論文 / 著書情報
Article / Book Information

Title	Augmented Cyclic Consistency Regularization for Unpaired Image-To-Image Translation
Author	Takehiko Ohkawa, Naoto Inoue, Hirokatsu Kataoka, Nakamasa Inoue
Journal/Book name	Proceedings of ICPR 2020 25th International Conference on Pattern Recognition, , , pp. 362-369
Pub. date	2021, 5
DOI	https://doi.org/10.1109/ICPR48806.2021.9412082
Copyright	(c)2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Note	This file is author (final) version.

Augmented Cyclic Consistency Regularization for Unpaired Image-to-Image Translation

Takehiko Ohkawa*, Naoto Inoue*, Hirokatsu Kataoka†, Nakamasa Inoue‡
*The University of Tokyo, Japan,
†National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki, Japan,
‡Tokyo Institute of Technology, Japan.
Email: ohkawa-t@iis.u-tokyo.ac.jp, inoue@hal.t.u-tokyo.ac.jp, hirokatsu.kataoka@aist.go.jp, inoue@c.titech.ac.jp

***Abstract*—Unpaired image-to-image (I2I) translation has received considerable attention in pattern recognition and computer vision because of recent advancements in generative adversarial networks (GANs). However, due to the lack of explicit supervision, unpaired I2I models often fail to generate realistic images, especially in challenging datasets with different backgrounds and poses. Hence, stabilization is indispensable for GANs and applications of I2I translation. Herein, we propose Augmented Cyclic Consistency Regularization (ACCR), a novel regularization method for unpaired I2I translation. Our main idea is to enforce consistency regularization originating from semi-supervised learning on the discriminators leveraging real, fake, reconstructed, and augmented samples. We regularize the discriminators to output similar predictions when fed pairs of original and perturbed images. We qualitatively clarify why consistency regularization on fake and reconstructed samples works well. Quantitatively, our method outperforms the consistency regularized GAN (CR-GAN) in real-world translations and demonstrates efficacy against several data augmentation variants and cycle-consistent constraints.**

I. Introduction

Image-to-image (I2I) translation aims to learn a function by mapping images from one domain to another. The I2I framework is applied to many tasks in the fields of machine learning and computer vision such as image-inpainting [1], super-resolution [2], [3], colorization [4], style transfer [5], [25], [26], [28], [27], domain adaptation [6], [7], [8], [9], [42], and person re-identification [10], [11], [12], [13], [14]. We face challenges either in collecting aligned image pairs for training (e.g., *summer* $\rightarrow$ *winter*) or its inexistence (e.g., *artwork* $\rightarrow$ *photo*); thus, most work focuses on unpaired I2I models under the assumption that paired data are not available. However, trade-offs arise in training stability due to the absence of paired supervision. Even more problematic, the unpaired setting is based on ill-posed problems having infinitely many solutions and multimodal outputs where a single input may correspond to multiple possible outputs. To handle this, models employ complex and disentangled architectures [22], [23], [24], which pose substantial difficulties from an optimization perspective.

In recent years, several variants of cycle-consistency constraints [42], normalization techniques [25], [26], [27], [28], [29], [30], and different latent space assumptions [20], [22], [24], [31] have been investigated to achieve semantic-aware cycles, control style information, and disentangle features. Despite these advances, stabilization is rarely discussed because the issue is avoided by dealing with narrow translations or utilizing refined datasets with similar poses and backgrounds. Furthermore, when it comes to real-world applications such as domain adaptation and person re-identification, data include various blurs, illuminations, or noise; therefore, training is not straightforward.

For GANs, several normalizations and gradient-based regularization techniques for the GAN discriminator have been studied such as batch normalization [32], layer normalization [33], spectral normalization [34], and the gradient

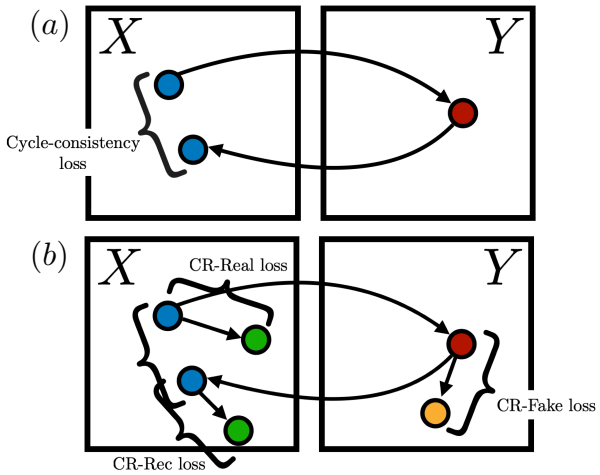


Fig. 1: **Illustration of our approach.** (a) CycleGAN [18] contains two generators $G : X \mapsto Y$ and $F : Y \mapsto X$, and cycle-consistency loss, which enforces pixel-wise matching in the form of L1 loss between real and reconstructed data. (b) Augmented Cyclic Consistency Regularization (ACCR) is an extension of consistency regularization on discriminators in unpaired I2I models leveraging each real $x \in X$ and reconstructed $F(G(x)) \in X$ sample pair (**Blue**), and augmented samples $T(x) \in X$, $T(F(G(x))) \in X$ (**Green**). $T(\cdot)$ denotes a semantics-preserving data augmentation function. A fake sample $G(x) \in Y$ is also employed (**Red**) translated from domain X and the augmented sample $T(G(x)) \in Y$ (**Yellow**). For simplicity, the other cycle is omitted.

This work was done when the first author was at the Tokyo Institute of Technology.

penalty [37]. However, [36] empirically revealed that simultaneous enforcement of both normalization and gradient-based regularization provides marginal gains or fails. Especially, I2I models adopt several normalization techniques [25], [26], [27], [28], [29], [30] and accordingly I2I models with state-of-the-art gradient-based regularization are associated with more theoretical uncertainty than standard GANs.

Zhang et al. [40] first put forward consistency regularized GAN (CR-GAN) whereby consistency regularization was introduced to the GAN discriminator from semi-supervised learning. CR-GAN surpasses gradient-based approaches, but CR-GAN is limited to real samples and regularization failures can occur using generated images for standard GANs.

