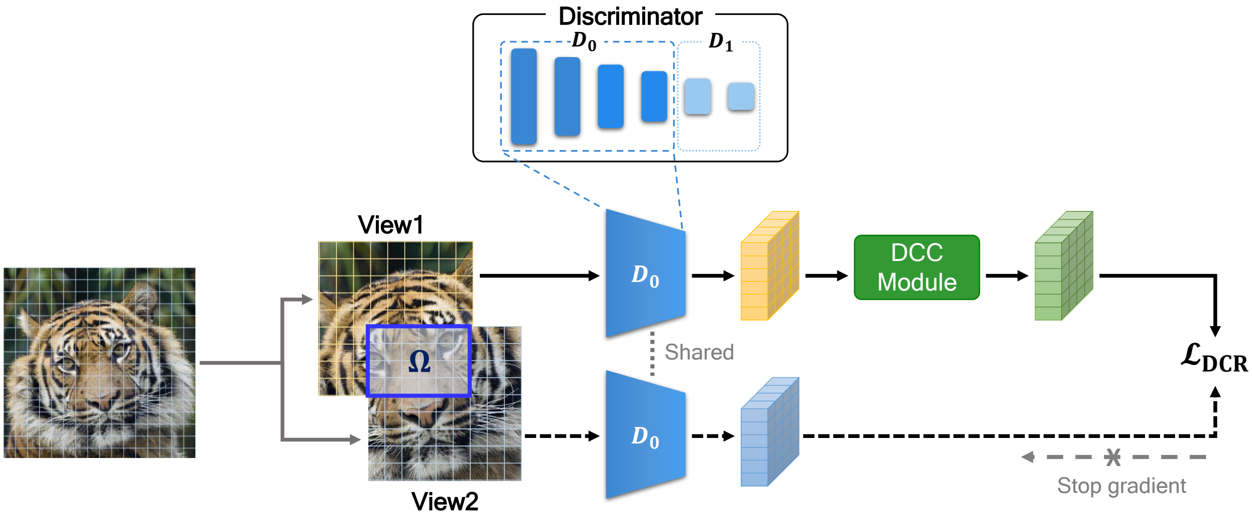

Figure 2. Illustration of the proposed DCR method. When updating a discriminator, two augmented views are randomly cropped from a single real image. The two views are then processed by the intermediate feature extraction network D_0, where D_0 is the first part of he discriminator while the remaining part is denoted by D_1. Note that D_1 is not used in our work. The DCR module is applied to one of the branches, and a stop-gradient operation is employed in the other one. The loss $\mathcal{L}_{\text{DCR}}$ is given by the similarity between the representations of two branches over the overlapping region Ω while $\tilde{\Omega}$ denotes the binary map indicating matching pairs of pixels.

sistency and visual harmony in spatial neighborhoods. This is achieved if the discriminator focuses on important features or regions for image-to-image translation instead of the background, as shown in Figure 1. Our main idea is illustrated in Figure 2, where the dense correspondence regularization is imposed on the intermediate layers of the discriminator.

We evaluate the proposed approach on various image-to-image translation scenarios such as CycleGAN [37], MUNIT [13], StarGANv2 [5], CUT [26], and FSeSim [36]. According to our experiments on the Horse $\rightarrow$ Zebra, Winter $\rightarrow$ Summer, Cat $\rightarrow$ Dog, and AFHQ datasets, the models with DCR consistently improve the FID scores compared to the models without DCR, which confirms that DCR indeed captures domain-specific characteristics effectively. For example, we manage to improve the FID score of CycleGAN [37] from 78.2 to 54.4, and that of MUNIT [13] from 102.3 to 59.9 on the Horse $\rightarrow$ Zebra dataset. Moreover, we also find out that DCR is particularly powerful with a small number of training data. Specifically, StarGANv2 [5] with DCR achieves the best FID score of 17.15 even if only 10% of a specific domain in the AFHQ dataset is used for training, while the best FID scores of StarGANv2 [5] are 22.63 and 17.86 with 10% and 100% of the examples in AFHQ.

We summarize our contributions as follows:

- We introduce a novel dense consistency regularization technique, referred to as DCR, for the discriminators of GANs, which facilitates high-fidelity image generation and translation.
- We show that DCR is effective to maintain structural and semantic consistency in the spatial neighborhoods of generated images.
- We empirically demonstrate that DCR achieves outstanding performance in various image-to-image translation scenarios.

In the rest of this paper, we first discuss closely related works to our approach in Section 2, and present our algorithm and implementation details in Section 3. Section 4 demonstrates the results from our experiments with their analysis, and Section 5 concludes this paper.

