RESEARCH-ARTICLE

Cross-Granularity Learning for Multi-Domain Image-to-Image Translation

HUIYUAN FU, Beijing University of Posts and Telecommunications, Beijing, Beijing, China

TING YU, Beijing University of Posts and Telecommunications, Beijing, Beijing, China

XIN WANG, Stony Brook University, Stony Brook, NY, United States

HUADONG MA, Beijing University of Posts and Telecommunications, Beijing, Beijing, China

Open Access Support provided by:

Beijing University of Posts and Telecommunications

Stony Brook University

Published: 12 October 2020

Citation in BibTeX format

MM '20: The 28th ACM International Conference on Multimedia
October 12 - 16, 2020
WA, Seattle, USA

Conference Sponsors:
SIGMM

Cross-Granularity Learning for Multi-Domain Image-to-Image Translation

Huiyuan Fu
Beijing University of Posts and Telecommunications
fhy@bupt.edu.cn

Ting Yu
Beijing University of Posts and Telecommunications
yuting97@bupt.edu.cn

Xin Wang
Stony Brook University
x.wang@stonybrook.edu

Huadong Ma
Beijing University of Posts and Telecommunications
mhd@bupt.edu.cn

ABSTRACT

Image translation across diverse domains has attracted more and more attention. Existing multi-domain image-to-image translation algorithms only learn the features of the complete image without considering specific features of local instances. To ensure the important instance to be more realistically translated, we propose a cross-granularity learning model for multi-domain image-to-image translation. We provide detailed procedures to capture the features of instances during the learning process, and specifically learn the relationship between style of the global image and the style of an instance on the image through the enforcing of the cross-granularity consistency. In our design, we only need one generator to perform the instance-aware multi-domain image translation. Our extensive experiments on several multi-domain image-to-image translation datasets show that our proposed method can achieve superior performance compared with the state-of-the-art approaches.

CCS CONCEPTS

• **Computing methodologies** → ***Machine learning algorithms***.

KEYWORDS

GAN; Image Translation; Image Generation

ACM Reference Format:
Huiyuan Fu, Ting Yu, Xin Wang, and Huadong Ma. 2020. Cross-Granularity Learning for Multi-Domain Image-to-Image Translation. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413656

1 INTRODUCTION

Image-to-image (I2I) translation aims at transferring an image from a source domain to a target one by learning a mapping between them, where the major representations of the source image is maintained. It has received significant attention in the computer vision field, because many vision and graphics issues can be formulated as I2I translation problems. Some example applications are image synthesis, image segmentation, image super-resolution, and image colorization.

Initial I2I translation is made between two domains. BicycleGAN [33] and Pix2Pix [13] are popular unified approaches that adopt the generative adversarial networks (GAN) [9] for image generation. It relies on the pairing of training data from two domains, which is often unavailable in the practical applications. As a result, unpaired I2I translation [18, 32] obtains increasing attentions and promising results. For example, CycleGAN [32] can translate an input image from the source domain to the target domain with the unpaired data by using the cycle consistency loss.

The two-domain image translation methods are limited in their scalability and robustness when handling more than two domains. Several recent efforts [3, 7, 27] have been made,

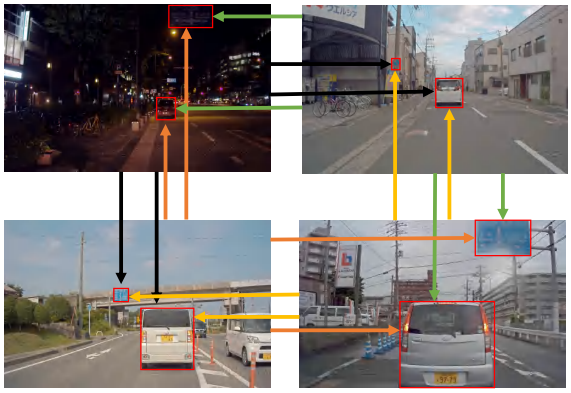

Figure 1: Illustration of the motivation of our cross-granularity learning method. The domain labels of images are Night, Cloudy, Sunny and Rainy in clockwise. Different colored arrows denote different global style codes. For multi-domain image-to-image translation, a wealth of relevant semantic information is contained between global and instance images. The style of instance images is associated with the global style but is not identical. Our method focuses on modeling the relationship between multi-domain global and instance images. Each pair of arrows is used to show that both the global style feature and the local instance content are sent to the instance generator to perform cross-granularity learning during training time.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413656

with StarGAN [7] as an example. They can perform I2I translation among multiple domains using only a single model, with the simultaneous training using multiple datasets from different domains. However, all existing multi-domain image translation approaches are based on the global learning of the entire input images, without considering the features of key objects (or instances) in the image to translate.

To increase the quality of multi-domain translation, we propose a cross-granularity multi-domain I2I translation framework that effectively capture the features of both the image to translate and the instances on the image, which allows the instance (or object) to be more accurately generated in the target domain. Besides incorporating the object features in all steps of training, we specifically learn the relationship of the image style and instance style, so that both the specific instances and the overall image are more realistically translated. The examples in Fig. 1 show the motivation of our cross-granularity learning method. A pair of arrows is used in each case to show that both the global style feature and the local instance content are sent to the instance generator to perform cross-granularity learning during the training time, while different colored arrows denote different global style codes.

We apply only one generator in our multi-domain I2I translation framework to perform instance-aware mapping in multi-domain translation. This will not only simplify our model structure, but also allow the instances and images to share some common features, so the translated instances can be more easily merged into the translated image. This simplify our model only need one model during inference time. To effectively capture the features of an important instance in the image to translate, we propose to enforce cross-granularity consistency during the training, that the relationship between the global style and object style can be effectively captured to ensure the translated image to be more realistic.