In this work, we propose augmented cyclic consistency regularization (ACCR), a novel regularization technique in unpaired I2I translation without gradient information which incorporates consistency regularization on the discriminators leveraging three types of samples: real, fake, and reconstructed images. We augment these data feeding to the discriminators and penalize sensitivity to perturbations. We show an intuitive illustration of our method in CycleGAN [18] in Fig. 1.

Qualitatively, I2I models guarantee quality of both fake and reconstructed samples due to faster learning and lower potential of mode collapse. Thus, we justify the use of these images. Quantitatively, our method outperforms the CycleGAN baseline, the CR-GAN method, and models with consistency regularization using fake and reconstructed samples respectively on MNIST $\leftrightarrow$ MNIST-M, MNIST $\leftrightarrow$ SVHN, Maps $\leftrightarrow$ Aerial photo, and Labels $\leftrightarrow$ Photo. ACCR-CycleGAN improves the baseline by 0.3% on MNIST $\rightarrow$ MNIST-M, 2.3% on MNIST-M $\rightarrow$ MNIST, 3.9% on MNIST $\rightarrow$ SVHN, and 3.7% on SVHN $\rightarrow$ MNIST as measured by classification accuracy on fake samples. In real-world translations, ACCR-CycleGAN surpasses the baseline by 26.4% on Maps $\rightarrow$ Photo, 22.9% on Photo $\rightarrow$ Maps, 29.8% on Photo $\rightarrow$ Labels as evaluated by mean squared error with ground-truth images. More importantly, even in real-world tasks, ACCR can always further improve the state-of-the-art CR technique. In extensive ablation studies, ACCR outperforms the CR-GAN method in other types of data augmentation and cycle-consistent constraints.

The contributions of our work are summarized as follows:

- We propose a novel, simple, and effective training stabilizer in unpaired I2I translation using real, fake, and reconstructed samples.
- We qualitatively explain why consistency regularization employing fake and reconstructed samples performs well in unpaired I2I models.
- Our ACCR quantitatively outperforms the CycleGAN baseline and the CR-GAN method in several datasets, for various cycle-consistent constraints, and with several commonly used data augmentation techniques, as well as combinations thereof.

II. Related Work

A. Image-to-Image Translation

To learn the mapping function with paired training data, Pix2Pix [16] applies conditional GANs using both a latent vector and the input image. The constraint is enforced by the ground truth labels or pairwise correspondence at the pixel-level. CycleGAN [18], DiscoGAN [19], and UNIT [20] employ a cycle-consistency constraint to simultaneously learn a pair of forward and backward mappings between two domains given unpaired training data, which is conditioned solely on an input image and accordingly produce one single output.

To achieve multimodal generation, BicycleGAN [17] injects noise into mappings between latent and target spaces to prevent mode collapse in the paired setting. In unpaired multimodal translations, augmented CycleGAN [21] also injects latent code in the generators, and concurrent work [22], [23], [24] adopts disentangled representations to produce diversified outputs.

In the field of domain adaptation, the CycleGAN framework is applied in cycle-consistent adversarial domain adaptation models [6], [9], [42] and I2I translation based domain adaptation [7], [8] is designed for semantic segmentation of the target domain images. Currently, GAN-based domain adaptation is introduced in person re-identification for addressing challenges in real-world scenarios. CycleGAN-based approaches [10], [11], [12], [13] are widely adopted to transfer pedestrian image styles from one domain to another. State-of-the-art DG-Net [14] makes use of disentangled architecture to encode pedestrians in appearance and structure spaces for implausible person image generation.

However, despite the wide range of use cases, unpaired I2I translation is more difficult from an optimization perspective because of the lack of supervision in the form of paired examples. Moreover, the latest multimodal methods incorporate domain-specific and domain-invariant encoders [22], [23], [24], [31]. These approaches often fail when the amount of training data is limited, or domain characteristics differ significantly [24]. It is problematic to learn separate latent spaces, larger networks, and unconditional generation where the latent vector can be simply mapped to a full-size image in contrast to the previous conditional cases. Therefore, our work mainly focuses on the stabilization of unpaired I2I translation.

In general, all the models share a problem whereby the generators cannot faithfully reconstruct the input images since I2I models are inherently one-to-many mappings. For instance, in the translation of *semantic labels* $\rightarrow$ *photo*, original colors, textures, and lighting are impossible to fully recover and stochastically vary because the details are lost in the label domain. This is also the case for all other translations such as *map* $\leftrightarrow$ *photo*, and *summer* $\leftrightarrow$ *winter*, as well as digits. In our work, we make use of this drawback for improving the diversity of fake and reconstructed images in consistency regularization.

B. Consistency Regularization

Consistency regularization was first proposed in the semi-supervised learning literature [52], [53]. The fundamental idea is simple: a classifier should output similar predictions for unlabeled examples even after they have been randomly perturbed. The random perturbations contain data augmentation [48], [49], stochastic regularization (e.g. Dropout [51]) [52], [53], and adversarial perturbations [55]. Analytically, consistency regularization enhances the smoothness of function prediction [54], [55].

CR-GAN [40] introduces consistency regularization in the GAN discriminator and improves state-of-the-art FID scores for conditional generation. In addition, CR-GAN outperforms gradient-based regularizers: Gradient Penalty [37], DRAGAN [38] and JS-Regularizer [39]. However, the CR-GAN method on generated images often fails. We seek to explore this limitation and demonstrate the effectiveness of adding consistency regularization which employs both fake and reconstructed images.

III. Proposed Method

A. Preliminaries

The goal of unpaired I2I translation is to learn the mapping within two domains X_1 and X_2 given training data $\{x_{1i}\}_{i=1}^{N}$ and $\{x_{2i}\}_{i=1}^{M}$ where $x_{1i} \in X_1$ and $x_{2i} \in X_2$. We denote data distributions in two domains $x_1 \sim p_{data}(x_1)$, $x_2 \sim p_{data}(x_2)$, generators $G_1 : X_1 \rightarrow X_2$, $G_2 : X_2 \rightarrow X_1$, and discriminators D_1, D_2, where D_1 learns to distinguish real data $\{x_1\}$ from fake data $\{G_2(x_2)\}$, D_2 learns differences $\{x_2\}$ from $\{G_1(x_1)\}$. The objective consists of an adversarial loss [15] and a constraint term $\mathcal{C}$ to encourage generators to produce samples that are structurally similar to inputs, and avoid excessive hallucinations and mode collapse that would increase the loss.