2. Related Work

This section reviews existing regularization methods imposed on the discriminator of GANs and presents generic dense representation learning techniques applicable to discriminator regularization. We also discuss existing approaches in image-to-image translation, which is the primary target task of the proposed regularizer.

2.1. Regularization for Discriminator

GAN [8] is a well-known generative model particularly effective for image generation and translation tasks. The generator is trained to produce realistic images deceiving the discriminator while the discriminator learns to distinguish between fake images obtained from the generator and real ones sampled from training data. The great advance in network architectures of GANs capacitates the generation of more realistic images, but GANs still suffer from inherent stability issues in training, especially high sensitivity to hyperparameters originated from the non-convexity of the min-max objective function.

The issue has been address in various studies, which include integration of the normalization method [24] or regularization via gradient penalization [10, 19, 29]. The regularization for discriminators turns out to stabilize training and improve performance [15, 16, 32, 34, 35]. We hypothesize that the main reason for the improvement is good representation in the discriminator side, which is crucial to distinguish real images from fake ones and eventually increases the quality of the generator. In particular, [34] introduces a simple consistency regularization (CR) to discriminators, and obtains substantially enhanced quality of generated images with reduced computational cost compared to gradient-based regularization techniques [10, 19, 29].

To learn more informative representations by optimizing discriminators, self-supervised learning approaches have been employed [3, 15, 16, 32]. For instance, [3] incorporates the auxiliary rotation loss for self-supervision, by which both the real and generated images are classified into one of the relevant rotation angles, $\{0°, 90°, 180°, 270°\}$. ContraD [15] distinguishes two independent real images as well as a pair of real and fake images to mitigate the overfitting problem in discriminators while learning more powerful representations in them. The promising results from these works imply that the learned representations in a discriminator play a crucial role in improving the performance of GANs in practice.

2.2. Dense Representation Learning

Image-level prediction tasks, *e.g.*, image classification, often achieve significant performance improvement by incorporating self-supervised visual representation learning via instance discrimination, which maximizes the similarity between two augmented images while optionally decreasing the similarity between different ones [2, 4, 9, 11, 33]. However, such instance-level contrastive learning methods may be suboptimal as a pretext task due to the lack of their spatial sensitivity.

To address the limitation, dense self-supervised learning approaches have been proposed, where they take into account the pixel-level similarity between two augmented images [27, 28, 30, 31]. View-Agnostic Dense Representation (VADeR) [27] adopts the pretext task that pulls the features at the overlapping locations of two different views cropped from a single image, while making the features from the non-overlapping regions apart from each other. However, VADeR relies heavily on a large number of negative pairs and consequently incurs high computational cost. On the other hand, [31] proposes a pixel-to-propagation consistency (PixPro) regularization without negative pairs, which encourages the pixels in the spatial neighborhood to have similar representations. Spatially Consistent Representation Learning (SCRL) [28] enforces the consistency of the features corresponding to the same objects that are identified in two different views of a single image.

2.3. Image-to-Image Translation

Unpaired image-to-image translation techniques are divided into two categories depending on the use of cycle consistency loss, which facilitates learning an inverse mapping from the target domain to the source domain [18, 37]. The loss is defined in either the image domain [5, 18, 21, 37] or the latent space [13, 22, 23] to preserve the key attributes between input and output images. Although the learned mapping with the cycle consistency is reliable enough to provide high-quality outputs in image-to-image translation, the translated images may contain too much information of the input images for its effective reconstruction, resulting in undesirable outputs.

To address this issue, distance-based loss [1] and geometric consistency [7] have been adopted for the translation from the source to the target without using its inverse mapping. The contrastive learning framework is adopted to preserve the content of the original image in the translated image [26], where a patch-wise contrastive loss is proposed to maintain the correspondence between the source and target images. A structure consistency loss is employed to enforce the self-similarity between the source and target images [36]. CUT [26] and F/LSeSim [36] deliver better results than the algorithms with the cycle consistency loss by taking advantage of cross-domain similarity functions. However, these algorithms focus only on the comparison between the original and translated images with no consideration about the representations in their discriminators. Since we believe that the discriminator representation reflecting the target domain distribution accurately is a crucial component to generate high-fidelity images in image-to-image translation, we propose a dense consistency regularization strategy via self-supervised learning.

3. Dense Consistency Regularization (DCR)

This section presents our main algorithm, especially the technical details of dense consistency regularization module. We also discuss several implementation issues of the proposed approach.

