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

Minsu Ko\textsuperscript{1}\textsuperscript{*} Eunju Cha\textsuperscript{1}\textsuperscript{*} Sungjoo Suh\textsuperscript{1} Huijin Lee\textsuperscript{1} Jae-Joon Han\textsuperscript{1} Jinwoo Shin\textsuperscript{2} Bohyung Han\textsuperscript{3}

\textsuperscript{1}Samsung Advanced Institute of Technology (SAIT), South Korea
\textsuperscript{2}Korea Advanced Institute of Science and Technology (KAIST), South Korea
\textsuperscript{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.

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 the 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 $\Omega$ while $\tilde{\Omega}$ denotes the binary map indicating matching pairs of pixels.

...consistency 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 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 addressed 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 category. The auxiliary rotation loss for self-supervision, by which both the real and generated images are classified into one category 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 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 category. The auxiliary rotation loss for self-supervision, by which both the real and generated images are classified into one category 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 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].

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

$$D = D_1 \circ D_0.$$  

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., $128 \times 128$) 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 quality of dense representations, it is important to identify the proper levels of representations for the regularization. We provide the further details about our approach. We provide the further details about our implementation in Section A of the supplementary file.

The proposed DCR module, denoted by $R(\cdot)$, consists of two $1 \times 1$ 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 $r_2 := D_0(x_2)$. Given the overlapping region $\Omega$ of two views $x_1$ and $x_2$, we define the negative cosine similarity of their corresponding features, which is given by

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

where $\hat{\Omega}$ 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

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

where $F_{sg}(\cdot)$ 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

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

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

$$L_D = L_{\text{disc}} + \lambda \cdot L_{\text{DCR}}, \quad (4)$$

where $\lambda$ 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, $L_{\text{DCR}}$, in (3) based on two local features $D_0(x_1)$ and $D_0(x_2)$ for the overlapping region. To compute $\text{sim}_{nc}(\cdot, \cdot)$, 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.

---

$^1$ $F_{sg}(z)$ means that $z$ is frozen as a constant for backpropagation.
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—shaping 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 SND-CGAN [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 in Section 4.5.
on the Horse $\rightarrow$ Zebra and the Winter $\rightarrow$ Summer dataset. The Cat $\rightarrow$ 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 $60\,K \times 32 \times 32$ images with 10 labels, $50\,K$ for training and $10\,K$ 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 $\rightarrow$ Zebra, Winter $\rightarrow$ Summer, and Cat $\rightarrow$ 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 $\rightarrow$ 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 $\rightarrow$ 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
from a specific domain in the training set to 30\%
gate this, we randomly reduce the proportion of real data
the proposed DCR effectively reflects the local context of
ing DCR to various existing models, we wonder whether
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 investiga-
tion 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 Star-
GANv2 [5] is relatively large under few data scenarios.
Since data were randomly selected, the FID varied depend-
ing 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 effect-
ively captures the local context of the target domain.

Figure 4 illustrates the reference-guided image-to-
image translation results for the AFHQ dataset, 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 par-
cular, 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 ap-
propriate 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 pow-
erful regularization for efficient training of small dataset as
well as for translation quality when combined with existing
algorithms for image-to-image translation task.

4.4. DCR with Unconditional GANs

Since our DCR regularizes consistency of the discrimi-
nator, 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 reg-
ularization of representations is appropriate. Table 3 shows
the quantitative results that DCR more improves FID than
instance-level method. Indeed, the results show the possi-
bility that dense and instance-level consistency regulariza-
tion 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 pro-
posed method affect performance, we conduct an ablation
study. We perform experiments on CycleGAN [37] model
with Horse→Zebra and Cat→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 preser-
vation of the shape. We measure a FID when we integrate
DCR to different representation of CycleGAN’s discrimina-
tor. 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 be-
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 representa-
tion, because it requires more semantic information com-
pared 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.

<table>
<thead>
<tr>
<th></th>
<th>Horse $\rightarrow$ Zebra</th>
<th>Cat $\rightarrow$ Dog$^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CycleGAN [37]</td>
<td>77.2</td>
<td>86.5</td>
</tr>
<tr>
<td>CycleGAN + DCR (layer2)</td>
<td>49.2</td>
<td>70.5</td>
</tr>
<tr>
<td>CycleGAN + DCR (layer3)</td>
<td>51.4</td>
<td>59.9</td>
</tr>
<tr>
<td>CycleGAN + DCR (layer4)</td>
<td>51.1</td>
<td>57.8</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>0</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>FID</td>
<td>77.2</td>
<td>56.7</td>
<td><strong>51.1</strong></td>
<td>51.4</td>
<td>57.9</td>
</tr>
</tbody>
</table>

Table 5. Ablation study on distance threshold $\tau$ for the proposed DCR method. We conducted the ablation study on CycleGAN [37] model with Horse $\rightarrow$ 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 $\tau$ in (5) to identify the positive pairs for the feature correspondence $\tilde{\Omega}$. To choose the optimal threshold value $\tau$, we experimented with various values $\tau \in \{0.3, 0.5, 0.7, 0.9\}$ on CycleGAN [37] model with Horse $\rightarrow$ Zebra dataset. Table 5 reports the quantitative comparison of the performance at the various distance threshold $\tau$.

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 $\tau$ to 0.5. Therefore, the distance threshold $\tau$ 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


