Published: 16 May 2024
Online AM: 04 April 2024
Accepted: 20 March 2024
Revised: 19 February 2024
Received: 09 December 2023

Citation in BibTeX format

RESEARCH-ARTICLE

# Multi-Domain Image-to-Image Translation with Cross-Granularity Contrastive Learning

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

JIN LIU, 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

ACM Transactions on Multimedia Computing, Communications and Applications, Volume 20, Issue 7 (July 2024)
https://doi.org/10.1145/3656048
EISSN: 1551-6865

# Multi-Domain Image-to-Image Translation with Cross-Granularity Contrastive Learning

HUIYUAN FU, Beijing University of Posts and Telecommunications, Beijing, China
JIN LIU, Beijing University of Posts and Telecommunications, Beijing, China
TING YU, Beijing University of Posts and Telecommunications, Beijing, China
XIN WANG, Stony Brook University, Stony Brook, United States
HUADONG MA, Beijing University of Posts and Telecommunications, Beijing, China

The objective of multi-domain image-to-image translation is to learn the mapping from a source domain to a target domain in multiple image domains while preserving the content representation of the source domain. Despite the importance and recent efforts, most previous studies disregard the large style discrepancy between images and instances in various domains, or fail to capture instance details and boundaries properly, resulting in poor translation results for rich scenes. To address these problems, we present an effective architecture for multi-domain image-to-image translation that only requires one generator. Specifically, we provide detailed procedures for capturing the features of instances throughout the learning process, as well as learning the relationship between the style of the global image and that of a local instance in the image by enforcing the cross-granularity consistency. In order to capture local details within the content space, we employ a dual contrastive learning strategy that operates at both the instance and patch levels. Extensive studies on different multi-domain image-to-image translation datasets reveal that our proposed method outperforms state-of-the-art approaches.

CCS Concepts: • **Computing methodologies** → *Artificial intelligence*; *Computer vision;*

Additional Key Words and Phrases: Image-to-image translation, GAN, cross-granularity, contrastive learning, multi-domain

**ACM Reference Format:**
Huiyuan Fu, Jin Liu, Ting Yu, Xin Wang, and Huadong Ma. 2024. Multi-Domain Image-to-Image Translation with Cross-Granularity Contrastive Learning. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 7, Article 228 (May 2024), 21 pages. https://doi.org/10.1145/3656048

This work is supported in part by the NSFC under No.62272059, the National Key R&D Program of China under No.2023YFF0904800, the Beijing Nova Program under No.20230484406, the Innovation Research Group Project of NSFC (61921003), and the 111 Project (B18008).

Authors' addresses: H. Fu, J. Liu, T. Yu, and H. Ma (Corresponding author), State Key Laboratory of Networking and Switching Technology, Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing, Beijing, China, 100876; e-mails: fhy@bupt.edu.cn, ljin@bupt.edu.cn, 2436904681@qq.com, mhd@bupt.edu.cn; X. Wang, Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, New York, United States, 11794; e-mail: x.wang@stonybrook.edu.

ACM 1551-6865/2024/05-ART228
https://doi.org/10.1145/3656048

## 1 INTRODUCTION

**Image-to-Image (I2I)** translation has been exploited to transfer external styles while preserving the internal contents of images and has drawn lots of attention in recent years. I2I [17, 19, 24, 30, 40, 41, 45, 60, 61, 64, 65] is commonly used in image inpainting, image/video colorization, image dehazing, style transfer, super-resolution, and so on. Meanwhile, as an effective data augmentation form, I2I brings convenience to downstream intelligent monitoring scene tasks [35–38].

Initial I2I translation methods concentrate on two domains using paired training data. BicycleGAN [69] and Pix2Pix [21] are two well-known unified approaches that adopt the **generative adversarial networks (GAN)** [14] for image generation, while it is hard to find paired images in practical applications. As a consequence, unpaired I2I translation approaches [12, 39, 66, 68] are gaining attraction and showing promise. For example, CycleGAN [68] is designed to train two-domain translations using cycle-consistency losses and produces impressive results. When dealing with multi-domain translation, these two-domain image translation methods are limited in scalability and robustness. Several recent studies [3, 10, 57] make remarkable progress in tackling these challenges. For instance, StarGAN [10] generates images by using a single model to train multi-domain datasets and referencing images from various domains. However, these methods primarily carry out the global style transfer for the whole input image, neglecting the translation of critical instance (or object) features in the image.

Due to the relevance of spatial texture in the image or the singleness of instance, methods that concentrate on transferring styles or attributes of the entire image work well in content-simple scenarios, but the absence of instance-level translation makes them suffer in content-rich cases (e.g., car, pedestrian, and traffic light in the road traffic scene). On the other hand, compared to the extraction of global scene texture, the content details of local instance objects are difficult to preserve, and the style of an instance is close to but not identical to the general style of the entire image. Several methods [2, 51] learn the relationship between the image style and instance style. In the instance-aware image-to-image translation approach (INIT) [51], the input image is treated as a disentangled representation of shared content space and domain-specific style space, and the fine-grained objects are incorporated into all steps of training to translate both the overall image and specific instances realistically. However, INIT requires training a generator for each pair of domains and then employs cyclic reconstruction loss by integrating the content and style codes of the global image or instance, which results in network and storage space redundancy.

To solve the aforementioned problems, as shown in Figure 1, we propose a cross-granularity learning strategy that effectively captures the features of both the image and the corresponding instance (or object). Besides incorporating the instance features into all steps of training, we specifically learn the relationship between the image style and instance style by enforcing cross-granularity consistency so that both the specific instances and the overall image are realistically translated. Simultaneously, we introduce a dual contrastive learning module that preserves the local content details of global images at the instance and patch levels. Our method trains an encoder to capture the useful representation by learning the relationship between different granularities of positive and negative pairs. Therefore, the generator learns the fine-grained domain-irrelevant contents (e.g., the shape of a car or the texture of the license plate of a car) to produce sharper and more distinct instances. Meanwhile, our multi-domain I2I translation framework just requires one generator to perform the instance-aware mapping. This not only simplifies our model structure but also allows instances and global images to share certain common features so that it is easier to incorporate the generated instances into the translated image.

We extend our previous work [13] from the following three aspects: First, we design a contrastive learning module that preserves the rich content from multiple instances while accurately generating the instances in the target domain at instance and patch levels. Second, we introduce

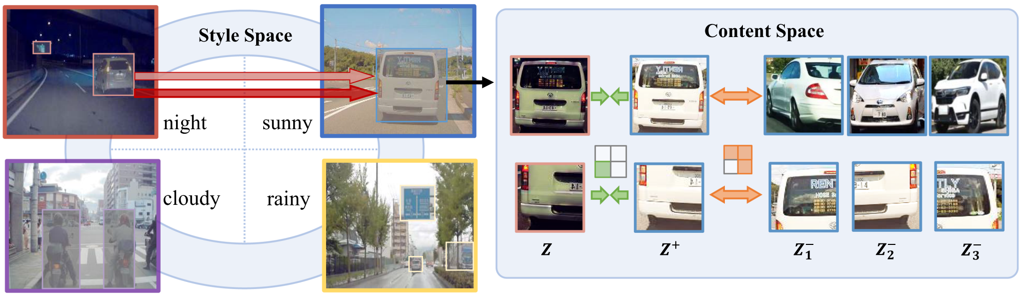


Fig. 1. Illustration of the motivation of our cross-granularity contrastive learning method for multi-domain I2I translation. Various colors signify corresponding image or instance styles. In the style space, the style of instance is related to but not identical to the equivalent global style. For more precise style transferring, both the global style feature and the local instance content are sent to the instance generator. Images and instances in the content space include a wealth of structured semantics. For retaining local details, a generated instance or patch of the output $Z$ should attract its corresponding input instance or patch $Z^+$ while repelling the others $Z^-$. During training, we apply cross-granularity contrastive learning to model the relationship among multi-domain image-level, instance-level, and patch-level objectives.

multilayer dual contrastive loss to encourage content preservation in unpaired I2I translation. Third, we validate extensive methods for both qualitative and quantitative evaluations. Moreover, we provide further insight into the trained models via subjective results, such as visualization.