3.1. Motivation

The role of the discriminator in GANs is to distinguish real data from fake ones created by the generator and provide the generator with the proper feedback for producing realistic images. Contrary to discriminative tasks such as image classification, image generation requires pixel-level predictions in its output. Hence, the discriminator should be able to capture the local context of an output image for the high-fidelity to the target domain in the image-to-image translation task. Spatial sensitivity in the representation learning has been introduced by [31], which measures the consistency of spatially overlapping pixels for more discriminative learning around object boundaries. To make the discriminator have spatial sensitivity, we design a task for the local feature similarity measure and discuss its details in the rest of this section. To obtain mid-level local features, we decompose a discriminator D into two subnetworks denoted by D_0 and D_1 such that

$$D = D_1 \circ D_0. \tag{1}$$

As in most visual representation learning, we start by sampling two augmented views x_1 and x_2 from an image x. The two views are resized to a fixed resolution (*e.g.*, 128×128) and passed through the shared feature extractor D_0.

To verify our hypothesis, we visualize the output feature map of the discriminator of CUT [26] in Figure 1, where DCR focuses on the foreground area more effectively than vanilla CUT [26] and CUT with CR [34] while suppressing the activations in the background region. This result implies that DCR is helpful to improve the quality of generated images, especially around object boundaries.

3.2. DCR Module

DCR is motivated by SimSiam [4], which utilizes only positive pairs for contrastive learning and employs the stop-gradient technique to prevent collapse to the trivial solution. Note that, since image generation task needs to learn the distribution of a target domain, sampling negative examples from the target domain dataset is not straightforward. One can introduce an additional dataset to obtain negative examples, but the selection of the negative dataset is tricky because it requires sophisticated and comprehensive supervision to check various attributes of the dataset.

The proposed DCR module, denoted by $R(\cdot)$, consists of two 1×1 convolutional layers and a LeakyReLU activation between the convolutions. The output feature map size of the DCR module is identical to that of its input (*e.g.*, $W \times H \times C$) to maintain the spatial information. Suppose that we have the intermediate representations of two augmented images as $r_1 := R(D_0(x_1))$ and $z_2 := D_0(x_2)$. Given the overlapping region Ω of two views x_1 and x_2, we define the negative cosine similarity of their corresponding features, which is given by

$$\text{sim}_{\text{nc}}(r_1, z_2; \tilde{\Omega}) \equiv \sum_{\{(i,j)|\tilde{\Omega}(i,j)=1\}} -\frac{r_1[i]}{\|r_1[i]\|_2} \cdot \frac{z_2[j]}{\|z_2[j]\|_2}, \tag{2}$$

where $\tilde{\Omega}$ is the binary map representing feature correspondences, $[\cdot]$ is used to specify the index corresponding to a particular location in a feature map, and $\|\cdot\|_2$ indicates ℓ_2-norm. Following [4], the DCR loss is given by

$$\mathcal{L}_{\text{DCR}} = \frac{1}{2}\text{sim}_{\text{nc}}(r_1, F_{\text{sg}}(z_2)) + \frac{1}{2}\text{sim}_{\text{nc}}(r_2, F_{\text{sg}}(z_1)), \tag{3}$$

where $F_{\text{sg}}(\cdot)$ is a stop-gradient layer[1].

Since we expect the discriminator to extract more useful information from images in the target domain, we apply DCR only to real images. Although the application of DCR to generated images would be helpful for learning better representations in the discriminator, we believe that this regularization is not necessarily helpful for the better simulation of the target domain distribution(Refer to Section **C** in the supplementary).

3.3. Objective of Discriminator

The objective of the discriminator in the standard GAN is given by

$$\mathcal{L}_{\text{disc}} = -\mathbb{E}_{x,y}[\log D_y(x)] - \mathbb{E}_{x,y}[\log(1 - D_y(G(x)))],$$

where $D_y(\cdot)$ denotes the output of the discriminator corresponding to domain y. The proposed approach jointly minimizes the standard GAN loss and the DCR loss, which is given by

$$\mathcal{L}_{\text{D}} = \mathcal{L}_{\text{disc}} + \lambda \cdot \mathcal{L}_{\text{DCR}}, \tag{4}$$

where λ is a hyperparameter set to 1 in our experiments.

3.4. Implementation Details

This subsection discusses a couple of crucial design issues of our approach. We provide the further details about our implementation in Section A of the supplementary file.

Location of dense representation We impose DCR to the output of the final residual block or the input of the final convolution layer in the discriminator. Since the performance gain with the proposed DCR depends heavily on the quality of dense representations, it is important to identify the proper levels of representations for the regularization. We conduct ablation studies by varying the locations for DCR within the network. More details about this issue are discussed in Section 4.5.