Our contributions of this paper can be summarized as follows:

- We propose a cross granularity framework for high quality multi-domain image-to-image translation.
- We provide detailed procedures to incorporate the instance features into the learning, and enforce the cross-granularity consistency to guide the learning of relationship between instance style and image style.
- We evaluate the proposed approach with extensive qualitative and quantitative experiments to demonstrate that images generated by our model are more realistic.

Extensive experimental analyses show the competing results compared with state-of-the-art approaches.

The remainder of this paper is organized as follows: Section 2 gives a brief overview of the related works. In Section 3, we introduce our proposed Multi-domain image-to-image translation framework in details. Section 4 and Section 5 describe our the experimental setup and experimental results respectively, followed by a conclusion in Section 6.

2 RELATED WORK

Generative adversarial network. Generative adversarial networks(GAN) [9] was introduced as a framework for learning a data distribution with an unsupervised manner. The GAN is composed of two parts: a generator and a discriminator. The generator is used to map random noise to the images and the discriminator is used to determine that the input produced by the generator is real or fake. The training of GAN always involves a two-player minmax game. The forgery technology of the generator will become more and more powerful after this game between them, and the identification technology of the discriminator will become more and more effective, too. Various derivative models based on GAN are proposed for model structure improvement. This will further expand the theory and application of GANs. In order to solve the problem of training gradient disappearing, the Wasserstein GAN (WGAN) [2] is presented. The InfoGAN [6] is proposed by combining the information theory and GAN. It brings the concept of mutual information to represent their correlation degree. Besides, GAN is widely used in the field of image generation tasks [1, 22, 30]. In [23], the authors improve the training process of GAN for representation learning. In [31], the authors apply GAN in image manipulation to constrain the edited images to stay close to the manifold of real images. In [17], the authors propose Perceptual GAN to improve the quality of small object detection by updating the generator network and discriminator network repeatedly to achieve the super-resolution of small objects.

Image-to-Image Translation. Image-to-image translation can be considered as an image generation task [7, 13, 18, 32, 33]. Abstractly, it is a mapping problem between different visual domains. These problems are usually solved by specific methods, and there is no general approach. In recent years, numerous methods have been proposed for image-to-image translation task based on deep learning technology, with a lot of applications in image processing, computer vision, and graphics [8, 13, 26]. With GANs, some studies have achieved significant progress. Specifically, CycleGAN [32] proposes a cycle-consistency constraint to tackle unpaired image-to-image translation. StarGAN [7] solves multi-domain image-to-image translation with a single model only. INIT [24] proposes to use bounding box based on disentangle representation methods. MUNIT [12] and DRIT [15] learn diverse image translation maps by disentangling images into content and style codes. DRIT++ [16] presents a multi-domain multi-modality image translation method base on DRIT. HomoInterpGAN [4] presents homomorphic latent space interpolation to solve the multidisciplinary multimodal image translation problems. HomoInterpGAN [4] presents homomorphic latent space interpolation to solve the multidisciplinary multimodal image translation problems.

3 FRAMEWORK

We propose to concurrently exploit the translation at the global image level and the local instance level for more accurate multi-domain image-to-image translation. In our framework, we apply only one generator to perform instance-aware mapping in multi-domain translation. This will not only simplify our model structure but also allow the instances and images share some common features, so the translated instances can be more easily merged into the translated image. For the convenience of presentation, we use the words object and instance interchangeably.

In this section, we first introduce our basic learning framework, and then the detailed designs to take into account different factors. Let g and o denote the global image and object regions respectively, our translation model consists of two encoders E_g and E_o, two decoders G_g and G_o, and two discriminators D_g and D_o.

Fig. 2 shows both the training and testing processes. In the testing step of Fig. 2(a) and Fig. 2(b), we only need one decoder and one encoder to translate the image, either with the style given or with the style to translate randomly