In general, the main contributions of this work are summarized as follows:

- We propose a cross-granularity contrastive learning framework to perform high-quality multi-domain image-to-image translation.
- We introduce specific procedures for incorporating the instance features into the learning and enforcing cross-granularity consistency in order to guide the learning of the relationship between instance style and image style.
- We design a multilayer instance-level and patch-level contrastive learning module for preserving local details of the original images or instances.
- Extensive qualitative and quantitative experiments demonstrate the superiority of our proposed approach on standard benchmarks.

The remainder of this paper is organized as follows: Section 2 gives a brief overview of the related work. In Section 3, we present our proposed multi-domain I2I translation framework with cross-granularity contrastive learning in detail. Section 4 describes the experimental setup, results, ablation studies, and analysis, respectively, which is followed by a conclusion in Section 5.

## 2 RELATED WORK

### 2.1 Generative Adversarial Network

Generative adversarial networks (GAN) [14] have achieved outstanding results and are being employed in a variety of computer vision applications [63, 64] recently. GAN is presented as a framework for learning a data distribution in an unconditional or conditional manner using a generator and a discriminator. The generator attempts to generate fake but plausible images, while the discriminator tries to distinguish if the generated image is true or false. The GAN training process can be considered a two-player, zero-sum game. In this game process, the ability of the generator to forge real samples will become stronger, and the judgment of the discriminator will be more accurate as well. A Nash equilibrium solution is eventually found between the two players.

Various derivative models [1, 16, 34] based on GAN are proposed to improve the model structure, and enormous progress has been made in the theoretical research and application of GAN. For instance, the **Wasserstein GAN (WGAN)** [1] replaces the Jensen-Shannon divergence loss function with the Wasserstein distance, also known as the Earth Mover distance, to solve the gradient vanishing and mode collapse problems, and WGAN-GP [16] proposes a gradient penalty to enforce the Lipschitz constraint in WGAN. The Perceptual GAN [34] aims to achieve high-resolution generation of tiny objects by repeatedly updating the generator and discriminator. **Deep convolutional GAN (DCGAN)** [48] refers to a set of fully convolutional architecture guidelines for GAN, such as stride convolutions and batch normalization to obtain more stable training and better performance. Besides, GAN is widely used in the field of image generation tasks [1, 48]. In [50], the authors improve the training process of GAN for representation learning. In [67], the authors apply GAN in image manipulation to constrain the edited images to stay close to the manifold of real images. In [34], 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.

### 2.2 Image-to-Image Translation

Image-to-image translation can be considered an image generation task [9, 10, 21, 39, 55, 68, 69]. Abstractly, it is a mapping problem between different visual domains. I2I translation studies can be divided into two categories: unsupervised and supervised. Since it is difficult to obtain paired data in a specific scene, we focus on the challenge of unsupervised I2I translation with unpaired data. In recent years, numerous methods have been proposed based on deep learning technology, with a lot of applications in image processing, computer vision, and graphics [11, 21, 53]. Specifically, CycleGAN [68], DualGAN [58] and DiscoGAN [26] are proposed to train the deterministic one-to-one mapping models with cycle-consistency constraints. However, these methods are limited to two-domain mapping translation. Recently great efforts have been made to explore deeper into multi-domain translation. For example, StarGANs [9, 10] learn a mapping between different visual domains while satisfying the diversity of generated images and the scalability across multiple domains. Nevertheless, they only apply to situations involving face attribute translation in structured single occurrences. MUNIT [20] and INIT [51] are two typical methods closely related to our research. MUNIT trains diverse image translation maps by disentangled domain-invariant content spaces and domain-specific style codes. INIT introduces bounding boxes to achieve instance-aware, disentangled representations. TSIT [23] presents a two-stream framework based on multi-scale feature normalization from coarse to fine that adapts to various tasks and achieves compelling results. Wang et al. [54] propose a shared knowledge module to learn the common information among multi-domain pairs.

Our model achieves outstanding translation performance and succinct network architecture in multi-instance traffic scenarios with large semantic discrepancies. Compared to the above models, the primary differences are as follows: 1) We introduce cross-granularity learning to specifically learn the relationship of image style and instance style rather than only learning global disentangled representations; 2) We allow the instances and images to share common features by applying only one generator rather than achieving multiple mappings via redundant models.

### 2.3 Contrastive Learning

Self-supervised learning methods [42, 43] have sprung up in recent years. Contrastive learning [5, 6, 15] is an effective self-supervised learning method commonly used in discriminative tasks. In general, it is formulated by simultaneously maximizing the consistency between multiple transformed views of the same image (e.g., clipping, flipping, color transformation) while minimizing it between transformed views of different images in feature space. The encoder learns image-level representation rather than pixel-level generation via contrastive training. For I2I translation tasks, CUT [46] is the first work that utilizes patch-wise contrastive estimation on unpaired translation tasks by learning the relationship between an input image and the corresponding generated image. Jeong et al. [22] introduce a set of key-value storage structures with read/update operations to record the style changes of categories while enhancing the discrimination ability of memory with the proposed feature contrastive loss. Nevertheless, CUT only applies the global style to the whole image without considering the style differences between individual instances and the background or within an instance. [22] can be computationally expensive to update the representations in the memory bank as the representations get outdated quickly in a few passes.

We further carry forward the idea of contrastive learning under multi-instance traffic scenarios with complex semantics in this work. Specifically, we conduct contrastive learning at both the complex instance-level and the fine patch-level within the instance. We make our contrastive training process extensive by considering the content differences between and within individual instances simultaneously.

## 3 THE PROPOSED METHOD

In this section, we first introduce the basic components and functions of our cross-granularity contrastive learning approach, and then describe our detailed designs, along with a comprehensive loss function that considers different types of loss to ensure high image translation quality in an end-to-end manner. Our framework attempts to strike a good trade-off between model efficiency and image translation quality.

For the convenience of presentation, we use the words *object* and *instance* interchangeably. Furthermore, let $g$ and $o$ denote the global image and object regions, respectively.