1) Adversarial loss: The adversarial loss $\mathcal{L}_{\mathrm{GAN}}$ is employed to match the distribution of the fake images to the target image distribution, as written by

$$\begin{aligned}\mathcal{L}_{\mathrm{GAN}}\left(G_1, D_2\right) = \mathbb{E}_{x_2 \sim p_{\mathrm{data}}(x_2)}\left[\log D_2(x_2)\right] \\ + \mathbb{E}_{x_1 \sim p_{\mathrm{data}}(x_1)}\left[\log\left(1 - D_2(G_1(x_1))\right)\right].\end{aligned} \tag{1}$$

2) Unpaired I2I objective: Unpaired I2I translation requires the additional loss $\mathcal{C}$ to support forward and backward mappings between two domains. Thus, the full objective of unpaired I2I models (UI2I) is given by

$$\begin{aligned}\mathcal{L}_{\mathrm{UI2I}}\left(G_1, G_2, D_1, D_2\right) = \mathcal{L}_{\mathrm{GAN}}\left(G_1, D_2\right) + \mathcal{L}_{\mathrm{GAN}}\left(G_2, D_1\right) \\ + \mathcal{C}(G_1, G_2).\end{aligned} \tag{2}$$

CycleGAN [18] imposes a pixel-wise constraint in the form of cycle-consistency loss $\mathcal{L}_{\mathrm{CC}}$ [18] as the constraint term $\mathcal{C}$,

$$\begin{aligned}\mathcal{L}_{\mathrm{CC}}(G_1, G_2) = \lambda_1 \mathbb{E}_{x_1 \sim p_{\mathrm{data}}(x_1)}\left[\|G_2(G_1(x_1)) - x_1\|_1\right] \\ + \lambda_2 \mathbb{E}_{x_2 \sim p_{\mathrm{data}} x_2)}\left[\|G_1(G_2(x_2)) - x_2\|_1\right].\end{aligned} \tag{3}$$

B. Consistency Regularization for GANs

CR-GAN [40] proposes a simple, effective and fast training stabilizer introducing consistency regularization on the GANs discriminator. Assuming that the decision of the discriminator should be invariant to any valid domain-specific data augmentations, the sensitivity of the discriminator is penalized to randomly augmented data. It improves FID scores in conditional generations. The consistency regularization loss for discriminator D is given by

$$\mathcal{L}_{\text{CR-Real}}(D) = \mathbb{E}_{x \sim p_{\mathrm{data}(x)}}\left[\|D(x) - D(T(x))\|^2\right], \tag{4}$$

where $T(\cdot)$ denotes a stochastic data augmentation function, e.g., flipping the image horizontally or translating the image by a few pixels. However, [40] reports that an additional regularization using generated images is not always superior to the original CR-GAN method.

C. Augmented Cyclic Consistency Regularization

We propose augmented cyclic consistency regularization (ACCR) for stabilizing training in unpaired I2I models. ACCR enforces consistency regularization on discriminators leveraging real, fake, reconstructed, and augmented samples. The goal is to verify the effectiveness of consistency regularization, even where fake and reconstructed data are employed from datasets which include noise (e.g., SVHN or MNIST-M). An overview of ACCR-CycleGAN is shown in Fig. 2.

We define consistency regularization losses on discriminators D_1 and D_2 leveraging real, fake, and reconstructed data denoted by $\mathcal{L}_{\text{CR-Real}}$, $\mathcal{L}_{\text{CR-Fake}}$, and $\mathcal{L}_{\text{CR-Rec}}$, respectively. $\mathcal{L}_{\text{CR-Real}}$ which is identical to CR-GAN is written as

$$\begin{aligned}&\mathcal{L}_{\text{CR-Real}}(D_1, D_2) \\ &= \mathbb{E}_{x_1 \sim p_{\mathrm{data}}(x_1)}\left[\|D_1(x_1) - D_1(T(x_1))\|^2\right] \\ &+ \mathbb{E}_{x_2 \sim p_{\mathrm{data}}(x_2)}\left[\|D_2(x_2) - D_2(T(x_2))\|^2\right].\end{aligned} \tag{5}$$

Given fake samples $\{G_1(x_1)\}$, $\{G_2(x_2)\}$ and augmented samples $\{T(G_1(x_1))\}$, $\{T(G_2(x_2))\}$, $\mathcal{L}_{\mathrm{CR-Fake}}$ is written as

$$\begin{aligned}&\mathcal{L}_{\text{CR-Fake}}(D_1, D_2) \\ &= \mathbb{E}_{x_1 \sim p_{\mathrm{data}}(x_1)}\left[\|D_2(G_1(x_1)) - D_2(T(G_1(x_1)))\|^2\right] \\ &+ \mathbb{E}_{x_2 \sim p_{\mathrm{data}}(x_2)}\left[\|D_1(G_2(x_2)) - D_1(T(G_2(x_2)))\|^2\right],\end{aligned} \tag{6}$$

where $T(\cdot)$ denotes a stochastic data augmentation function which is semantics-preserving such as random crop, random rotation, or cutout [48]. Given reconstructed samples $\{G_2(G_1(x_1))\}$, $\{G_1(G_2(x_2))\}$ and augmented samples $\{T(G_2(G_1(x_1)))\}$, $\{T(G_1(G_2(x_2)))\}$, $\mathcal{L}_{\mathrm{CR-Rec}}$ is written as

$$\begin{aligned}&\mathcal{L}_{\text{CR-Rec}}(D_1, D_2) \\ &= \mathbb{E}_{x_1 \sim p_{\mathrm{data}}(x_1)}[\|D_1(G_2(G_1(x_1))) \\ &\qquad - D_1(T(G_2(G_1(x_1))))\|^2] \\ &+ \mathbb{E}_{x_2 \sim p_{\mathrm{data}}(x_2)}[\|D_2(G_1(G_2(x_2))) \\ &\qquad - D_2(T(G_1(G_2(x_2))))\|^2].\end{aligned} \tag{7}$$

By default, we use the random crop as $T(\cdot)$, and explore the effects of other functions in Section IV-E2.