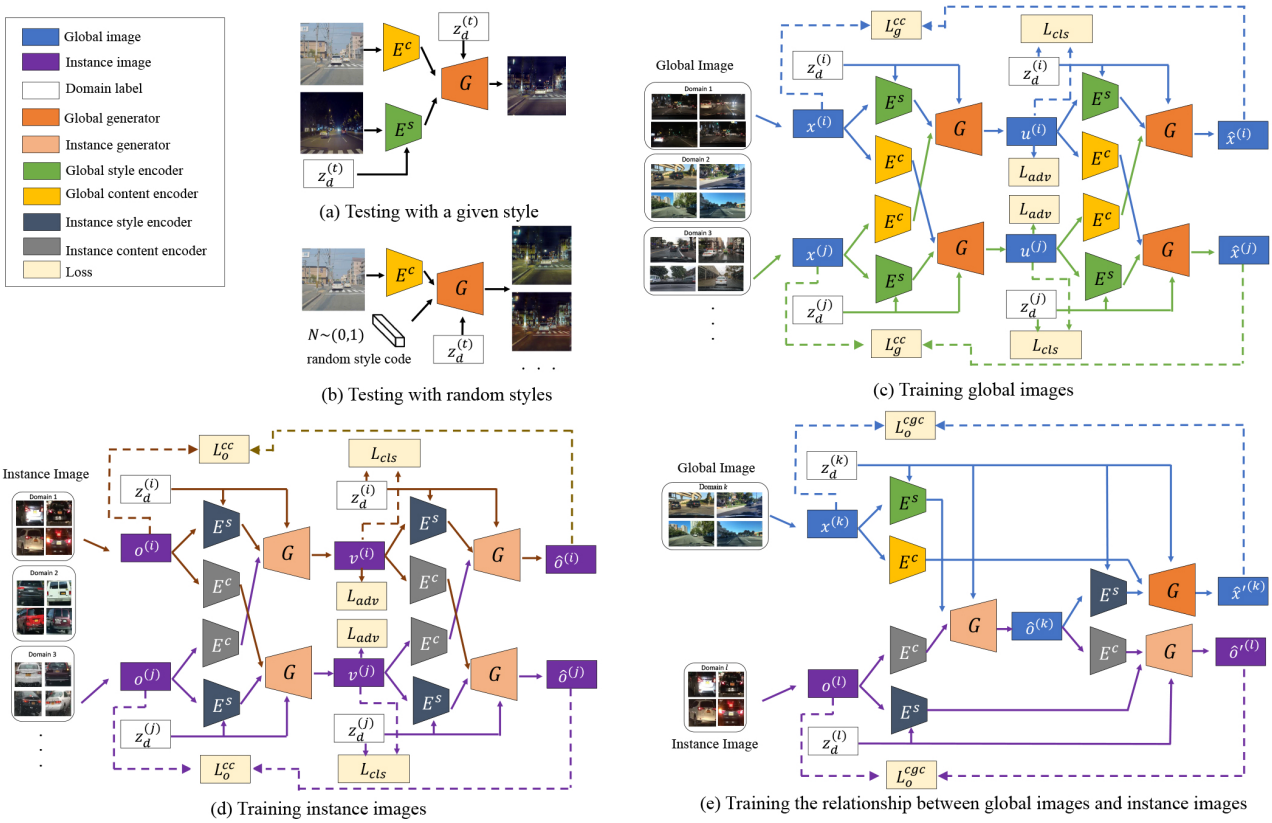

Figure 2: Proposed cross-granularity learning framework: (a) Generating an image according to the input style; (b) Generating multi-modal images following the random style codes sampled from a known distribution; (c) Applying the adversarial loss and cross-cycle consistency loss of the global image during the training; (d) Applying the adversarial loss and cross-cycle consistency loss of the instance in an image during the training; (e) Applying the cross-granularity loss during the training to capture the multi-domain global and instance relationship. The global images and instances cropped from the global images using the object coordinates are randomly selected during training.

generated. In Fig. 2(c) and Fig. 2(d), we show how the adversarial loss and cross-cycle consistency loss are considered in the training of global images and instance images. Fig. 2(e) applies the cross-granularity loss to regulate the learning of relationship between instance style and global image style.

3.1 Extraction of Features for Image Content and Image Style from Multiple Domains

In our framework, features are first extracted from the input global images and local instances and put into the content space and the style space respectively. Different from image translation between two domains, with only one generator, a special training scheme is needed for multi-domain translation. We exploit a random pair-wise training scheme. Given n domains$\{N_i\}_{i=1\sim n}$, we randomly select two domains $\{i, j\} \in N$, and sample two images $\{x^{(i)}, x^{(j)}\}$ with their domain labels represented as $\{z_d^{(i)}, z_d^{(j)}\}$. We also select pair of instances $\{o^{(i)}, o^{(j)}\}$ from the images of the two domains, and their domain labels are $\{z_d^{(i)}, z_d^{(j)}\}$.

For a global image, each encoder E_g can extract the input to form a content code c_g and a style code s_g, where $E_g = (E_g^c, E_g^s)$, $c_g = E_g^c(Img)$, $s_g = E_g^s(Img, z_d)$. Img denotes the input image, c_g and s_g are content and style features extracted from the global image, and z_d is the label of the domain. The process of extracting content and style for instances is similar to that for the images.

3.2 Multi-domain Image-to-image Translation

Given N-domain images, a straight-forward learning of mapping between every two domains will incur the complexity $N(N-1)$, which is very high when N is large. We choose to use only one generator and one discriminator instead. To ensure better learning and translation quality, we comprehensively consider loss components: adversarial loss, domain classification loss, reconstruction loss, cross cycle consistency loss and cross-granularity cycle consistency loss.

Adversarial Loss. The GAN loss is applied to enforce that the generated images and instances look to be realistic. We adopt the adversarial loss $\mathcal{L}_{adv}$, where the discriminators

D^g and D^o attempt to discriminate between real and synthetic images/instances in multiple domains. Our generators need three inputs: input images, random or reference style codes and target domain and instance labels to generate the images of the desired domains. An image could contain several instances, and allowing users to input the instances of interest will help increase the subjective quality of the image translated.

The discriminator consists of two parts, D_{src} to discriminate the input is real or fake and D_{cls} to determine the domain label of the input. The adversarial loss is:

$$\begin{aligned}\mathcal{L}_{adv} = \mathbb{E}_x[\log D_{src}(x)] + \\ \mathbb{E}_{x,s,z_d^t}[\log(1 - D_{src}(G(E_c(x), s, z_d^t)))]\end{aligned} \tag{1}$$

where x is the input global image or instance image, z_d^t is the target domain label, s is the style latent code. The generator G tries to minimize this objective, while the discriminator D tries to maximize it.

Domain Classification Loss. Our discriminators not only attempt to discriminate between real and synthetic images for global images and instance images(D_{src}), but also perform domain classification (D_{cls}). To achieve this goal, the objective functions of image generators and discriminators are independent. $\mathcal{L}_{cls}^r$ is used to optimize D for real images and $\mathcal{L}_{cls}^f$ is used to optimize G for fake images. The domain classification loss is:

$$\begin{aligned}\mathcal{L}_{cls}^r &= \mathbb{E}_{x,z_d^s}[-\log D_{cls}(z_d^s|x)] \\ \mathcal{L}_{cls}^f &= \mathbb{E}_{x,z_d^t}[-\log D_{cls}(z_d^t|G(E_c(x), s, z_d^t)]\end{aligned} \tag{2}$$

where z_d^s is the source domain label and z_d^t is the target domain label.