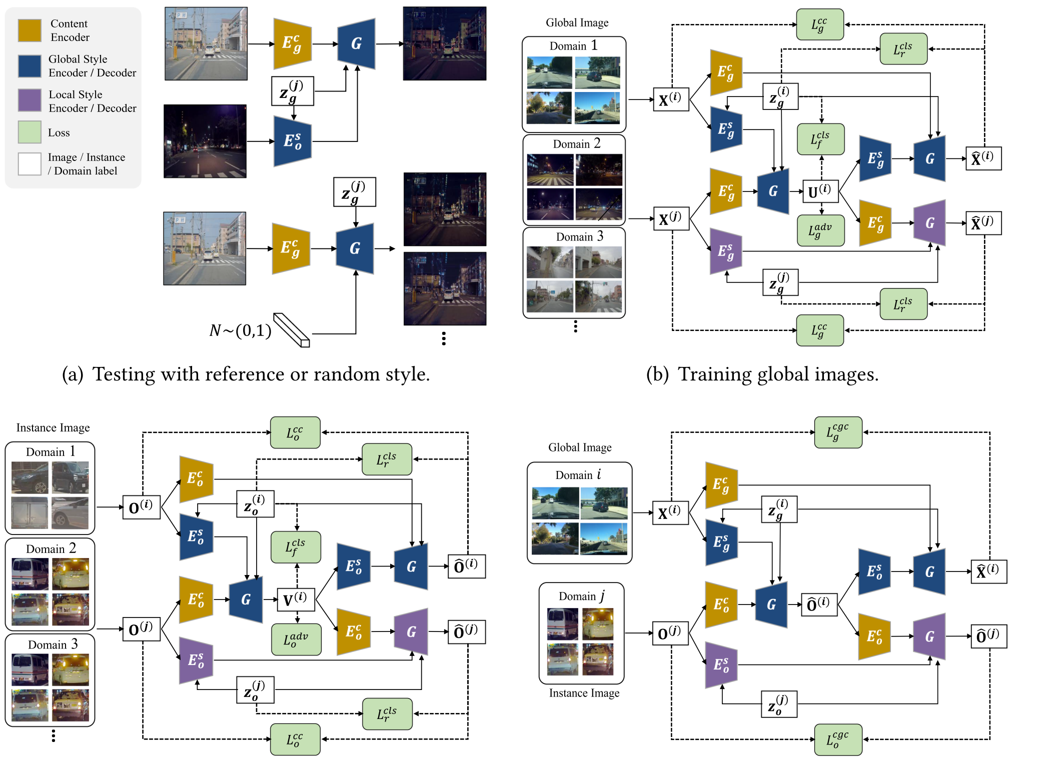
### 3.1 Extraction of Features for Image Content and Image Style from Multiple Domains

When referring to the I2I translation task for $N$ domains, we need to learn $N{\cdot}(N-1)$ mappings simultaneously. We choose to employ simply one generator and one discriminator instead. We exploit a random pair-wise training scheme for multi-domain translation.

During the training process, the Encoder first extracts features from the input global image and local instances respectively. The extracted features include content and style codes. Given $n$ domains $\{N_i\}_{i=1\sim n}$, we sample two images $\{x^{(i)}, x^{(j)}\}$ randomly selected from two domains $\{i, j\} \in N$ that are called source domain and target domain, respectively, and their corresponding domain labels are represented as $\{z_g^{(i)}, z_g^{(j)}\}$. We also select a pair of random instances $\{o^{(i)}, o^{(j)}\}$ from the random images of two domains, and their corresponding domain labels are $\{z_o^{(i)}, z_o^{(j)}\}$. The encoder $E_g$ first extracts the input image to form a content latent code $c_g$ and a style code $s_g$, where $E_g = (E_g^c, E_g^s)$, $c_g = E_g^c(img)$, $s_g = E_g^s(img, z_g)$. $img$ denotes the input source image, and $z_g$ is the domain label corresponding to $img$. Then the decoder $G_g$ utilizes the above content latent code $c_g$ and style code $s_g$ to generate the fake target image $img$ eventually, where $img' = G_g(c_g, s_g, z_g')$. $s_g$ denotes the style code from the given image or the prior distribution $\mathrm{q}(s_g) \sim \mathcal{N}(0, 1)$, and $z_g'$ denotes the target domain label. The process for the local instance is similar to the above.

### 3.2 Multilayer Dual Contrastive Learning

As shown in Figure 3, we deploy a multilayer dual contrastive learning strategy to facilitate the efficiency of the training process. We draw instances or patches internally from the input image as negatives and the corresponding generated image as positives, rather than externally from other images in the dataset.



(a) Testing with reference or random style.

(b) Training global images.

(c) Training instance images.

(d) Training the relationship between global images and instance images.

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

Taking global images as examples, we name the instance in the generated image as *query*, and the instance in the corresponding original input image as *key*. If the spatial locations of $k$-$q$ pairs are the same, the *key* is considered as *positives*, otherwise as *negatives*. Our goal is to maximize the mutual information between *query* and *positives* while minimizing it between *query* and *negatives* [44]. We deal with it as a $(M+1)$-way binary classification problem, where $M$ negative instances are sampled from the same input image at different locations. Firstly, we feed the input and output image pairs $\{x^{(i)}, x^{(j)}\}$ into the encoder $E_g^c$ to obtain the $\{c_g^{(i)}, c_g^{(j)}\}$ of the content latent space. Then we map them to the K-dimensional vector space through the fully connected network $H_g$ to obtain $\{q, k\}$. Finally, we calculate the contrastive loss of the embeddings by using the following noise contrastive estimation function [44]:

$$\mathcal{L}^{NCE}\left(q, k^+, k^-\right) = -\log\left(\frac{\exp(sim(q,k^+)/\tau)}{\exp(sim(q,k^+)/\tau + \sum_{i=1}^{M}\exp(sim(q,k^-)/\tau)}\right), \tag{1}$$

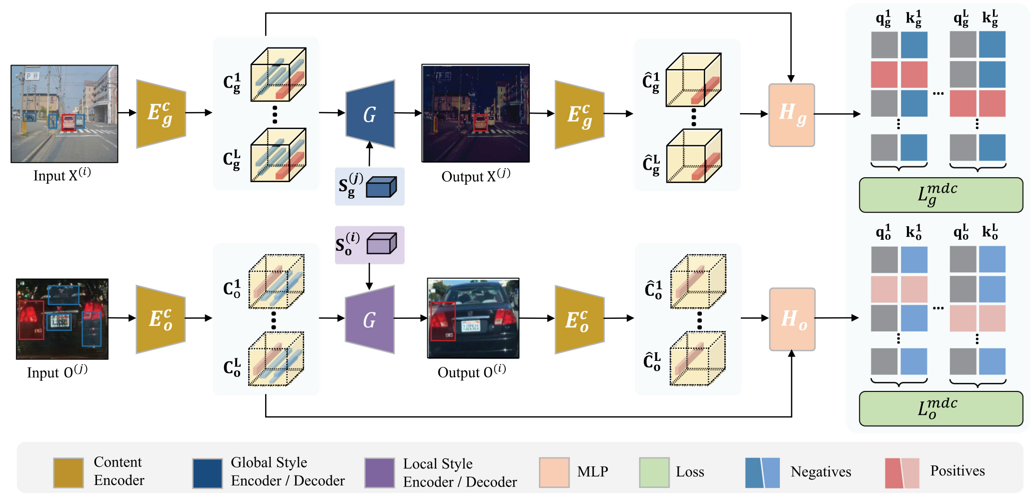


Fig. 3. Training with instance-level and patch-level contrastive learning schemes. The content encoders extract contents $C_g^L, C_o^L$ from the input image $X^{(i)}$ and object $O^{(j)}$ individually. Multilayer(1...L) residual structure of the content encoder outputs $\{(C_g^1, C_o^1), \ldots, (C_g^L, C_o^L)\}$. The multi-instance contents of output and input images pair $\{X^i, X^j\}$ are mapped into key-query items $\{(q_g^1, k_g^1), \ldots, (q_g^L, k_g^L)\}$ by $H_g$ where the red $k$ represents a **positive** item while the other blue ones represent the **negative** items of the red $q$. And vice versa for the patches of the instances pair $\{O^i, O^j\}$. The network aims to project the specific instances or patches of the output and input pairs to a shared embedding space.

where $q, k^+, k^-$ denote the embeddings of query, positives and negatives, respectively, which can be formulated as $q = H_g(E_g^c(x^{(j)})), k = H_g(E_g^c(x^{(i)}))$. The cosine similarity $sim(q, k)$ is represented as $sim(q, k) = q^T \cdot k/\|q\|\|k\|$. The temperature coefficient $\tau$ is a hyperparameter and set to 0.07 by default.