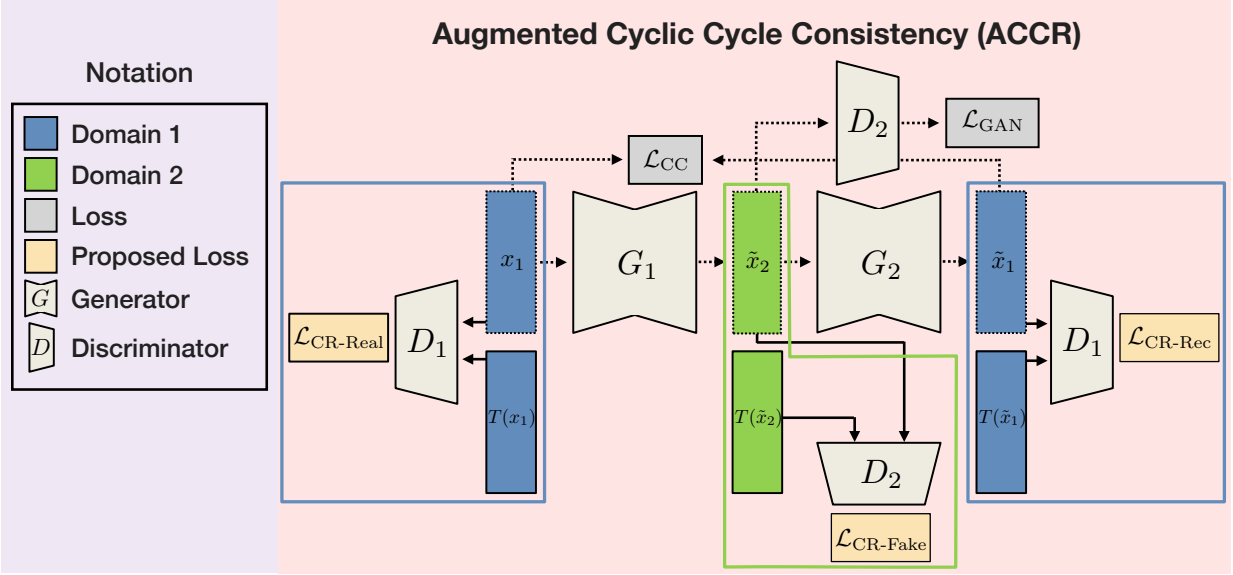


Fig. 2: **Overview of ACCR-CycleGAN.** CycleGAN [18] is depicted with Augmented Cyclic Consistency Regularization (ACCR), which consists of the CycleGAN architecture and consistency regularization losses on real, fake, and reconstructed images. The rectangles correspond to our proposed parts, $\mathcal{L}_{\text{CR-Real}}$, $\mathcal{L}_{\text{CR-Fake}}$, and $\mathcal{L}_{\text{CR-Rec}}$. $\mathcal{L}_{\text{GAN}}$ and $\mathcal{L}_{\text{CC}}$ indicate adversarial loss [15] and cycle-consistency loss [18], respectively. For expository purposes, the cycle in domain X_2 is omitted.

D. Full Objective

Finally, the objective of augmented cyclic consistency regularized unpaired I2I models (ACCR-UI2I) given unpaired data is written as

$$\begin{aligned}\mathcal{L}_{\text{UI2I}}^{\text{ACCR}}(G_1, G_2, D_1, D_2) = &\mathcal{L}_{\text{GAN}}(G_1, D_2) \\ &+ \mathcal{L}_{\text{GAN}}(G_2, D_1) \\ &+ \mathcal{C}(G_1, G_2) \\ &+ \lambda_{\text{Real}}(\mathcal{L}_{\text{CR-Real}}(D_1, D_2)) \\ &+ \lambda_{\text{Fake}}(\mathcal{L}_{\text{CR-Fake}}(D_1, D_2)) \\ &+ \lambda_{\text{Rec}}(\mathcal{L}_{\text{CR-Rec}}(D_1, D_2)), \quad (8)\end{aligned}$$

where the hyper-parameters λ_{Real}, λ_{Fake}, and λ_{Rec} control the weights of the regularization terms.

IV. Experiments

This section validates our proposed ACCR method in digit translations with various backgrounds (MNIST $\leftrightarrow$ MNIST-M and MNIST $\leftrightarrow$ SVHN) and real-world translations (Maps $\leftrightarrow$ Aerial photo and Cityscapes labels $\leftrightarrow$ Photo). First, we present details concerning the datasets and experimental implementation. Next, we conduct a quantitative analysis to demonstrate the performance on the translations and investigate the feature distance between real and augmented samples in discriminators for verifying the effect of ACCR. We then conduct a qualitative analysis for the generated samples and failure cases of the CR-GAN. Finally, we conduct ablation studies to compare consistency regularization utilizing fake and reconstructed images and explore the importance of choices with respect to data augmentation and cycle-consistent constraints.

A. Datasets

MNIST (M) $\leftrightarrow$ MNIST-M (MM): MNIST [43] contains centered, 28×28 pixel grayscale images of single-digit numbers on a black background, 60,000 images for training and 10,000 for validation. We rescale to 32×32 pixels and extend the channel to RGB. MNIST-M [44] contains centered, 32×32 pixel digits on a variant background which is substituted by a randomly extracted patch obtained from color photos from BSDS500 [45], 59,000 images for training and 1,000 for validation.

MNIST (M) $\leftrightarrow$ SVHN (S): We preprocess MNIST [43] as above. SVHN [46] is the challenging real-world Street View House Number dataset, much larger in scale than the other considered datasets. It contains 32×32 pixel color samples, 73,257 images for training and 26,032 images for validation. Besides varying the shape and texture, the images often contain extraneous numbers in addition to those which are labeled and centered.

Google Aerial Photo $\leftrightarrow$ Maps: The dataset consists of 3292 pairs of aerial photos and corresponding maps. In our experiments, we resize the images from 600×600 pixels to 256×256 pixels. We use 1098 images for training and 1096 images for testing.

Cityscapes Labels $\leftrightarrow$ Photo: The dataset contains 2975 pairs of training images from the Cityscapes training set [47] and its segmentation labels. In our experiments, we resize the images to 256×256 pixels. We use the Cityscapes validation set for testing.