Reconstruction Loss. Given an image following a data distribution, the reconstruction loss encourages to reconstruct the image after encoding and decoding.

$$\begin{aligned}\hat{Img} = G_g(E_g^c(Img),\ E_g^s(Img, z_d^{Img}),\ z_d^{Img}) \\ \hat{o} = G_o(E_o^c(o),\ E_o^s(o, z_d^o),\ z_d^o)\end{aligned} \tag{3}$$

where Img are global images and o are objects in the images, z_d^Img and z_d^o are global domain and instance domain labels. Given a latent code sampled from the latent distribution for images or instances, we can also reconstruct the content and style codes of the object or image after encoding the corresponding generated ones. Following show the example of reconstructing the object codes:

$$\begin{aligned}\hat{c}_o = E_o^c(G_o(c_o,\ s_g,\ z_d^g)) \\ \hat{s}_o = E_o^s(G_o(c_o,\ s_g,\ z_d^g), z_d^o)\end{aligned} \tag{4}$$

where c_o and s_g are instance content and global image style features, z_d^o and z_d^g are labels for the instance and global image. $\hat{s}_o$ represents the reconstructed instance-level style features. We formulate the reconstruction loss as:

$$\mathcal{L}_{recon}^k = \mathbb{E}_{k\sim p(k)}\left[\left\|\hat{k} - k\right\|_1\right] \tag{5}$$

where k can be Img, o, c or s. $p(k)$ denotes the distribution of data k.

Cross-cycle consistency loss. We adjust the cross-cycle consistency loss [15] to enforce that the translated images or instances are consistent with the input.

As shown in Fig. 2 (c), given a pair of images $\{x^{(i)}, x^{(j)}\}$ from two domains i and j, we would like to check if we could reconstruct the images by using the features from their translated images. We first encode the two images into $\{c_g^{x^{(i)}}, s_g^{x^{(i)}}\}$ and $\{c_g^{x^{(j)}}, s_g^{x^{(j)}}\}$, and then perform forward and backward translations, with the styles swapped in each translation. In the first translation, we generate $\{u^{(i)}, u^{(j)}\}$ and the second generates $\{\hat{x}^{(i)}, \hat{x}^{(j)}\}$, where $u^{(i)}$ and $\hat{x}^{(i)}$ belong to domain i, and $u^{(j)}$ and $\hat{x}^{(j)}$ belong to domain j. The translation on the instance branch shown in Fig. 2 (d) is similar to that of the global branch. After forward and backward translation, we should reconstruct the original global images $\{\hat{x}^{(i)}, \hat{x}^{(j)}\}$ and instance images $\{\hat{o}^{(i)}, \hat{o}^{(j)}\}$ and enforce the cross-cycle consistency through the loss constraint:

$$\begin{aligned}\mathcal{L}_g^{cc} = \mathbb{E}_{x^{(i)},x^{(j)}}\left[\left\|\hat{x}^{(i)} - x^{(i)}\right\|_1 + \left\|\hat{x}^{(j)} - x^{(j)}\right\|_1\right] \\ \mathcal{L}_o^{cc} = \mathbb{E}_{o^{(i)},o^{(j)}}\left[\left\|\hat{o}^{(i)} - o^{(i)}\right\|_1 + \left\|\hat{o}^{(j)} - o^{(j)}\right\|_1\right]\end{aligned} \tag{6}$$

where the forward translation gets $u^{(i)} = G_g(c_g^{x^{(j)}}, s_g^{x^{(i)}}, z_d^{(i)})$, and $v^{(i)} = G_o(c_o^{o^{(j)}}, s_o^{o^{(i)}}, z_d^{(i)})$. In the backward translation, we have $\hat{x}^{(i)} = G_g(c_g^{u^{(j)}}, s_g^{u^{(i)}}, z_d^{(i)})$, $\hat{o}^{(i)} = G_o(c_o^{v^{(j)}}, s_o^{v^{(i)}}, z_d^{(i)})$. Among these, $\{u^{(i)}, u^{(j)}\}$ and $\{v^{(i)}, v^{(j)}\}$ are images generated from the forward translation of the global image and the local instance respectively. $\{\hat{x}^{(j)}, \hat{o}^{(j)}\}$ is generated similar to $\{\hat{x}^{(i)}, \hat{o}^{(i)}\}$ and $\{u^{(j)}, v^{(j)}\}$ is generated similar to $\{u^{(i)}, v^{(i)}\}$.

Cross-granularity consistency loss(CGC). To effectively capture the features of an important instance in the image to translate, we propose a novel cross-granularity consistency loss to guide the cross-granularity learning. As the style of the image has an impact on the object style, this loss is applied to regularize the training, that is the relationship between the global style and object style can be effectively captured to ensure the translated image to be more realistic.

The process of finding the cross-granularity consistency loss is illustrated in Fig. 2(e). Given a pair of global image and instance image $\{x^{(k)}, o^{(l)}\}$, we encode them into $\{c_g^{x^{(k)}}, s_g^{x^{(k)}}\}$ and $\{c_o^{o^{(l)}}, s_o^{o^{(j)}}\}$. We feed the object content and the global style into the global generator to produce the instance image $\hat{o}^{(k)}$ with the global style in the forward translation. We further perform the backward translation by feeding the original global image content with the style of $\hat{o}^{(k)}$ into the global generator and feeding the original instance image style with the content of $\hat{o}^{(k)}$ into the object generator, so we can compare the regenerated $\hat{x}\prime^{(k)}$ and $\hat{o}\prime^{(l)}$ with the inputs $x^{(k)}$ and $o^{(l)}$. We formulate the cross-granularity consistency loss as follows:

$$\begin{aligned}\mathcal{L}_g^{cgc} = \mathbb{E}_{x^{(k)}}\left[\left\|\hat{x}\prime^{(k)} - x^{(k)}\right\|_1\right] \\ \mathcal{L}_o^{cgc} = \mathbb{E}_{o^{(l)}}\left[\left\|\hat{o}\prime^{(l)} - o^{(l)}\right\|_1\right]\end{aligned} \tag{7}$$

where $\hat{x}\prime^{(k)} = G_g(c_g^{x^{(k)}}, s_o^{\hat{o}^{(k)}}, z_d^{(k)})$, $\hat{o}\prime^{(l)} = G_o(c_o^{\hat{o}^{(k)}}, s_o^{o^{(l)}}, z_d^{(l)})$, $\hat{o}^{(k)}$ is the generated instance after the forward translation.