### 3.3 Multi-domain Image-to-image Translation

To guarantee a superior learning process and multi-domain I2I translation quality, our loss metric is designed to comprehensively contain the original adversarial loss, domain classification loss, reconstruction loss, cycle consistency loss, and our novel cross-granularity consistency loss and dual contrastive loss.

**Adversarial and Domain classification Losses.** To ensure that the generated images and instances are similar in distribution to the training samples from the target domain, the adversarial loss $\mathcal{L}^{adv}$ and domain classification loss $\mathcal{L}^{cls}$ are adopted in our I2I translation situation. Our generators require three inputs: input images, random or reference style codes, and target domain 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 improve the subjective quality of the image translated. The above losses are formulated as:

$$\mathcal{L}^{adv} = \mathbb{E}_x[\log D_{src}(x)] + \mathbb{E}_{x,s,z_d^t}\left[\log\left(1 - D_{src}\left(G\left(E_c(x), s, z_g^{(j)}\right)\right)\right)\right], \tag{2}$$

$$\begin{aligned} \mathcal{L}_r^{cls} &= \mathbb{E}_{x,z_d^s}\left[-\log D_{cls}\left(z_g^{(i)} \mid x\right)\right], \\ \mathcal{L}_f^{cls} &= \mathbb{E}_{x,z_d^s}\left[-\log D_{cls}\left(z_g^{(j)} \mid G\left(E_c(x), s, z_g^{(i)}\right)\right)\right], \end{aligned} \tag{3}$$

where $D_{src}$ and $D_{cls}$ discriminate real or synthetic images and perform domain classification, respectively. The generator $G$ tries to minimize this objective, while the discriminator $D$ tries to maximize it. The objective functions of image generators and discriminators are independent. $\mathcal{L}_r^{cls}$ is used to optimize $D$ for real images, and $\mathcal{L}_f^{cls}$ is used to optimize $G$ for fake images.

**Reconstruction and Cross-cycle Consistency Losses.** We utilize reconstruction and cycle consistency losses within a domain and between domains to facilitate the training that the reconstructed images or instances are consistent with the origin inputs [32]. We adopt the $\mathcal{L}_1$ norm to regularize the losses as follows:

$$\begin{aligned}\mathcal{L}_k^{rc} &= \mathbb{E}_{k\sim p(k)}[\|\hat{k}-k\|_1],\\ \mathcal{L}_g^{cc} &= \mathbb{E}_{x^{(i)},x^{(j)}}[\|\hat{x}^{(i)}-x^{(i)}\|_1+\|\hat{x}^{(j)}-x^{(j)}\|_1],\\ \mathcal{L}_o^{cc} &= \mathbb{E}_{o^{(i)},o^{(j)}}[\|\hat{o}^{(i)}-o^{(i)}\|_1+\|\hat{o}^{(j)}-o^{(j)}\|_1],\end{aligned} \tag{4}$$

The reconstruction loss encourages reconstructing images, objects, contents, and style codes. $i$ and $j$ represent source and target domains, respectively; $\hat{k}$ can be $\widehat{img}, \hat{o}, \hat{c}$ or $\hat{s}$, that comes from the reconstruction process of $k \rightarrow G(k, z^{(i)}) \rightarrow G(G(k, z^{(i)}), z^{(i)}) \rightarrow \hat{k}$, and $p(k)$ denotes the distribution of data $k$. $\hat{x}^{(i)}$ comes from the cycle translation process of $x^{(i)} \rightarrow G(E^c(x^{(i)}), E^s(x^{(i)}, z^{(i)}), z^{(j)}) \rightarrow u^{(j)} \rightarrow G(E^c(y^{(j)}), E^s(y^{(j)}, z^{(j)}), z^{(i)}) \rightarrow \hat{x}^{(i)}$. The same can be obtained for $\hat{o}^{(i)}$, and $u^{(j)}$ is the generated image from the source domain to the target domain.

**Cross-Granularity Consistency Loss (CGC).** We introduce a novel cross-granularity consistency loss in order to prevent the features between the global image style and the local instance style from being disturbed. We apply this loss to effectively capture the relationship of the style between the global and local, which ensures the generated image is more realistic. As illustrated in Figure 2(d), we first encode a pair of global image and instance $\{x^{(i)}, o^{(j)}\}$ as input and encode them into $\{c_g^x, s_g^{x^{(i)}}\}$ and $\{c_o^o, s_o^{o^{(j)}}\}$. The global style $s_g^{x^{(i)}}$ and instance content are assembled to the object generator to produce the generated instance $\hat{o}^{(i)}$ in the forward translation. We further perform the backward translation by encoding $\hat{o}^{(i)}$ into $\{c_o^{\hat{o}}, s_o^{\hat{o}^{(i)}}\}$, then input the original global image content $c_g^x$ with the style of $\hat{o}^{(i)}$ called $s_o^{\hat{o}^{(i)}}$ to the global generator while feeding the content of $\hat{o}^{(i)}$ called $c_o^{\hat{o}}$ with the original instance style $s_o^{o^{(j)}}$ into the object generator, and obtain the reconstructed $\{\hat{x}^{(i)}, \hat{o}^{(j)}\}$ finally. We express the cross-granularity consistency loss as:

$$\mathcal{L}^{cgc} = \mathbb{E}_{x^{(i)},o^{(j)}}[\|\hat{x}^{(i)}-x^{(i)}\|_1+\|\hat{o}^{(j)}-o^{(j)}\|_1], \tag{5}$$

where $\hat{x}^{(i)} = G_g(c_g^x, s_o^{\hat{o}^{(i)}}, z_g^{(i)})$, $\hat{o}^{(i)} = G_o(c_o^{\hat{o}}, s_o^{o^{(j)}}, z_g^{(j)})$.

For different granularities, both global images and local instances can achieve translation tasks independently. The only difference is the input size. For cross-granularity, based on the content features of the instance, we hope to achieve better translation results with the help of global style features that are richer than the instance style. It can be regarded as a data (style feature) augmentation behavior to make the generator with stronger generalization and adaptive performance. More specifically, the global style vector with richer information is integrated into the instance content features through general **adaptive instance normalization (AdaIN)**, making the instance translation results better.