DCR loss computation and positive pair selection We measure the DCR loss, $\mathcal{L}_{\text{DCR}}$, in (3) based on two local features $D_0(x_1)$ and $D_0(x_2)$ for the overlapping region. To compute $\text{sim}_{\text{nc}}(\cdot,\cdot)$, we adopt the approach described in PixPro [31]. The position and scale of each pixel in the two feature maps are first estimated and transformed to the original image space. Then, we compute the distances between all pairs of positions in the feature map and normalize the distances considering the estimated scales.

[1] $F_{\text{sg}}(z)$ means that z is frozen as a constant for backpropagation.

Method		Horse → Zebra		Winter → Summer		Cat → Dog (AFHQ)		AFHQ
Metric		FID↓	D&C ↑	FID↓	D&C↑	FID↓	D&C↑	FID↓
CycleGAN [37]	–	78.2±1.0	0.56±0.14 / 0.73±0.12	80.9±4.6	0.88±0.03 / **0.82**±0.06	85.9	0.54 / **0.48**	–
	DCR	**54.4**±3.3	**0.72**±0.01 / **0.89**±0.00	**74.6**±1.7	**0.91**±0.00 / **0.82**±0.06	71.0	**0.55** / 0.46	–
CUT [26]	–	43.2±2.3	0.73±0.06 / 0.87±0.02	77.8±0.5	0.56±0.12 / 0.51±0.24	76.2	0.38 / 0.41	–
	DCR	**34.0**±0.4	**0.96**±0.10 / **0.90**±0.0	**73.2**±0.5	**0.89±0.15 / 0.65±0.24**	68.4	**0.54 / 0.48**	–
FSeSim [36]		45.2±4.8	0.75±0.14 / 0.83±0.04	86.5±4.0	0.66±0.05 / 0.81±0.02	87.3	0.20 / 0.07	–
	DCR	**36.7**±1.4	**0.89**±0.02 / **0.89**±0.02	**74.5**±0.2	**0.83**±0.00 / **0.86**±0.02	**73.5**	**0.34 / 0.10**	–
MUNIT [13]	–	102.3±4.3	**0.29**±0.11 / 0.43±0.09	97.0±0.5	0.12±0.03 / 0.16±0.06	104.4	0.21 / 0.32	61.6
	DCR	**59.9**±0.1	0.28±0.04 / **0.44**±0.07	**91.2**±0.4	**0.19**±0.06 / **0.31**±0.01	**88.2**	**0.33 / 0.42**	**56.0**
StarGANv2 [5]	–	19.7	1.38 / 0.68	42.2	0.23 / **0.39**	44.2	0.93 / **0.69**	18.1 (16.2*)
	DCR	**19.4**	**1.65 / 0.73**	**39.7**	**0.27** / 0.38	**33.0**	**0.98** / 0.66	**17.4**
DRIT++ [22]	–	88.5	0.21 / 0.35	93.1	0.24 / 0.37	110.9	**0.25** / 0.19	–
	DCR	**67.1**	**0.29 / 0.44**	**82.6**	**0.28 / 0.44**	**107.8**	0.19 / **0.20**	–

Table 1. Quantitative comparison in terms of FID scores and D&C for various image-to-image translation models on the Horse → Zebra, Winter → Summer, Cat → Dog, and AFHQ datasets. Standard deviations are calculated from two runs. For the StarGAN v2, due to the inherent uncertainty of the model using the random latent codes, we present the average performance from our reproductions while the reported score in [5] is 16.2 on the AFHQ dataset.

PixPro [31] proposes to select positive pairs based on the fixed threshold for the distance in the whole batches. However, DCR can be applied to any locations in the dense representation map, we should consider feature map size as an additional factor. Hence, the positive pairs for the feature correspondence are identified by $\tilde{\Omega}$, whose elements are given values as follows:

$$\tilde{\Omega}(i,j) = \begin{cases} 1, & \text{if } \operatorname{dist}(i,j) \leq \tau \cdot s_f \\ 0, & \text{otherwise} \end{cases}, \tag{5}$$

where $\operatorname{dist}(\cdot,\cdot)$ denotes the normalized distance between of two locations within an image, and τ and s_f indicate the hyperparameters for the threshold value and the spatial resolution of the feature map, respectively. We set τ to 0.5 in our experiment and present its impact on the accuracy in Section 4.5.

4. Experiments

We verify the effectiveness of DCR in three different aspects: (a) image-to-image translation performance with one-sided and two-sided translation models, (b) benefit of the proposed regularizer with limited availability of training data, and (c) applicability to unconditional GANs. We also conduct a few ablation studies to show the robustness of the proposed approach.