Reconstructing the global image and all instances using forward translation and backward translation would be time

consuming and difficult, as an image may include a lot of instances. We choose to randomly select an image and instance pair from the global image set and the instance set instead of a global image with all its instances. During the training, k and l are each randomly select from the domains night, cloud, sunny or rainy.
The full objective function of our framework is:

$$\begin{aligned}\mathcal{L}_D &= -\alpha\mathcal{L}_{adv} + \beta\mathcal{L}^r_{cls} \\ \mathcal{L}_G &= \alpha\mathcal{L}_{adv} + \beta\mathcal{L}^f_{cls} + \lambda_{rec}\Big(\mathcal{L}^{Img}_{rec} + \mathcal{L}^o_{rec} \\ &\quad + \mathcal{L}^{cc}_g + \mathcal{L}^{cgc}_g + \mathcal{L}^{cc}_o + \mathcal{L}^{cgc}_o\Big) + \mathcal{L}^s_{rec} + \mathcal{L}^c_{rec}\end{aligned} \tag{8}$$

where α, β and λ_{rec} are the hyper-parameters. Our goal is to optimize G and D by minimizing $\mathcal{L}_D$ and $\mathcal{L}_G$ respectively.

4 EXPERIMENTAL SETUP

We implement our proposed learning framework with the following considerations.

Improving the GAN Training. The training process of generative adversarial network is generally difficult. In order to stabilize the training process and generate higher quality images, we replace Eq. (1) with Wasserstein GAN objective with the gradient penalty [1, 10] defined as:

$$\begin{aligned}\mathcal{L}_{adv} = &\mathbb{E}_x[D_{src}(x)] - \mathbb{E}_{x,s,z}[D_{src}(G(E_c(x), s, z))] \\ &- \lambda_{gp}\,\mathbb{E}_{\hat{x}}[(\|\nabla_{\hat{x}} D_{src}(\hat{x})\|_2 - 1)^2]\end{aligned} \tag{9}$$

where $\hat{x}$ is point sampled along straight lines between points in the faked image distribution and the real image distribution, s is the style code and z is the target domain label. The training process of WGAN-GP relies on hyper-parameters λ_{gp} heavily, we adopt an adaptive weighting scheme [5] for WGAN-GP to facilitate the training process.

Training setting. Our model consists of content encoder E^c, style encoder E^s, multilayer perceptron M, generator G and discriminator D. The content encoder E^ccontains 3 convolutional layers and 4 residual blocks. Each convolutional layer uses 4×4 filters with stride 2, except that the first convolutional layer uses 7×7 filters with stride 1. Each residual block includes two convolution layers with 3×3 filters and stride 1. The style encoder consists of 5 convolutional layers, a global pooling layer and a fully connected layer with 8 units. Except that the first convolutional layer uses 7×7 filters with stride 1, the rest of convolutional layers use 4×4 filters with stride 2. M consists of three fully connected layers with 256, 256 and 4096 units. The generator is composed of 4 residual blocks and two 2×2 nearest-neighbor upsampling layers followed by a 5×5 convolutional layer and stride 1. Each residual block includes two convolution layers with 3×3 filters and stride 1. The discriminator is adapted from PatchGAN [12] which includes 6 convolutional layers with 4×4 filters and stride 2, a convolutional layer with 1x1 filters and stride 1 and a convolutional layer with 4×4 filters and stride 1. The last two convolutional layers are used to calculate the adversarial loss and classification loss respectively. We apply Instance Normalization (IN) [12] to the content encoder and Adaptive Instance Normalization (AdaIN) [12] to the generator. We use ReLU activations in the generator and Leaky ReLU with slope 0.2 in the discriminator.

We implement our model with PyTorch [21]. We initialize the weights of Encoders and Decoders according to the Kaiming normal distribution, and initialize the weights of discriminator with a Gaussian distribution. We optimize the model using Adam [20] with batch size 2 for both INIT and BDD100k datasets. The learning rate is set to 0.0001, and is decreased by a half for each 100,000 iterations. In all the experiments, we use the following hyper-parameters: $\alpha = 1$, $\beta = 1$, $\lambda_{rec} = 10$.

Table 1: Grouped Domain of BDD100K. We divide the datatset into four domains according to the weather style sunny, night, cloudy and rainy. We also count the number and labels of each domain.

Dim	Weather Style	Total	Domain Labels
1	Night	22884	Clear weather, Night time
2	Sunny	12451	Clear weather, Day time
3	Cloudy	4262	Cloudy weather, Day time
4	Rainy	2521	Rainy weather, Day time

Table 2: Inception Score(IS) (higher, better) on INIT Dataset.

Method	Night	Cloudy	Rainy	Average
MUNIT	1.03	1.46	1.26	1.25
DRIT++	1.15	1.26	1.34	1.25
INIT	1.35	1.29	1.44	1.36
HomoInterpGAN	1.07	1.06	1.06	1.06
StarGAN	1.66	1.29	1.44	1.47
Ours w/ $\mathcal{L}^{cgc}$	**1.70**	**1.65**	**1.52**	**1.62**
Ours w/o $\mathcal{L}^{cgc}$	1.62	1.26	1.16	1.35

4.1 Datasets

We take INIT dataset to evaluate our models. In order to verify the generality of our method, we rebuild BDD100K dataset to perform image-to-image translation tasks.

INIT Dataset [24]. It contains two resolutions, 1280×1920 and 3000×4000. Due to the limitations of GPU memory and the long running time, we only use 1280×1920 resolution ones to form a set of 38836 images for training and 7434 for testing. The weather domains include sunny, night, cloudy and rainy (overcast weather with wet road). we further divide the datatset into four domains according to the weather style sunny, night, cloudy and rainy. Our rebuilt dataset is composed of 35548 images for training and 6570 images for testing.