**Multilayer Dual Contrastive Loss (MDC).** Our model aims at key and complex instances (e.g., car, pedestrian) within the image and patches (e.g., license plate, lighting) within the instance in the content space. We apply the contrastive learning strategy to the immediate latent contents of the multilayer residual structure (see Figure 4), $l \in \{1, 2, \ldots, L\}$ in the generator. It encourages the

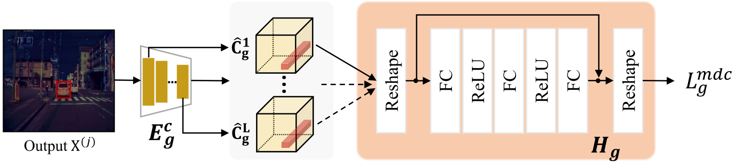


Fig. 4. The details of multilayer residual structure.

generated global images or instances to better preserve the content details of the original input. The formula is expressed as follows:

$$\mathcal{L}^{mdc} = \mathbb{E}_{x\sim X,o\sim O} \sum_{l=1}^{L} \left[ \sum_{s=1}^{s} \mathcal{L}_s^{NCE} \left( q_s, k_s^+, k_{S\backslash s}^- \right) + \sum_{p=1}^{P} \mathcal{L}_p^{NCE} \left( q_p, k_p^+, k_{P\backslash p}^- \right) \right], \tag{6}$$

where $q_s = H_g^l(E_g^c(x^{(j)}))$ and $q_p = H_o^l(E_o^c(o^{(j)}))$ denote the $s$-th query embedding of instance and the $p$-th query embedding of patch, respectively. Similarly, $k_s^+ = H_g^l(E_g^c(x^{(i)}))$ and $k_p^+ = H_o^l(E_o^c(o^{(i)}))$ denote the $s$-th positive embedding of instance and the $p$-th positive embedding of patch. $k_{S\backslash s}^-$ and $k_{P\backslash p}^-$ are negative embeddings and formulated similarly to the above. We set the number of layers L to 4, and $S = 256, P = 64$ by default.

Contrastive learning has been an effective tool in unsupervised visual representation learning. Image-to-image translation is a typical unsupervised task, and we hope to explore the potential of contrastive learning. We find it pertinent to use it in a multilayer, patchwise fashion. Specifically, as $I(X, Y) = H(X) - H(X|Y)$, cycle consistency is substantiated to be the upper bound of conditional entropy $H(X|Y)$ [33]. Therefore, the conditional entropy should be minimized. The output is encouraged to be more dependent on the input. As entropy $H(X)$ is a constant that is independent of the generator, the objective is maximizing mutual information $I(X, Y)$. Meanwhile, inspired by InfoGAN [4], we use NCE loss to achieve contrastive learning on high-dimensional image spaces. In addition, drawing negative patches internally from within the input image, rather than externally from other images in the dataset, makes it challenging to distinguish and forces the patches to preserve the content of the input better.

The objective function of our full-size model for multi-domain I2I translation is:

$$\begin{aligned} \mathcal{L}_D =& \alpha \left( -\mathcal{L}^{adv} + \mathcal{L}_r^{cls} \right), \\ \mathcal{L}_G =& \alpha \left( \mathcal{L}^{adv} + \mathcal{L}_f^{cls} \right) + \lambda \mathcal{L}^{mdc} + \beta \left( \mathcal{L}_k^{rc} + \mathcal{L}_g^{cc} + \mathcal{L}_o^{cc} + \mathcal{L}^{cgc} \right), \end{aligned} \tag{7}$$

where $\alpha, \beta, \lambda$ are the hyper-parameters for balancing the corresponding loss terms. We aim to optimize our model by minimizing $\mathcal{L}_D$ and $\mathcal{L}_G$, respectively.

## 4 EXPERIMENTS

### 4.1 Implementation Details

In order to stabilize the training process and achieve better performance, we replace the vanilla GAN in Equation (2) with Wasserstein GAN [1] with the gradient penalty [16] added into the objective function:

$$\mathcal{L}^{adv} = \mathbb{E}_x[D_{src}(x)] - \mathbb{E}_{x,s,z}[D_{src}(G(E_c(x), s, z))] - \lambda_{gp}\mathbb{E}_{\hat{x}}[(\|\nabla_{\hat{x}} D_{src}(\hat{x})\|_2 - 1)^2], \tag{8}$$

where $\hat{x}$ is sampled along straight lines between the real and generated image distribution, $s$ and $z$ denote the style code and the target domain label, respectively. We adopt an adaptive weighting scheme for gradient penalty to facilitate the training process [8]. $\lambda_{gp}$ is initially set to 10. The

Table 1. Quantitative Evaluation on INIT Dataset (a) and BDD100K Dataset (b) We Measure FID, IS and LPIPS Performances for Different Methods

**(a) INIT Dataset**

| Method | FID (↓) | | | Mean | IS (↑) | | | Mean | LPIPS (↑) | | | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *night* | *cloudy* | *rainy* | | *night* | *cloudy* | *rainy* | | *night* | *cloudy* | *rainy* | |
| MUNIT [20] | 101.1 | 30.76 | 62.26 | 64.70 | 1.03 | 1.46 | 1.26 | 1.25 | 0.25 | 0.21 | 0.14 | 0.20 |
| DRIT++ [32] | 112.9 | 25.75 | 54.30 | 64.32 | 1.15 | 1.26 | 1.34 | 1.25 | 0.06 | 0.05 | 0.05 | 0.05 |
| INIT [51] | 91.87 | 49.20 | 59.10 | 66.72 | 1.35 | 1.29 | 1.44 | 1.36 | 0.28 | 0.27 | 0.12 | 0.22 |
| HGAN [7] | 111.4 | 60.09 | 98.61 | 90.02 | 1.07 | 1.06 | 1.06 | 1.06 | 0.09 | 0.09 | 0.10 | 0.09 |
| StarGAN v2 [10] | 71.97 | 19.47 | 52.90 | 48.11 | 1.39 | 1.48 | 1.10 | 1.32 | 0.20 | 0.04 | 0.05 | 0.10 |
| CoMoGAN [47] | 86.31 | 21.12 | 45.97 | 51.13 | 1.66 | 1.57 | 1.59 | 1.61 | 0.24 | 0.20 | 0.14 | 0.19 |
| Kim et al. [25] | 82.64 | 25.45 | 54.85 | 54.31 | 1.61 | 1.33 | 1.57 | 1.50 | 0.28 | **0.28** | 0.20 | 0.25 |
| Xie et al. [56] | 121.6 | 47.22 | 68.91 | 79.24 | 1.56 | 1.55 | 1.32 | 1.48 | 0.05 | 0.07 | 0.10 | 0.07 |
| Ours (w/o MDC) | 76.34 | 14.26 | 47.43 | 46.01 | 1.70 | **1.65** | 1.52 | 1.62 | 0.31 | 0.19 | 0.18 | 0.23 |
| Ours (w/ MDC) | **60.04** | **14.08** | **44.73** | **39.62** | **1.76** | 1.62 | **1.89** | **1.76** | **0.32** | 0.25 | **0.21** | **0.26** |