4.1. Experiment Setup

We analyze our method mainly on image-to-image translation since the verification of the desired properties in an output image is more straightforward in conditional GAN models. The image-to-image translation task typically involves two distinct problems—shape deformations and texture changes, we evaluate the performance of the proposed approach in both aspects. Since DCR is a generic consistency regularization technique for the discriminator of a GAN, we test its applicability to unconditional GAN models. Note that an unconditional GAN model maps the predefined latent distribution to the distribution in the target domain. Hence, we consider an unconditional GAN task as a special case of a conditional GAN problem with a latent source domain while the source domain of the image-to-image translation is defined by the images in the corresponding training dataset.

Tested models Existing unpaired image-to-image translation approaches belong to either the two-sided or one-sided framework. The two-sided framework exploits both forward and backward mappings between the source and target domains. We apply DCR to CycleGAN [37], which is one of the most representative works in the two-sided framework. We also adopt MUNIT [13] and StarGANv2 [5], which employ the cycle consistency loss at the feature level and the pixel level, respectively. In addition, we apply DCR to DRIT++ [22], which utilizes disentangled representations for image-to-image translation. As the one-sided baseline models, we employ CUT [26] based on the contrastive patch relation and FSeSim [36] based on the structure similarity. For unconditional GANs, we employ SNDCGAN [24] as the baseline and also augment ContraD [15], a recently proposed contrastive regularization method at the instance level, to achieve additional performance boosting.

Datasets and metrics The datasets for image-to-image translation tasks need to contain images with geometric deformations or texture changes across domains. We carry out extensive experiments to verify the effectiveness of DCR on three commonly used datasets for the image-to-image translation. The tasks related to texture changes are evaluated

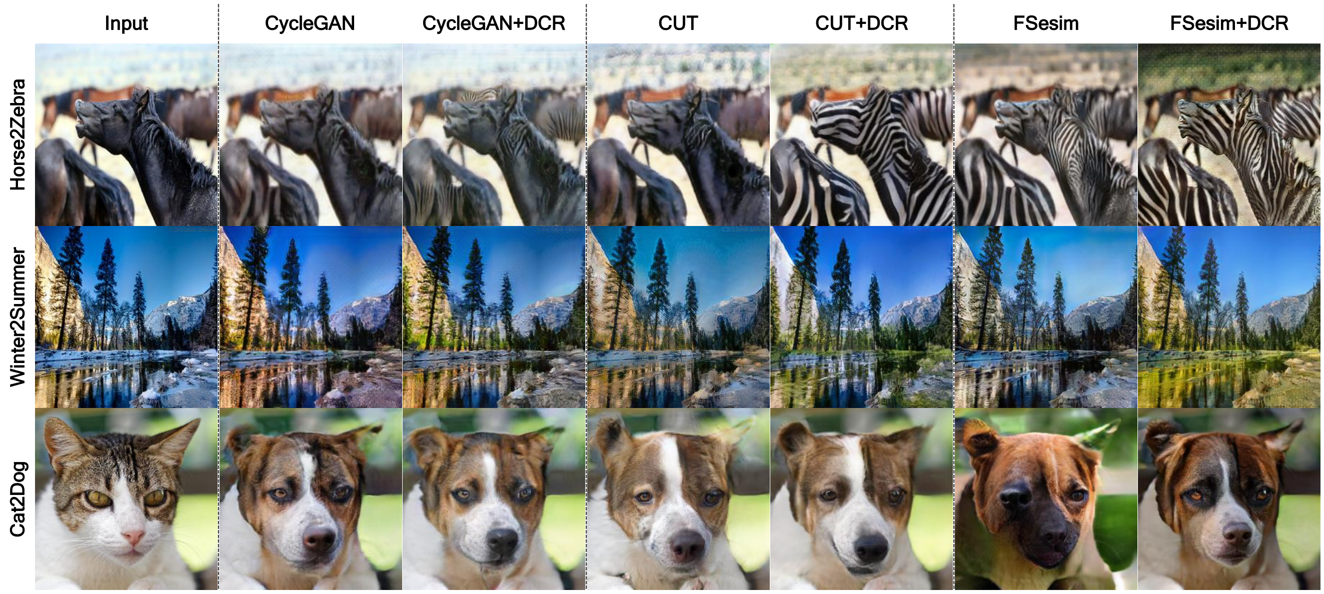

Figure 3. Qualitative comparison of image-to-image translation results on the Horse → Zebra, Winter → Summer, and Cat → Dog datasets. CycleGAN [37], CUT [26], and FSeSim [36] are employed as the baseline models and the proposed DCR is integrated into those models. The proposed DCR can add the appropriate patterns of the zebra on the horse and more realistic images compared to the baseline algorithms.

on the Horse → Zebra and the Winter → Summer dataset. The Cat → Dog dataset from AFHQ is employed to test the shape deformations and geometric transformations.For the image-to-image translation with multiple domains, we also use the AFHQ dataset [5], high-quality animal faces with large intra- and inter-class variations. For unconditional GANs, we utilize CIFAR-10 [20], which contains $60K$ 32×32 images with 10 labels, $50K$ for training and $10K$ for testing. We measure a Frechet Inception Distance (FID) [12] as a quantitative metric to evaluate generation quality and how accurate the target distribution is. We also report the density and coverage (D&C) [25], which simultaneously calculate diversity and fidelity of generated results.