BDD100K Dataset [28]. It consists of 100K images with labels. As shown in Table 1, we further divide the datatset into four domains according to the weather style sunny, night, cloudy and rainy. Our rebuilt dataset is composed of 35548 images for training and 6570 images for testing. we further divide the datatset into four domains according to the weather style sunny, night, cloudy and rainy. Our rebuilt dataset is composed of 35548 images for training and 6570 images for testing. We further divide the datatset into four domains according to the weather style sunny, night, cloudy and rainy.

4.2 Evaluation Metrics

To validate our approach, we consider the following three metrics.

Inception Scors (IS). We use the Inception Score (IS) [23] to evaluate the diversity of generated images. A higher inception score means the generated images have higher diversity.

Table 3: FID (lower, better) on INIT Dataset.

Method	Night	Cloudy	Rainy	Average
MUNIT	101.09	30.76	62.26	64.70
DRIT++	112.92	25.75	54.30	64.32
INIT	91.87	49.20	59.10	66.72
HomoInterpGAN	111.37	60.09	98.61	90.02
StarGAN	100.71	15.81	52.66	56.39
Ours w/ $\mathcal{L}^{cgc}$	**76.34**	**14.26**	**47.43**	**46.01**
Ours w/o $\mathcal{L}^{cgc}$	93.87	18.10	52.75	54.91

Table 4: Inception Score(IS) (higher, better) on BDD100K Dataset.

Method	Night	Cloudy	Rainy	Average
MUNIT	1.22	**1.34**	1.12	1.23
DRIT++	1.14	1.14	1.14	1.14
INIT	1.16	1.27	1.20	1.21
HomoInterpGAN	1.01	1.01	1.01	1.01
StarGAN	1.17	1.15	1.25	1.19
Ours w/ $\mathcal{L}^{cgc}$	**1.50**	1.32	**1.26**	**1.36**
Ours w/o $\mathcal{L}^{cgc}$	1.43	1.22	1.11	1.25

Table 5: FID (lower, better) on BDD100K Dataset.

Method	Night	Cloudy	Rainy	Average
MUNIT	59.98	48.93	81.56	63.49
DRIT++	103.19	48.60	50.31	67.37
INIT	43.63	119.20	52.84	71.89
HomoInterpGAN	123.58	65.99	76.91	88.83
StarGAN	73.90	41.19	54.52	56.54
Ours w/ $\mathcal{L}^{cgc}$	**40.14**	**34.53**	**34.05**	**36.24**
Ours w/o $\mathcal{L}^{cgc}$	48.89	44.95	68.93	54.26

We fine-tune our Inception-V3 model on four domain labels of our INIT dataset and BDD100K dataset, and follow MUNIT using 100 input images and 100 samples per input to calculate inception scores.

Learned Perceptual Image Patch Similarity (LPIPS). In order to measure the diversity of translation, we calculate the Learned Perceptual Image Patch Similarity (LPIPS) [29], which has been demonstrated to correlate well with human perceptual similarity. In our setting, we randomly select 19 pairs of translation outputs via different random style codes from 100 input images of our test set, which amounts to 1900 pairs in total according to [12]. We also use the ImageNet-pretrained AlexNet [14] to extract features. If the generator has more diversity, the LPIPS value will be large.

The Frechet Inception Distance (FID). The Frechet Inception Distance (FID) [11] has become an important metric to evaluate the quality of images translated with the generative adversarial models. The real image dataset and generated image dataset are mapped into a feature space by the pre-trained InceptionV3 [25] model to calculate the distance at the feature level.

Table 6: Model size of baselines and our method on INIT and BDD100K dataset.

Method	Model Size
MUNIT	179MB × 4
DRIT++	620MB
INIT	179MB × 4
HomoInterpGAN	220MB
StarGAN	205MB
Ours	**119MB**

5 EXPERIMENTAL RESULTS

First, we introduce several baseline approaches for image-to-image translation task. Then, we evaluate the model on INIT and BDD100K datasets. Finally, we compare our model with baselines case by case. In addition, we also analyze the performance of image translation on the instance level and verify our model's ability to disentangle global style and instance style codes.

5.1 Baselines

We adopt MUNIT [12] and INIT [24], two recent methods proposed for image translation between two domains, as our baseline models. To apply them for image translation over N domains, we train these two models for every pair of domains. We also compare our model with multi-domain translation methods StarGAN [7], HomoInterpGAN [4] and DRIT++ [16].

MUNIT [12]. This method disentangles image representation into content codes and style codes. It combines random style codes sampled from style codes with content codes to perform unpaired multimodal image translation. MUNIT also swaps content-style pairs to perform latent reconstruction.

INIT [24]. INIT performs image translation between two domains by transferring styles of global images and its corresponding instance area. INIT requires an encoder and decoder for each domain and uses cyclic reconstruction by associating the style codes of pairs of instances and global images.

StarGAN [7]. For multi-domain translation, StarGAN only needs a single model and target domain labels to perform $N-$domains$N \geq 2$ image translation task. StarGAN can also perform multiple datasets with different domains by using mask vectors.

HomoInterpGAN [4]. HomoInterpGAN performs unpaired multi-domain multi-modal image translation by homomorphic latent space interpolation. It can generate intermediate images between two domains by selecting the paths properly.

DRIT++ [16]. DRIT++ is a multi-domain version of DRIT. The main idea of DRIT is to divide the latent space into content codes and attribute codes which is similar to MUNIT. DRIT++ uses domain labels for multi-domain training. The discriminator will perform domain classification, in addition to determining the generated image is real or fake.