**(b) BDD100K Dataset**

| Method | FID (↓) | | | Mean | IS (↑) | | | Mean | LPIPS (↑) | | | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *night* | *cloudy* | *rainy* | | *night* | *cloudy* | *rainy* | | *night* | *cloudy* | *rainy* | |
| MUNIT[20] | 59.98 | 48.93 | 81.56 | 63.49 | 1.22 | 1.34 | 1.12 | 1.23 | **0.30** | 0.19 | **0.28** | **0.26** |
| DRIT++[32] | 103.2 | 48.60 | 50.31 | 67.37 | 1.14 | 1.14 | 1.14 | 1.14 | 0.06 | 0.05 | 0.05 | 0.05 |
| INIT[51] | 43.63 | 119.2 | 52.84 | 71.89 | 1.16 | 1.27 | 1.20 | 1.21 | 0.25 | **0.27** | 0.13 | 0.21 |
| HGAN[7] | 123.6 | 65.99 | 76.91 | 88.83 | 1.01 | 1.01 | 1.01 | 1.01 | 0.18 | 0.21 | 0.27 | 0.22 |
| StarGAN v2[10] | 30.30 | 38.57 | 43.61 | 37.49 | 1.53 | 1.44 | **1.90** | 1.62 | 0.17 | 0.08 | 0.15 | 0.13 |
| CoMoGAN[47] | 44.35 | 50.39 | 62.75 | 52.50 | 1.41 | 1.56 | 1.44 | 1.47 | 0.14 | 0.19 | 0.15 | 0.16 |
| Kim et al. [25] | 57.30 | 51.07 | 45.50 | 51.29 | 1.47 | 1.39 | 1.40 | 1.42 | 0.19 | 0.22 | 0.15 | 0.19 |
| Xie et al. [56] | 128.5 | 41.64 | 81.00 | 88.71 | 1.44 | 1.25 | 1.38 | 1.36 | 0.08 | 0.11 | 0.12 | 0.10 |
| Ours (w/o MDC) | 40.14 | 34.53 | **34.05** | 36.24 | 1.50 | 1.32 | 1.26 | 1.36 | 0.26 | 0.23 | 0.22 | 0.24 |
| Ours (w/ MDC) | **30.26** | **32.77** | 34.52 | **34.11** | **1.67** | **1.65** | 1.71 | **1.68** | 0.26 | 0.26 | 0.25 | **0.26** |

We perform multi-domain translation for each domain pair. In each case, the best performance indices are highlighted in boldface fonts, and the second-best ones are highlighted in underlined fonts. Our results attain the best results.

model performs gradient regularization and accumulation every 50 iterations, and $\lambda_{gp}$ is updated once according to the gradient value with a double increase or half decrease.

Our model includes content encoder $E^c$, style encoder $E^s$, multilayer perceptron $M$, generator $G$ and discriminator $D$. The discriminator is adapted from PatchGAN which is composed of 6 convolutional layers with $4 \times 4$ filters and stride 2 and followed by the last two convolutional layers that calculate the adversarial loss and classification loss respectively. We make use of **Instance Normalization (IN)** to the content encoder $E^c$ and Adaptive Instance Normalization (AdaIN) [20] to the residual blocks of generator $G$. We apply ReLU activations in the generator and Leaky ReLU with slope 0.2 in the discriminator $D$. Furthermore, we apply $H$, a two-layer MLP with 256 units behind the multiple residual layers of the generator to map the content to the contrastive feature space. Inspired by CUT [46], we set the number of negative instances M with 64 for global (image) level and 16 for local (instance) level.

In the experiment, the weights of Encoders and Decoders use Kaiming uniform distribution, while the Discriminator sets Gaussian distribution in the initialization. We optimize the model using Adam [27] with the batch size 2 and the initial learning rate is set to 0.0001 and decreased

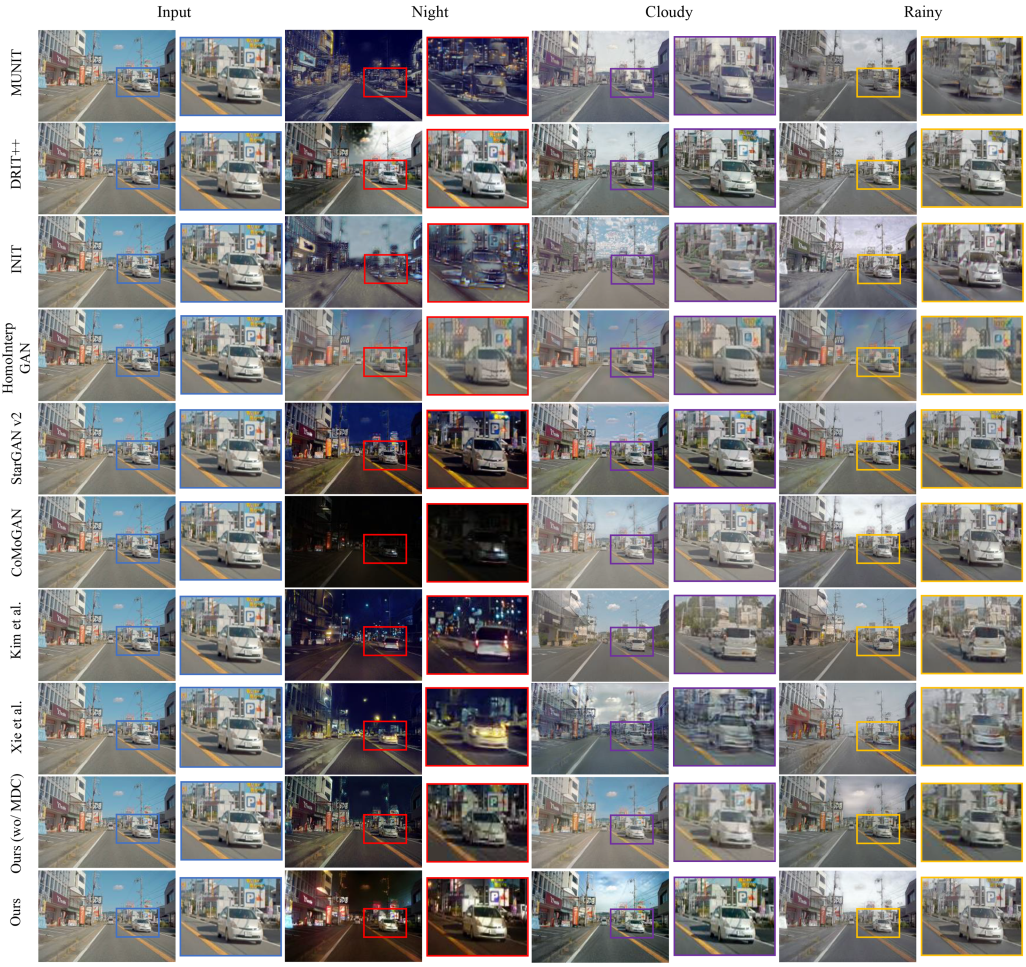


Fig. 5. Qualitative comparison of existing I2I translation methods on INIT datasets. For each baseline method, we present multi-domain outputs for the same input. Our results preserve instance details well and look realistic.

by half for every 100,000 iterations. We refer to the conference paper [13] to set $\alpha, \beta, \lambda$ to $1, 10, 1$ respectively.

## 4.2 Datasets

We conduct experiments on two standard datasets for instance-aware I2I, INIT dataset [51] and BDD100K dataset [59], to verify the generality of our method. The datasets provide object bounding box annotations to achieve instance acquisition. All of the images are resized to 256 × 256 in the experiments.

- **INIT Dataset** [51] is a large-scale street scene centric dataset with object bounding box annotations for car, person, and traffic sign. It consists of two types of resolution images, 1,208 × 1,920 and 3,000 × 4,000. We take 1,208 × 1,920 resolution ones to form a set of 38,836 images for training and 7,434 for testing. We further divide the dataset into four domains according to the weather style sunny, night, cloudy and rainy (overcast weather with wet roads).