4.2. DCR for Image-to-Image Translation

For the evaluation of DCR with various existing image-to-image translation models, we use the official implementation of each model and incorporate DCR into it. We employ Horse → Zebra, Winte → Summer, and Cat → Dog (AFHQ) to evaluate the models for a single domain. Since StarGANv2 [5] and MUNIT [13] can handle multiple domains, they are also tested on the AFHQ dataset. We trained three MUNIT [13] models for each direction and computed the average of FIDs followed by [5].

Table 1 presents the comprehensive results and demonstrates consistent improvements over all the baseline models on the tested datasets. For StarGANv2 [5] using AFHQ dataset, we would like to note that we reported the average of the best FID scores from 3 trials. There are some gaps between the reproduced results and the reported ones in the original paper [5]. This may come from the underlying randomness due to the use of random vectors for latent guided translation. Therefore, we compare the performance using the reproduced results.

It is noteworthy that significantly improved FID scores are achieved by the proposed DCR in the same setup with the baseline models without modifying the hyperparameters. The DCR loss turns out to be effective for shape deformations, which is validated by consistently improved results on the Cat → Dog dataset in terms of FID. As described in Section 3.4, we apply DCR to the input of the last convolution layer, which is more advantageous for shape deformation tasks according to our analysis presented in Section 4.5. We also employ the recently introduced metric, D&C [25], and confirm consistently improved performance compared to the base algorithms except few exceptions.

Figure 3 illustrates uni-modal image-to-image translation results using the baseline models, CycleGAN [37], CUT [26], and FSeSim [36], and the ones with the DCR integration into these methods. It is worth mentioning that we observe more realistic local semantics and structures in the generated images with DCR compared to the baseline models. In the case of the Horse → Zebra dataset, CUT [26] fails to provide an image with the desired zebra style while the integration of DCR into CUT [26] is effective to generate a more zebra-like image from the given horse image. Overall, DCR consistently provides better results for various datasets compared to the baseline models. Refer to

	StarGANv2 [5]	StarGANv2 [5] + DCR
10 %	22.63 ± 3.70	17.15 ± 0.89
30 %	19.08 ± 4.88	16.88 ± 1.17
100 %	17.86 ± 0.54	16.72 ± 0.59

Table 2. Quantitative comparison of the best FID score for the effect of DCR with few target data. We randomly selected 10%, 30%, and 100% of the wild domain of AFHQ dataset. The best FID score is the average of the best FID score for cat-to-wild translation and the best FID score for dog-to-wild translation. We report the mean and standard deviation of best FIDs across 3 trials.

Section B in the supplementary document for more qualitative results from uni-modal and multi-modal image-to-image translation models.

4.3. DCR with Few Target Data

Although we achieved the FID improvement by applying DCR to various existing models, we wonder whether the proposed DCR effectively reflects the local context of target domain even under few data scenarios. To investigate this, we randomly reduce the proportion of real data from a specific domain in the training set to 30%, and 10%. We perform experiments with the StarGANv2 [5] and the AFHQ dataset, consisting of dog, cat, and wild domains, as a baseline. We only reduce the wild domain, which has various intra-variations (fox, cheetah, lion, and tiger). For a fair comparison, we report the average performance across three trials. Quantitative and qualitative results are shown in Table 2 and Figure 4, respectively.

We observe that the FID score variance of StarGANv2 [5] is relatively large under few data scenarios. Since data were randomly selected, the FID varied depending on the similarity of the selected data to the test data. However, the proposed DCR shows less FID variance score than baseline. This implies that the proposed method effectively captures the local context of the target domain.

Figure 4 illustrates the reference-guided image-to-image translation results for the AFHQ datset, where we only used 10% of the data for the wild domain, while we utilized the overall data for other domains. It is worth mentioning that StarGAN v2 [5] with DCR provides significantly better translated images compared to the baseline. In particular, due to the small amount of data in the wild domain, the transformed images from the cat and dog images cannot reflect the style of the cheetah image and became lion images. However, the proposed DCR properly encoded the style from the cheetah image, and translated into the appropriate cheetah images. In addition, the proposed DCR encourages the network to generate the translated images while maintaining the geometry of the source images as well. These results clearly confirm that the DCR is a powerful regularization for efficient training of small dataset as well as for translation quality when combined with existing algorithms for image-to-image translation task.