5.2 Results

Qualitative evaluation. Fig. 3 shows the weather transfer results on INIT dataset and BDD100K dataset. As shown in Fig. 3, the generated cloudy image of StarGAN is almost the same with the input. Paying special attention to instances in the images, our model can better learn the important

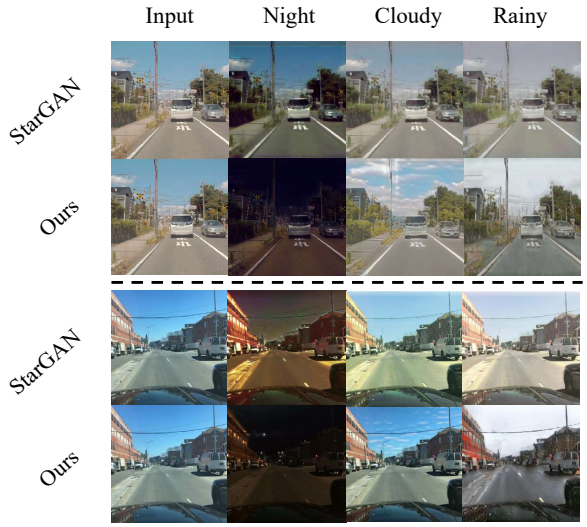

Figure 3: Multi-domain Image-to-Image translation results of our method compared with StarGAN. The generated images of StarGAN suffer from only transferring global style, while ours overcome this problem and make the generated images more natural and indistinguishable from the real images.

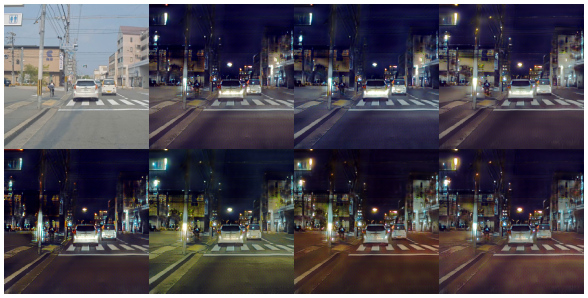

Figure 4: The multi-modal results of our method on INIT dataset. The top-left is the input image and others are the generated output images by using randomly sampled style codes.

features to flexibly generate images for different style translation. Fig. 4 shows the multi-modal results on sunny to night translation task.

Quantitative evaluation. The quantitative evaluation results of INIT dataset are in Table 2, Table 3 and Table 7, and the results of BDD100K dataset are in Table 4, Table 5 and Table 8, respectively. "w/ $\mathcal{L}^{cgc}$" and "w/o $\mathcal{L}^{cgc}$" denote that we train our model with and without considering the cross-granularity consistency loss respectively. For IS, our method achieve higher average scores for three different tasks on both datasets, which proves that the generated images of our method are more diverse and photorealism. For FID scores, our method is significantly lower than baselines on both two datasets indicating that images generated by ours are more similar to the real target domain images. For LPIPS, our model without CGC loss surpasses others and model with CGC loss is slightly lower than the best baseline method. CGC loss makes FID scores higher and LPIPS lower which indicates CGC loss improves the quality of generated images at the cost of possibly resulting in a decrease in the diversity of generated images. Table 6 shows that the model sizes of different methods and our method surpasses all baselines since our network architecture takes advantage of MUNIT and DRIT++ and only uses one generator and one discriminator.

Figure 5: Impact of considering the proposed cross-granularity consistency loss. First row: input sunny images. Second and third rows: input images are translated from sunny to the night images using our method without and with considering the cross-granularity consistency loss respectively.

Figure 6: Qualitative results of our methods compared with StarGAN on instance level Day to Night translation. The first row is the input objecs. The second and third rows is translated object images from StarGAN and ours respectively.

Ablation Studies. We perform ablation studies to assess the effectiveness of the visual generation mechanism. We evaluate the performance of our approach with and without the proposed cross-granularity consistency loss. As shown in Table 2 and Table 4, the average inception scores are improved by 20.5% on INIT dataset and 8.5% on BDD100K dataset thanks to our consideration of cross-granularity consistency loss $\mathcal{L}^{cgc}$. Table 3 and Table 5 show the cross-granularity consistency loss $\mathcal{L}^{cgc}$ brings our model FID score improvement of 16.2% on INIT dataset and 33.2% on BDD100K dataset. In the qualitative evaluations of Fig. 5, both the global night images and instance night images generated by our model are more realistic and natural, which verifies the effectiveness of incorporating our proposed cross-granularity consistency loss.

Table 7: LPIPS (higher, better) on INIT Dataset.

Method	Night	Cloudy	Rainy	Average
MUNIT	0.25	0.21	0.14	0.20
DRIT++	0.06	0.05	0.05	0.05
INIT	0.28	0.27	0.12	0.22
HomoInterpGAN	0.09	0.09	0.10	0.09
StarGAN	0.07	0.02	0.04	0.04
Ours w/ $\mathcal{L}^{cgc}$	0.31	0.19	0.18	0.23
Ours w/o $\mathcal{L}^{cgc}$	**0.39**	**0.26**	**0.23**	**0.29**

Table 8: LPIPS (higher, better) on BDD100k Dataset.

Method	Night	Cloudy	Rainy	Average
MUNIT	0.30	0.19	**0.28**	0.26
DRIT++	0.06	0.05	0.05	0.05
INIT	0.25	0.27	0.13	0.21
HomoInterpGAN	0.18	0.21	0.27	0.22
StarGAN	0.17	0.29	0.06	0.17
Ours w/ $\mathcal{L}^{cgc}$	0.26	**0.23**	0.22	0.24
Ours w/o $\mathcal{L}^{cgc}$	**0.33**	**0.23**	0.26	**0.27**

5.3 Analysis

Day→ Night Instance-level Comparison. We compare our model with StarGAN on the instance-level in Fig. 6. We find that our model can generate more diverse instance and more details. It appears that StarGAN just changes the lightness of instance and fails to change the style of instances.