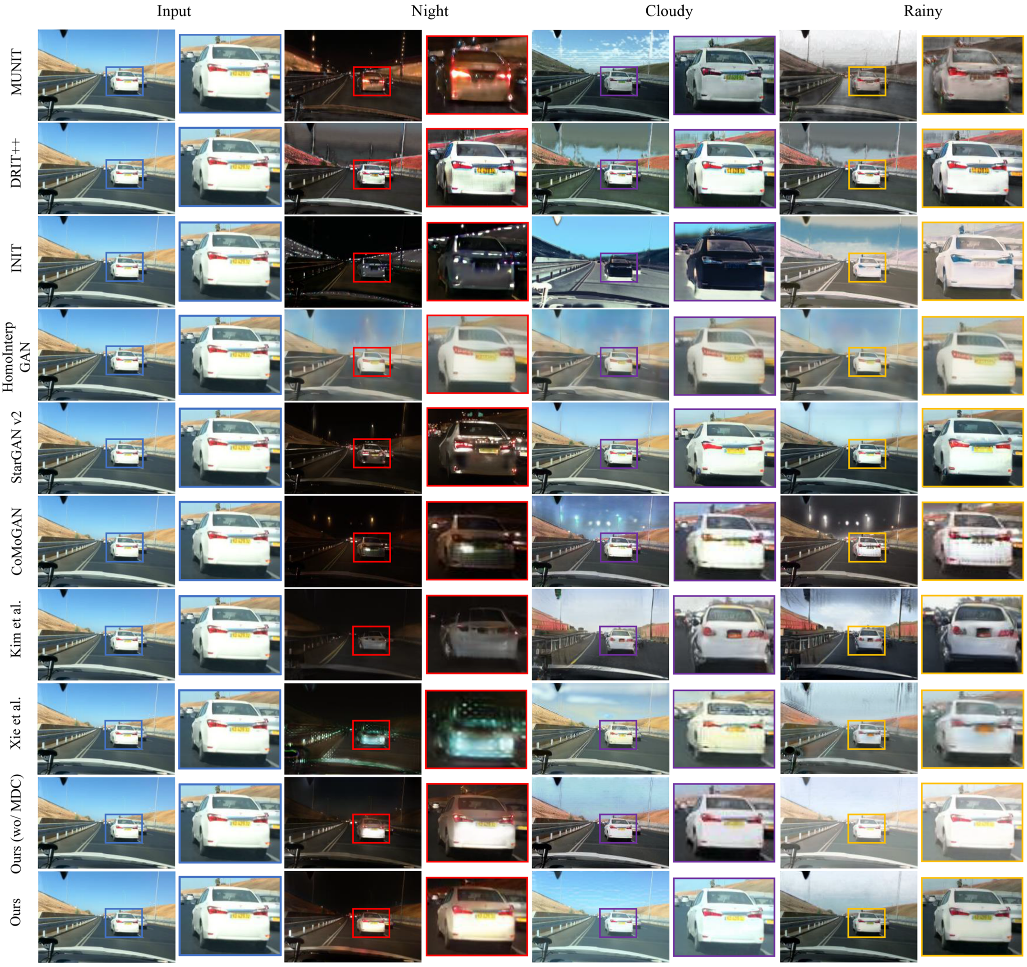


Fig. 6. Qualitative comparison of existing I2I translation methods on BDD100K datasets. For each baseline method, we present multi-domain outputs for the same input. Our results preserve instance details well and look realistic.

- **BDD100K Dataset** [59] is an open driving dataset based on urban streets and consists of 100K images with labels. The resolution of all images in this dataset is 1,208 × 1,920. There are many kinds of weather domains including sunny, cloudy, rainy, snowy and foggy in different hours. We further divide the dataset into four domains, including the weather style sunny, night, cloudy and rainy. Our rebuilt dataset is composed of 35,548 images for training and 6,570 images for testing with the bounding box annotations for ten object classes.

### 4.3 Metrics

We introduce the following three metrics to quantitatively evaluate the synthesized images by our proposed method.

- **Frechet Inception Distance (FID) [18].** FID considers the distance between the distribution of the real image and the generated image. This measurement respectively extracts the 2,048-dim feature vectors of real images and the generated images from the pre-trained

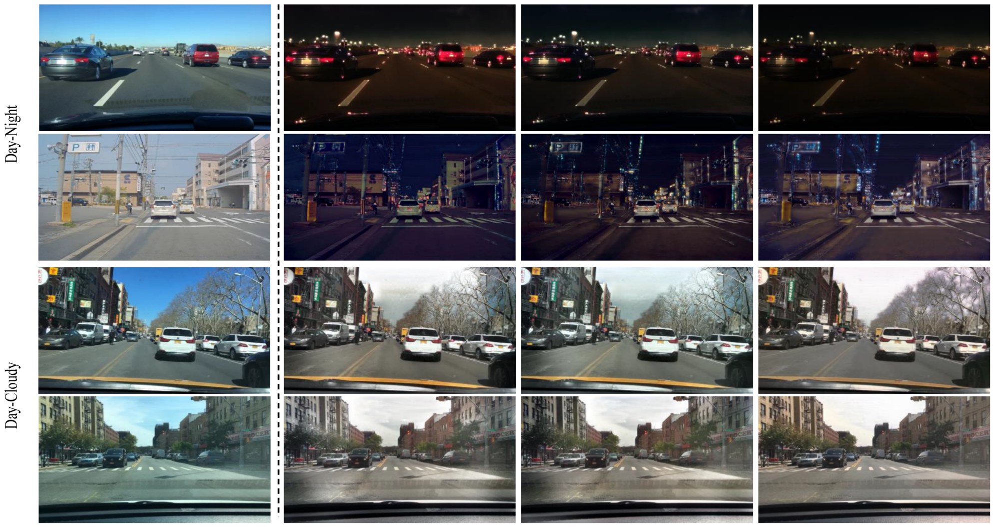


Fig. 7. Results of multi-modal image translation on INIT dataset and BDD100K dataset. For the given input sunny images (first from left), we use randomly sampled style codes to generate night images.

Inception-V3 [52] network that removes the classifier in the last layer. A lower score denotes a better quality of generated images. We calculate the FID scores between real test images and generated images.

- **Inception Scores (IS) [50].** We use the IS to evaluate the diversity of generated images and follow the settings in our original paper. Specifically, we first fine-tune the Inception-V3 model on four domain category labels of our INIT dataset and BDD100K dataset, then utilize 100 input images to generate 100 samples per input to calculate Inception Scores. A higher Inception Score indicates a higher generated diversity.
- **Learned Perceptual Image Patch Similarity (LPIPS) [62].** LPIPS is demonstrated to correlate well with a human perceptual similarity. It measures the average LPIPS distance between the pair of randomly translated samples from the same input. In the experiments, we select 100 images from the test set and generate 19 pairs of translated images via different random style codes for each image, to obtain 1,900 pairs of images [20]. Then the generated images are mapped into feature spaces by the pre-trained AlexNet [28] to calculate the distance. The higher the LPIPS score, the better the diversity and authenticity of generated images.

We conduct experiments on six baseline approaches for multi-domain I2I translation task, including MUNIT [20], DRIT++ [32], INIT [51], HGAN [7], StarGAN v2 [10], ComoGAN [47], Kim et al. [25] and Xie et al. [56], to perform qualitative and quantitative comparison with our proposed method. Note that "Ours (wo/ MDC)" means our conference paper's method.