TABLE I: **Digit translations.** Direction indicates source → target translation direction. We measure the impact of the proposed method, ACCR, with respect to classification accuracy (%) on fake samples in the target domain by fixed classifiers and compare with the CycleGAN [18] baseline and CR-CycleGAN. * indicates statistically significant differences at the level of significance $\alpha = 0.05$. For ablation we further experiment using CycleGAN with CR-Fake and CR-Rec.

Model	M → MM	MM → M	M → S	S → M
CycleGAN	97.7 ± 0.3	92.2 ± 1.2	47.1 ± 3.1	28.2 ± 0.9
CR-CycleGAN	97.7 ± 0.6	94.3 ± 0.5*	43.7 ± 4.1	29.6 ± 0.7
CR-CycleGAN + CR-Fake (Ours)	97.7 ± 0.3	93.8 ± 0.7*	46.3 ± 4.7	$\mathbf{31.9 \pm 3.0}$*
CR-CycleGAN + CR-Rec (Ours)	97.6 ± 1.2	94.2 ± 1.4*	48.5 ± 3.0	30.5 ± 1.7*
ACCR-CycleGAN (Ours)	$\mathbf{98.0 \pm 0.5}$	$\mathbf{94.5 \pm 0.5}$*	$\mathbf{51.0 \pm 5.2}$	$\mathbf{31.9 \pm 1.6}$*

TABLE II: **Maps ↔ Aerial photograph.** We evaluate the generation quality for fake samples by mean square error (MSE) with ground-truth images.

Model	Maps → Photo	Photo → Maps
CycleGAN	0.159	**0.023**
CR-CycleGAN	0.120	**0.023**
ACCR-CycleGAN (Ours)	**0.117**	**0.023**

TABLE III: **Cityscapes labels ↔ Photograph.** We evaluate the generation quality for fake samples by mean square error (MSE) with ground-truth images.

Model	Labels → Photo	Photo → Labels
CycleGAN	0.728	0.057
CR-CycleGAN	0.625	0.065
ACCR-CycleGAN (Ours)	**0.561**	**0.040**

B. Implementation

1) Network architecture: We adopt architecture for our networks based on Hoffman et al [6]. The generator consists of two slide-2 convolutional layers followed by two residual blocks and then two deconvolution layers with slide $\frac{1}{2}$. The discriminator network consists of PatchGAN [16] with 5 convolutional layers. For all digit experiments, we use a variant of LeNet [43] architecture with 2 convolutional layers and 2 fully connected layers for 32×32 pixel images.

2) Training details: In terms of $\mathcal{L}_{\mathrm{GAN}}$, we replace binary cross-entropy loss by a least-squares loss [35] to stabilize GANs optimization as per [18]. For digit experiments, we exploit the Adam solver [50] to optimize the objective with a learning rate of 0.0002 on the generators and 0.0001 on the discriminators, and first (second) moment estimates of 0.5 (0.999). We train for the first 10 epochs and then linearly decay the learning rate to zero over 20 epochs. Moving on, λ_{real} is set to 1 and λ_{fake} and λ_{rec} linearly increase from zero to half of λ_{real} because higher quality and diversified samples are guaranteed in the latter part of the training. We set the magnitude of cycle-consistency λ_{cyc} as 10 in MNIST and 0.1 in MNIST-M and SVHN. For real-world translations (Maps ↔ Aerial photo and Labels ↔ Photo), we exploit the hyper-parameter setting of CycleGAN [18]. By default, random crop is adopted as the stochastic data augmentation function at digit experiments. Color jitter is utilized for the photo domains in the real-world translations (Maps ↔ Aerial photo and Labels ↔ Photo), since it preserves the geometric structure of maps and segmentation labels, and tends to occur in the real world situations.

3) Evaluation details: For evaluation of all digit translations, we train revised LeNets [43] in MNIST, MNIST-M, and SVHN, which reach classification accuracies of 99.2%, 97.5%, and 91.0%, respectively. We fix these classifiers for the tests, experiment 5 times with different random seeds, and report classification accuracies (%) on fake samples. In Maps ↔ Aerial photo and Labels ↔ Photo translations, we measured the quality and level of detail for fake samples by mean square error (MSE) with ground-truth images.

C. Quantitative Analysis

Our proposed method is compared against CR-GAN [40] in Table I. We conduct experiments on CycleGAN [18] as a baseline, CR-CycleGAN, a CycleGAN with consistency regularization using real samples, and our ACCR-CycleGAN on MNIST ↔ MNIST-M and MNIST ↔ SVHN. ACCR-CycleGAN outperforms CycleGAN and CR-CycleGAN in digit translations. To confirm the statistical significance of the results, we conduct paired sample T-tests by comparing our proposal with the baseline. CR + CR-Fake, CR + CR-Rec, and ACCR demonstrate statistically significant differences at the level of significance $\alpha = 0.05$ for the translations of MNIST-M → MNIST and SVHN ↔ MNIST. The performance of the MNIST → MNIST-M seems saturated. Therefore, we next validate ACCR for the more complex tasks.

We conduct experiments on two real-world I2I tasks: Maps ↔ Aerial photo and Labels ↔ Photo on Cityscapes. The results are shown in Table II, Table III, Fig. 4, and Fig. 5. Although the performance of the three models is saturated at Photo → Maps (Table II), ACCR has the lowest (best) scores in all the other translations. Hence, ACCR is generic regularization for real-world problems.

To identify the sensitivity of the discriminators to the augmented data, we calculate the mean squared error (MSE) in the feature space between the real and augmented data as shown in Table IV. ACCR and CR decrease the distance to the baseline, and ACCR exerts a greater impact than the baseline and CR, especially in MNIST-M. Therefore, the discriminators of ACCR are robust to inputs' perturbation and give further semantics-aware feedback to the generators in adversarial training.

D. Qualitative Analysis

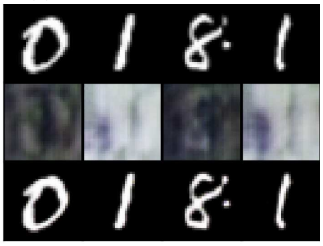
Unpaired I2I problems are innately ill-posed and thus could have infinite solutions. Here we show generated samples in

TABLE IV: **Feature distance between real and augmented samples.** We report the mean squared error (MSE) at the penultimate layer of a discriminator between test and augmented data.