Method	CIFAR-10
SNDCGAN [24]	97.4
SNDCGAN+ContraD [15]	10.9
SNDCGAN+Our	8.6
SNDCGAN+ContraD+Our	**7.7**

Table 3. Quantitative comparison of the best FID score on unconditional image generation.

4.4. DCR with Unconditional GANs

Since our DCR regularizes consistency of the discriminator, it is natural to study for unconditional GAN. We take SNDCGAN [24] as our baseline model and compare with a recent ContraD [15] which is instance-level contrastive learning based regularization method on CIFAR10 dataset for simplicity.

The motivation of our work is the hypothesis that image generation requires pixel-level prediction and the dense regularization of representations is appropriate. Table 3 shows the quantitative results that DCR more improves FID than instance-level method. Indeed, the results show the possibility that dense and instance-level consistency regularization technique can boost the each others performance by fusing ContraD and DCR. However, the role of instance-level and dense-level consistency regularization is still open area and we believe it deserves further study.

4.5. Ablation Study

To better understand how the hyper-parameters of proposed method affect performance, we conduct an ablation study. We perform experiments on CycleGAN [37] model with Horse$\rightarrow$Zebra and Cat $\rightarrow$ Dog(ImageNet [6]) dataset.

Where to regularize One of important choices in our algorithm is where to apply the proposed regularization. We conduct experiments on two types of tasks that require shape deformation task and texture translation with preservation of the shape. We measure a FID when we integrate DCR to different representation of CycleGAN's discriminator. The results are shown in Table 4.

The quantitative result shows proposed DCR improves the performance wherever applied to any representations. However improvement gap shows different aspect of behaviour at two types of dataset. The texture translation task shows better performance when representation is closer to pixel-level. On the other hand, shape deformation task provides better performance with the higher level representation, because it requires more semantic information compared to other tasks. In order to perform on overall tasks,

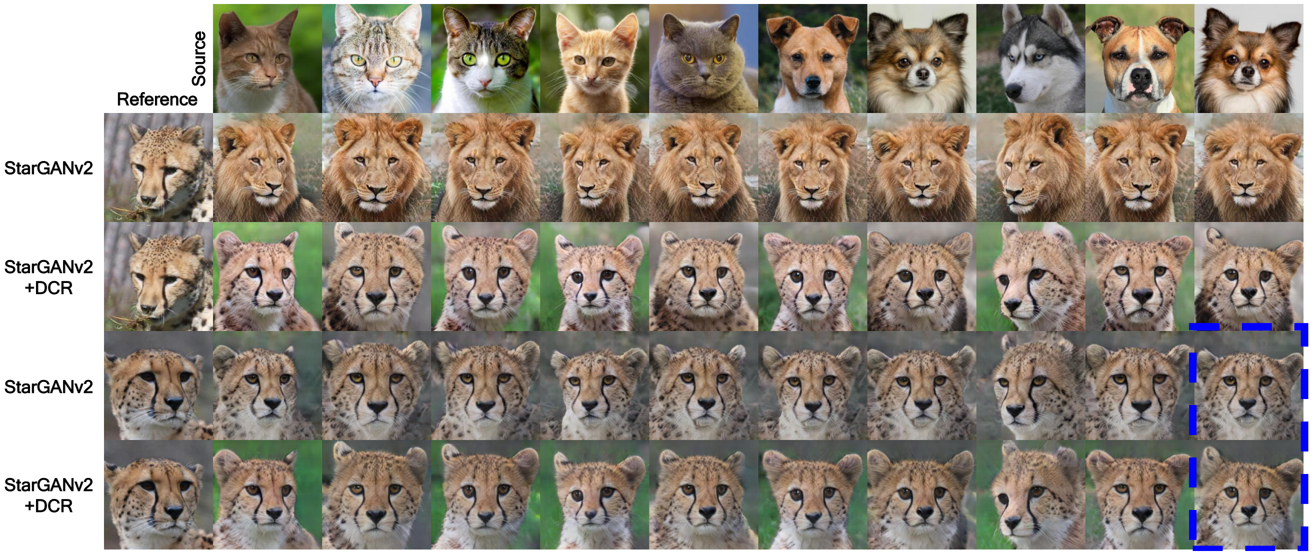

Figure 4. Qualitative comparison of reference guided samples on StarGANv2 and ours. Dashed line at the right corner of the image shows that the proposed DCR synthesizes more similar structure of the source image than baseline.