Qualitative Comparison. We compare the results of multi-domain image translation for the same input image case by case using the BDD100K dataset as shown in Fig. 7. Since the data distribution of BDD100K is very complex, some baselines such as MUNIT, INIT will fail to perform image translation. They may bring small artifacts around the object area. The other multi-domain baselines such as DRIT ++, HGAN and StarGAN may fall into mode collapse. Obviously, our proposed method is more realistic.

Disentangle the global and instance style. We follow [24] which randomly sample 100 images and instances in the test set of each domain and use t-SNE tools [19] to visualize global and instance style codes in multiple domains to verify that the global and instance style are distinguishable enough to disentangle. The results are shown in Fig. 8. There is a remarkable margin (In this paper, the margin means the distance between the local and global features) between global and instance style codes extracted from the same domain images and instances, which demonstrates the effectiveness of our model.

6 CONCLUSION

In this paper, we propose a cross-granularity learning model that can enable more effective multi-domain image-to-image translation. Besides, the detailed procedures to incorporate the instance features into the learning have been provided. To guide the learning of relationship between the image style and the instance style, we present the cross-granularity consistency. Extensive experimental analysis on some multi-domain image-to-image translation datasets show the competing results compared with state-of-the-art approaches.

Figure 7: Case-by-case multi-domain image translation comparison on BDD100K dataset. For each method we present multi-domain outputs for the same input.

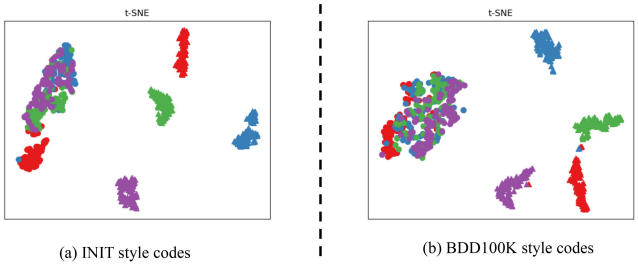

Figure 8: T-SNE tools [19] are used to visualize global and instance features in the two-dimensional space. Each high-dimensional data point is given a location on a two-dimensional map. (a) The style codes extracted by our model on INIT dataset. (b) The style codes extracted by our model on BDD100K dataset.

7 ACKNOWLEDGEMENT

This work is supported in part by the Beijing Natural Science Foundation (L191004), the National Natural Science Foundation of China under No.1720106007 and No.61872047, the National Key R&D Program of China under No.2017YFB1003000, the NSFC-Guangdong Joint Found under No.U1501254 and the 111 Project (B18008).

REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*. 214–223.

[2] Bottou L Arjovsky M, Chintala S. 2017. Wasserstein GAN. arXiv preprint arXiv: 1701.07875.

[3] Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. 2019. Learning Disentangled Semantic Representation for Domain Adaptation. *International Joint Conference on Artificial Intelligence* (2019), 2060–2066.

[4] Ying-Cong Chen, Xiaogang Xu, Zhuotao Tian, and Jiaya Jia. 2019. Homomorphic Latent Space Interpolation for Unpaired Image-To-Image Translation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2408–2416.

[5] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. 2018. Deep Photo Enhancer: Unpaired Learning for Image Enhancement From Photographs With GANs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[6] Houthooft R SchulmanJ Sutskever I Abbeel P Chen X, Duan Y. 2016. InfoGAN: interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the Neural Information Processing Systems*. 2172–2180.

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2017. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. *arXiv preprint arXiv:1711.09020* (2017), 8789–8797.

[8] David Eigen and Rob Fergus. 2015. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture. In *ICCV*. 2650–2658.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *Advances in neural information processing systems*. 5767–5777.

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[12] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal Unsupervised Image-to-image Translation. In *The European Conference on Computer Vision (ECCV)*.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *CVPR*. 1125–1134.

[14] Hinton G Krizhevsky A, Sutskever I. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of Advances in neural information processing systems*.

[15] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2018. Diverse Image-to-Image Translation via Disentangled Representations. In *The European Conference on Computer Vision (ECCV)*. 35–51.

[16] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. 2019. DRIT++: Diverse Image-to-Image Translation via Disentangled Representations. *arXiv preprint arXiv:1905.01270* (2019).

[17] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. 2017. Perceptual generative adversarial networks for small object detection. In *Conference on Computer Vision and Pattern Recognition*. 1222–1230.

[18] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised Image-to-Image Translation Networks. In *NIPS*. 700–708.

[19] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[20] Kingma D P and Ba J. 2014. Modular generative adversarial networks. In *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980.

[21] Chintala S et al Paszke A, Gross S. 2017. Automatic differentiation in pytorch. (2017).

[22] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

[23] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in neural information processing systems*. 2234–2242.

[24] Zhiqiang Shen, Mingyang Huang, Jianping Shi, Xiangyang Xue, and Thomas Huang. 2019. Towards Instance-level Image-to-Image Translation. In *Conference on Computer Vision and Pattern Recognition*.

[25] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2818–2826.

[26] Xiaolong Wang and Abhinav Gupta. 2016. Generative Image Modeling Using Style and Structure Adversarial Networks. In *ECCV*. 318–335.

[27] Xuewen Yang, Dongliang Xie, and Xin Wang. 2018. Crossing-Domain Generative Adversarial Networks for Unsupervised Multi-Domain Image-to-Image Translation. 374–382.

[28] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. 2018. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687* (2018).

[29] Efros A A et al Zhang R, Isola P. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 586–695.

[30] Junbo Zhao, Michael Mathieu, and Yann LeCun. 2016. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126* (2016).

[31] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. 2016. Generative visual manipulation on the natural image manifold. In *The European Conference on Computer Vision (ECCV)*.

[32] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision (ICCV)*. 2223–2232.

[33] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. 2017. Toward Multimodal Image-To-Image Translation. In *NIPS*. 465–476.