Feature Distance (MSE)	MNIST test	MNIST-M test
CycleGAN	41.3 ± 13.4	46.4 ± 4.5
CR-CycleGAN	$\mathbf{35.1 \pm 6.1}$	35.4 ± 8.3
ACCR-CycleGAN (Ours)	36.9 ± 4.9	$\mathbf{29.6 \pm 8.8}$

(a) **MNIST → MNIST-M**

(b) **MNIST-M → MNIST**



(c) **MNIST → SVHN**

(d) **SVHN → MNIST**

Fig. 3: **Translation results of digit datasets.** In each translation result, Top: real samples, Middle: fake samples, Bottom: reconstructed samples. Reconstructed images at (b) MNIST-M → MNIST have a high variety of backgrounds due to the one-to-many projection of unpaired I2I translation. ACCR uses these samples for regularization. In the SVHN translations, the model often fails to generate reliable samples because of a large domain shift.

Fig. 3. It seems impossible to determine only one mapping from a grayscale to a color background in the translation from real to fake on MNIST → MNIST-M (Fig. 3a) and the reconstruction on MNIST-M → MNIST (Fig. 3b). However, we leverage the stochastic property as diversified samples of consistency regularization. Indeed, the meaningful regularization corresponds to either CR-Fake or CR-Rec. Therefore, the combination of both CR-Fake and CR-Rec (ACCR) is reasonable.

Furthermore, CR-GAN [40] reported that consistency regularization on generated samples (CR-Fake) does not always lead to improvements. By investigating this limitation, we found that standard GANs fail to produce recognizable samples at the initial and end steps because the GANs are unable to fully capture the data distribution and may cause mode collapse, respectively. However, unpaired I2I translation induces these problems to a lesser extent due to image conditioning and the constraint term $\mathcal{C}$. Hence, I2I models can preserve semantics even at the first and end epochs and this

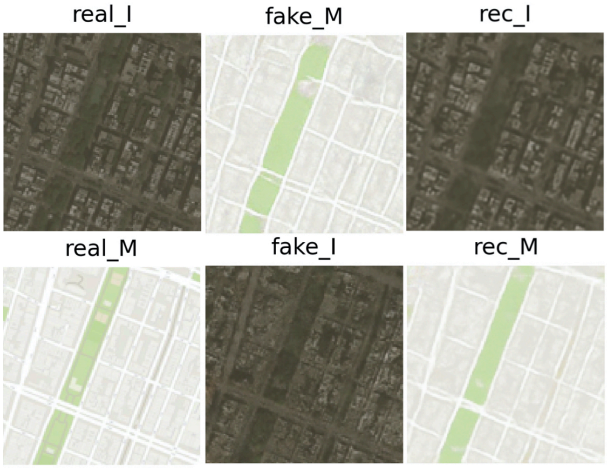


Fig. 4: **Translation results of Aerial Photo ↔ Maps.** The dataset contains paired images. The ground-truth image of a fake map (*fake_M*) is *real_M*, and vice versa.

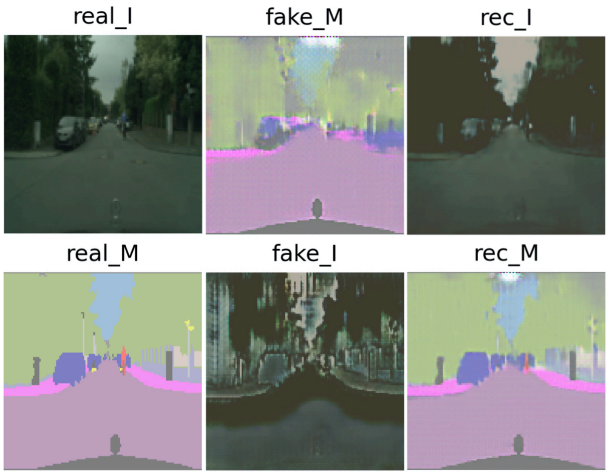


Fig. 5: **Translation results of Cityscapes Labels ↔ Photo.** The dataset contains paired images. The ground-truth image of a fake mask (*fake_M*) is *real_M*, and vice versa.

justifies using fake and reconstructed images for consistency regularization.

E. Ablation Studies

1) Comparison with CR-Fake, CR-Rec, and ACCR: To explore the effect of CR-Fake, CR-Rec, and ACCR, we compare each model on MNIST ↔ MNIST-M and MNIST ↔ SVHN as shown in Table I. As shown in Fig. 3, the generation fails in a fake image of (c) MNIST → SVHN and a reconstructed image of (d) SVHN → MNIST due to complexity of the SVHN domain. Thus, CR + CR-Fake in MNIST → SVHN and CR + CR-Rec in SVHN → MNIST do not improve so much. Although CR-Fake and CR-Rec are sometimes inferior to CR, but ACCR always achieves the best. The gains of ACCR are discussed in Section IV-D.

2) Comparison with other data augmentation: We compare several augmentation techniques in semantics-preserving ways

TABLE V: **Comparison of different types of data augmentation.** We experiment with 7 types of image augmentation: (1) randomly cropping images by a few pixels, (2) randomly rotating images by a few degrees, (3) combination of random cropping and random rotation, (4) applying cutout [48], (5) applying random erasing [49], (6) randomly changing the brightness, contrast, and saturation of the images, and (7) a combination of random cropping, rotation, and color jitter. All the models in the experiment are based on CycleGAN.

Data Augmentation	Method	M → MM	MM → M
(1) Random Crop	CR	97.7 ± 0.6	94.3 ± 0.5
	ACCR (Ours)	$\mathbf{98.0 \pm 0.5}$	$\mathbf{94.5 \pm 0.5}$
(2) Random Rotation	CR	97.9 ± 0.3	93.6 ± 1.3
	ACCR (Ours)	$\mathbf{98.1 \pm 0.1}$	$\mathbf{94.4 \pm 0.3}$
(3) Crop&Rotation	CR	98.1 ± 0.3	92.7 ± 1.1
	ACCR (Ours)	$\mathbf{98.2 \pm 0.2}$	$\mathbf{94.3 \pm 0.2}$
(4) Cutout	CR	97.2 ± 0.6	93.1 ± 1.3
	ACCR (Ours)	$\mathbf{97.4 \pm 0.4}$	$\mathbf{94.0 \pm 0.9}$
(5) Random Erasing	CR	96.6 ± 1.2	92.5 ± 1.5
	ACCR (Ours)	$\mathbf{97.3 \pm 1.0}$	$\mathbf{93.6 \pm 0.8}$
(6) Color Jitter	CR	97.2 ± 0.8	95.0 ± 0.3
	ACCR (Ours)	$\mathbf{97.7 \pm 0.3}$	$\mathbf{95.2 \pm 0.1}$
(7) Crop&Rotation&Jitter	CR	97.7 ± 0.6	$\mathbf{94.5 \pm 0.4}$
	ACCR (Ours)	$\mathbf{97.8 \pm 0.4}$	$\mathbf{94.5 \pm 0.3}$

TABLE VI: **Comparison of other cycle-consistent constraints.** [42] proposed Relaxed Cyclic Adversarial Learning (RCAL) as an extension of cycle-consistency [18], which enforces feature-wise cycle-consistency by using task specific models.