	Horse → Zebra	**Cat → Dog***
CycleGAN [37]	77.2	86.5
CycleGAN + DCR (layer2)	49.2	70.5
CycleGAN + DCR (layer3)	51.4	59.9
CycleGAN + DCR (layer4)	51.1	57.8

Table 4. Ablation study on the location of dense representation for the proposed DCR method. The layer number indicates which output of this layer is used as a dense representation. The asterisk (*) denotes the experiment using the examples in ImageNet.

τ	0	0.3	0.5	0.7	0.9
FID	77.2	56.7	**51.1**	51.4	57.9

Table 5. Ablation study on distance threshold τ for the proposed DCR method. We conducted the ablation study on CycleGAN [37] model with Horse → Zebra dataset. $\tau = 0$ indicates the baseline model without the proposed DCR.

we select the representation either the output of final residual block or the input of final convolution layer.

How to identify positive pairs One of the major hyperparameters in DCR is the distance threshold τ in (5) to identify the positive pairs for the feature correspondence $\tilde{\Omega}$. To choose the optimal threshold value τ, we experimented with various values $\tau \in \{0.3, 0.5, 0.7, 0.9\}$ on CycleGAN [37] model with Horse→Zebra dataset. Table 5 reports the quantitative comparison of the performance at the various distance threshold τ.

The results in Table 5 shows a consistent improvement over the baseline model ($\tau = 0$). This verify the effectiveness of the proposed DCR in providing better translated images. As shown in Table. 5, we achieve the best result when we set the distance threshold τ to 0.5. Therefore, the distance threshold τ is set to 0.5 in all experiments.

The further investigation to analyze the performance gain by the proposed DCR can be found in Supplementary Section **C**. We conducted the ablation studies to understand the effect of stop-gradient and the reason we apply the DCR only to the cropped regions of generated images, not to the entire images or real images.

5. Conclusion

We presented a novel regularization technique, referred to as dense consistency regularization (DCR). The proposed approach enforces the consistency between the representations of the overlapping regions in two different views from the same image. DCR is suitable for the tasks that require dense prediction and can be incorporated into various existing conditional and unconditional GAN models. According to our experiments for image-to-image translation and unconditional image generation tasks, DCR achieved outstanding performance consistently. Moreover, DCR captures a local context in the target domain effectively with only a small fraction of data and it also leads to extra performance gains through the combination with instance-level regularization methods. Refer to Section E of the supplementary document for discussions about potential negative societal impacts and limitation.

References

[1] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. *Advances in Neural Information Processing Systems*, 30, 2017. 3

[2] 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. 3

[3] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-supervised gans via auxiliary rotation loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12154–12163, 2019. 1, 3

[4] 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. 3, 4

[5] 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. 2, 3, 5, 6, 7

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 7

[7] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2427–2436, 2019. 3

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1, 2

[9] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 3

[10] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NIPS*, 2017. 3

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 6

[13] 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. 2, 3, 5, 6

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[15] Jongheon Jeong and Jinwoo Shin. Training GANs with stronger augmentations via contrastive discriminator. In *International Conference on Learning Representations*, 2021. 1, 3, 5, 7

[16] Minguk Kang and Jaesik Park. ContraGan: contrastive learning for conditional image generation. *arXiv preprint arXiv:2006.12681*, 2020. 1, 3

[17] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020. 1

[18] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pages 1857–1865. PMLR, 2017. 3

[19] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*, 2017. 3

[20] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 6

[21] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 3

[22] 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. 3, 5

[23] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017. 3

[24] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. 3, 5, 7

[25] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, 2020. 6

[26] 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. 1, 2, 3, 4, 5, 6

[27] Pedro O Pinheiro, Amjad Almahairi, Ryan Y Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. In *NeurIPS*, 2020. 3

[28] Byungseok Roh, Wuhyun Shin, Ildoo Kim, and Sungwoong Kim. Spatially consistent representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1153, 2021. 3

[29] Kevin Roth, Aurélien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *NIPS*, 2017. 3

[30] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 3

[31] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021. 3, 4, 5

[32] Ning Yu, Guilin Liu, Aysegul Dundar, Andrew Tao, Bryan Catanzaro, Larry S Davis, and Mario Fritz. Dual contrastive loss and attention for GANs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6731–6742, 2021. 1, 3

[33] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021. 3

[34] Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. Consistency regularization for generative adversarial networks. In *International Conference on Learning Representations*, 2020. 1, 3, 4

[35] 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. 3

[36] 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. 2, 3, 5, 6

[37] 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, 3, 5, 6, 7, 8