Model	M → S	S → M
RCAL	68.2 ± 10.9	47.5 ± 3.5
CR-RCAL	67.1 ± 11.4	55.7 ± 1.4
ACCR-RCAL (Ours)	$\mathbf{72.0 \pm 8.0}$	$\mathbf{57.8 \pm 9.4}$

(i.e., random crop, random rotation, cutout [48], random erasing [49], color jitter, and combinations thereof), as shown in Table V. ACCR tends to outperform the CR-GAN method in commonly used data augmentation techniques.

3) Comparison with other cycle-consist constraints: Table VI also shows results of an experiment with CycleGAN with Relaxed Cyclic Adversarial Learning (RCAL), which is a much looser constraint than having consistency in the pixel space, to verify our regularization with feature-level cycle-consistent constraints. RCAL is a naive extension of CycleGAN to the semantic-aware cycles using task-specific classifiers. ACCR-RCAL surpasses the RCAL baseline and CR-RCAL. Therefore, ACCR does not limit the choice of the constraint in pixel space. Rather, it is compatible with feature-wise cycle-consistent models.

4) Training speed: In terms of computational cost, we measure the actual update speeds of the discriminators for ACCR-CycleGAN with NVIDIA Tesla P100 in Table VII. ACCR marginally increases the forward pass of the discriminators compared with CR. ACCR-CycleGAN is around 1.5 times faster than CycleGAN with Gradient Penalty [37]. We observe that CycleGAN with Gradient Penalty sometimes degrades from the baseline as observed in [36], [40].

TABLE VII: **Training speed of discriminator updates.** We report the actual training speed of discriminator updates for CycleGAN on MNIST ↔ MNIST-M with NVIDIA Tesla P100. GP denotes CycleGAN with Gradient Penalty [37].

Method	W/O	GP	CR	ACCR (Ours)
Speed (step/s)	1871 ± 14	786 ± 8	1740 ± 25	1176 ± 28

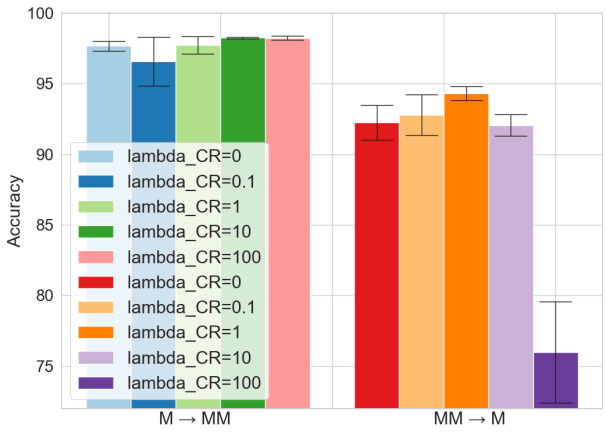


Fig. 6: **Different choices of λ_{Real}.** lambda_CR stands for the value of λ_{Real}. The bars of $\lambda_{\text{Real}} = 0$ denote the CycleGAN baseline and the others are CR-CycleGANs.

5) Lambda sensitivity: An experimental limitation of CR-CycleGAN and ACCR-CycleGAN lies in a lack of robustness for different lambda hyper-parameters. We conduct experiments for different hyper-parameters of CycleGAN and CR-CycleGAN where λ_{Real} ranges from 0.1 to 100. The results are shown in Fig 6. Due to the complexity of dual GANs optimization and original regularization terms of I2I models (e.g., consistency regularization), we cannot find the robustness in I2I translation as the significant results that CR-GAN reported [40], especially when λ_{Real} is large. For ACCR-CycleGAN, we fixed the hyper-parameters that recorded the best in CR-CycleGAN. By the observation that real samples are more reliable than fake and reconstructed data for consistency regularization, we set $\lambda_{\text{Real}} > \lambda_{\text{Fake}} = \lambda_{\text{Rec}}$ rather than $\lambda_{\text{Real}} = \lambda_{\text{Fake}} = \lambda_{\text{Rec}}$. This adjustment is also workable in real-world datasets.

V. Conclusion

In this paper, we propose a novel, simple, and effective training stabilizer ACCR in unpaired I2I translation. We demonstrate the effectiveness of adding consistency regularization using both fake and reconstructed data. In experiments, our ACCR outperforms the baseline and the CR-GAN method in several digit translations and real-world translations. Furthermore, the proposed method surpasses the CR-GAN in various situations where the cycle-consistent constraint and the data augmentation function are different.

Acknowledgments: We thank Takuma Yagi and Hiroaki Aizawa for comments on an earlier version of the manuscript. This work was partially supported by JSPS KAKENHI Grant Number 19K22865 and a hardware donation from Yu Darvish, a Japanese professional baseball player for the Chicago Cubs of Major League Baseball.

REFERENCES

[1] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context Encoders: Feature Learning by Inpainting. *Proc. CVPR*, 2014.
[2] C. Dong, C. C. Loy, K. He, and X. Tang. Image Super-Resolution Using Deep Convolutional Networks. *Proc. TPAMI*, 2016.
[3] J. Kim, J. K. Lee, and K. M. Lee. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. *Proc. CVPR*, 2016.
[4] R. Zhang, P. Isola, and A. A. Efros. Colorful Image Colorization. *Proc. ECCV*, 2016.
[5] L. A. Gatys, A. S. Ecker, and M. Bethge. Image Style Transfer Using Convolutional Neural Networks. *Proc. CVPR*, 2016.
[6] J. Hoffman, E. Tzeng, T. Park, J.Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. *Proc. ICML*, 2018.
[7] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim. Image to Image Translation for Domain Adaptation. *Proc. CVPR*, 2018.
[8] W. Hong, Z. Wang, M. Yang, and J. Yuan. Conditional Generative Adversarial Network for Structured Domain Adaptation, *Proc. CVPR*, 2018.
[9] P. Russo, F. M. Carlucci, T. Tommasi, and B. Caputo. From source to target and back: symmetric bi-directional adaptive GAN. *Proc. CVPR*, 2018.
[10] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. *Proc. CVPR*, 2018.
[11] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye, and J. Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. *Proc. CVPR*, 2018.
[12] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person Transfer GAN to Bridge Domain Gap for Person Re-Identification. *Proc. CVPR*, 2018.
[13] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. *Proc. CVPR*, 2019.
[14] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz. Joint Discriminative and Generative Learning for Person Re-identification. *Proc. CVPR*, 2019.
[15] I. J. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. *Proc. NeurIPS*, 2014.
[16] P. Isola, J.Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *Proc. CVPR*, 2017.
[17] J.Y Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward Multimodal Image-to-Image Translation. *Proc. NeurIPS* 2017.
[18] J.Y Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proc. ICCV*, 2017.
[19] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. *Proc. ICML*, 2017.
[20] M. Y. Liu, T. Breuel, and J. Kautz. Unsupervised Image-to-Image Translation Networks. *Proc. NeurIPS*, 2017.
[21] A. Almahairi, S. Rajeswar, A. Sordoni, P. Bachman, and A. Courville. Augmented CycleGAN: Learning Many-to-Many Mappings from Unpaired Data. *Proc. ICML*, 2018.
[22] X. Huang, M. Y. Liu, S. Belongie, and J. Kautz. Multimodal Unsupervised Image-to-Image Translation. *Proc. ECCV*, 2018.
[23] L. Ma, X. Jia, S. Georgoulis, T. Tuytelaars, and L. V. Gool. Exemplar Guided Unsupervised Image-to-Image Translation with Semantic Consistency. *Proc. ICLR*, 2019.
[24] H. Y. Lee, H. Y. Tseng, J. B. Huang, M. K. Singh, and M. H. Yang. Diverse Image-to-Image Translation via Disentangled Representations. *Proc. ECCV*, 2018.
[25] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv preprint arXiv:1607.08022*, 2016.
[26] V. Dumoulin, J. Shlens, and M. Kudlur. A Learned Representation for Artistic Style. *Proc. ICLR*, 2016.
[27] X. Huang, and S. Belongie. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. *Proc. ICCV*, 2017.
[28] H. Nam and H. E. Kim. Batch-Instance Normalization for Adaptively Style-Invariant Neural Networks. *Proc. NeurIPS*, 2018.
[29] T. Park, M. Y. Liu, T. C. Wang, and J. Y. Zhu. Semantic Image Synthesis with Spatially-Adaptive Normalization. *Proc. CVPR*, 2019.
[30] J. Kim, M. Kim, H. Kang, and K. Lee. U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation. *Proc. ICLR*, 2020.
[31] W. Wu, K. Cao, C. Li, C. Qian, and C. C. Loy. TransGaGa: Geometry-Aware Unsupervised Image-to-Image Translation. *Proc. CVPR*, 2019.
[32] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
[33] J. Ba, J. R. Kiros, and G. E. Hinton. Layer Normalization. *arXiv preprint arXiv:1607.06450*, 2016.
[34] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral Normalization for Generative Adversarial Networks. *Proc. ICLR*, 2018.
[35] X. Mao, Q. Li, H. Xie, R. Y.K. Lau, and Z. Wang, S. P. Smolley. Least Squares Generative Adversarial Networks. *Proc. ICCV*, 2017.
[36] K. Kurach, M. Lucic, X. Zhai, M. Michalski, and S. Gelly. A Large-Scale Study on Regularization and Normalization in GANs. *Proc. ICML*, 2019.
[37] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved Training of Wasserstein GANs. *Proc. NeurIPS*, 2016.
[38] N. Kodali, J. Abernethy, J. Hays, and Z. Kira. On Convergence and Stability of GANs. *arXiv preprint arXiv:1705.07215, 2017.*
[39] K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann. Stabilizing Training of Generative Adversarial Networks through Regularization. *Proc. NeurIPS*, 2017.
[40] H. Zhang, Z. Zhang, A. Odena, and H. Lee. Consistency Regularization for Generative Adversarial Networks. *Proc. ICLR*, 2020.
[41] A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *Proc. ICLR*, 2016.
[42] E. H. Asl, Y. Zhou, C. Xiong, and R. Socher. Augmented Cyclic Adversarial Learning for Low Resource Domain. Adaptation. *Proc. ICLR*, 2019.
[43] Y. LeCun, L. Bottou, Y. Bengio, and P. Haner. Gradient-Based Learning Applied to Document Recognition. *Proc. of the IEEE, 86(11):2278-2324, 1998. 5.*
[44] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-Adversarial Training of Neural Networks. *Proc. JMLR*, 2015.
[45] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour Detection and Hierarchical Image Segmentation. *Proc. PAMI*, 2011.
[46] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and Andrew Y. Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. *Proc. NeurIPSW*, 2011.
[47] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. *Proc. CVPR*, 2016.
[48] T. DeVries, and G. W. Taylor. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv preprint arXiv:1708.04552*, 2017.
[49] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random Erasing Data Augmentation. *arXiv preprint arXiv:1708.04896*, 2017.
[50] D. P. Kingma, and J. Ba. Adam: A Method for Stochastic Optimization. *Proc. ICLR*, 2015.
[51] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 2014.
[52] M. Sajjadi, M. Javanmardi, and T. Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Proc. NeurIPS*, 2016.
[53] S. Laine, and T. Aila. Temporal Ensembling for Semi-Supervised Learning. *Proc. ICLR*, 2017.
[54] W. Hu, T. Miyato, S. Tokui, E. Matsumoto, and M. Sugiyama. Learning Discrete Representations via Information Maximizing Self-Augmented Training. *Proc. ICML*, 2017.
[55] T. Miyato, S. Maeda, M. Koyama, and S. Ishii. Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. *Proc. TPAMI*, 2018.