### 4.4 Baselines

To apply MUNIT and INIT for image translation over $N$ domains, we train these two models for every pair of domains. Specifically, MUNIT decomposes image representation into domain-invariant content codes and domain-specific style codes. It performs unpaired multi-modal image translation by combining random/reference style codes and target content codes. INIT proposes an instance-aware I2I translation method that takes into account both the global images and local instances to improve the translation quality. Furthermore, it collects a large-scale benchmark and

Table 2. Quantitative Results on Average for Ablation Studies

| Method | INIT | | | BDD100K | | |
|---|---|---|---|---|---|---|
| | FID (↓) | IS (↑) | LPIPS (↑) | FID (↓) | IS (↑) | LPIPS (↑) |
| Baseline | 54.91 | 1.35 | 0.29 | 54.26 | 1.25 | 0.27 |
| w/ CGC | 46.01 | 1.62 | 0.23 | 36.24 | 1.36 | 0.24 |
| w/ MDC | 43.86 | 1.47 | **0.31** | 49.63 | 1.59 | **0.27** |
| Ours | **39.62** | **1.76** | 0.26 | **34.11** | **1.68** | 0.26 |

Baseline means our method without cross-granularity contrastive learning strategy.

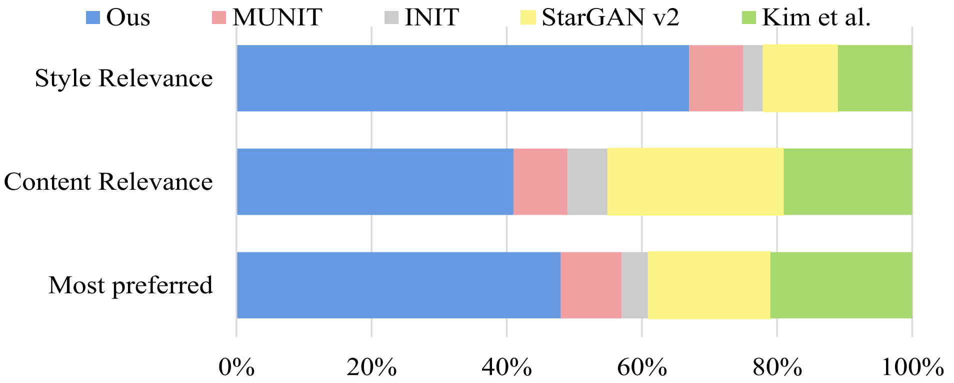


Fig. 8. User preference results. Our method is most preferred for overall quality, semantic consistency and style relevance, compared to MUNIT [69], INIT [51], StarGAN [10], and Kim et al. [25].

we adopt the dataset in the experiments. We also compare our model with multi-domain translation methods StarGAN v2, HGAN, DRIT++. As a multi-domain version of DRIT [31], DRIT++ generates diverse images via disentangled representations with cross-cycle consistency loss. It employs domain labels for multi-domain training. The discriminator perform domain classification, in addition to determining the generated image is real or fake. HGAN generates intermediate images between the two domains by proper homomorphic latent space interpolation and applies to multi-modal translation. StarGAN v2 as an improved version that only requires a single model and reference/random domain style codes to perform the image translation tasks. CoMoGAN learns continuous translations under the guidance of naive physics-inspired models. Kim et al. introduce to learn a style-specific representation and the prototype by combining a style encoder and a discriminator, as well as a data augmentation strategy for the style space. Xie et al. propose a simple way to encourage the network to find the shortest path between two domains.

### 4.5 Qualitative Evaluation

We compare our method with all the baselines one by one to verify the performance. The corresponding results on diverse datasets are shown in Figure 5 and Figure 6. All the methods compare the effects of the generated images of three different styles one by one using the same input. It can be seen that the texture quality of the global images generated by MUNIT [20], Kim et al. [25] and INIT [51] are generally good, but the generated local instances from both are blurry and far inferior to ours. For the DIRT++ [32] and HGAN [7], there is a common mode collapse phenomenon that the generated images are similar and difficult to distinguish. The style migration of StarGAN v2 [10] makes no obvious change in the details such as the color of the sky, the shape of clouds, or the humidity of the road. To conclude, most of the aforementioned

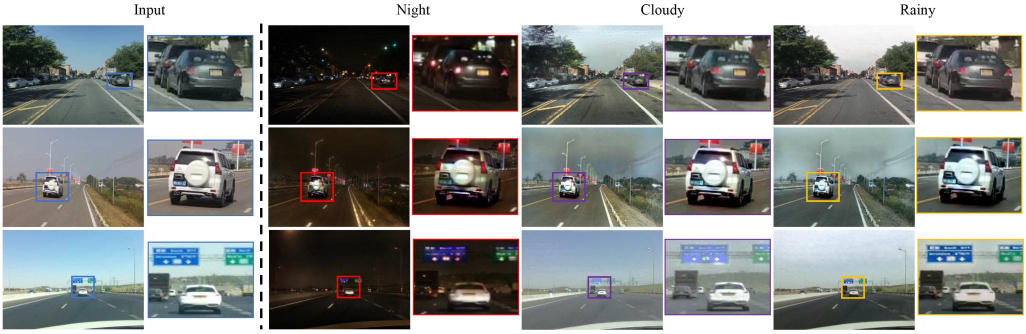


Fig. 9. Visual quality results of our method in the real-world scenarios. Our method shows the naturally overall quality and style relevance.

baselines may bring small artifacts around the object area or fail to change the style of instances since ignoring the relationship of the global image style and local instance style. Our method generally presents realistic details and high-quality images, both in terms of textures and colors.

We visualize the multi-modal results by using randomly sampled style codes on sunny→night translation task in Figure 7, our model can better learn the critical features to generate images for different styles flexibly. It shows that our model is capable of generalization.

### 4.6 Quantitative Evaluation

Table 1 shows the quantitative evaluation results of several I2I translation methods adopted on INIT dataset and BDD100K dataset. “w/ MDC” and “w/o MDC” denote that we train our model with and without considering the multilayer dual contrastive loss, respectively. We implement the tasks with StarGAN v2 [10], MUNIT [20] based on their released code while re-implementing with others based on their papers. For each baseline, we report the scores of three metrics for every domain mapping. The best performance items of each metric are shown in bold, and the second best ones are highlighted in underlined fonts. It is worth mentioning that MUNIT and INIT train one model for every pair of domains, which are not affected by each other.

Among them, the average FID scores of the three image translation tasks of our method are 13.89% and 5.88% lower than our method without dual contrastive learning scheme, and 17.65% and 8.87% lower than the baselines with the optimal FID averages of the two datasets, respectively. It proves that the image styles generated by our model are more similar to those in the target domain. For Inception scores, our performance is improved by 8.64% and 19.12%, which indicates more diversity and photorealism. It can be seen that the diversity improvement brought by our method is not obvious from the LPIPS. On average our method outperforms the baseline by quite a margin. Table 3 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.

### 4.7 User Preference

We conducted a user study to evaluate the visual quality of synthesized images on mixed INIT [51] and BDD100K [59] datasets. Ten subjects were shown with 40 different samples generated by our model or baselines. The experiments are carried out with the following questions: “Which do you think has better image quality in overall/similar content to source domain/relevant style to target domain?", summarized in Figure 8. Our method ranks the first in each case, especially on style relevance and overall preference.

Table 3. Model Size of Baselines and our Method

| Method | MUNIT [20] | DRIT++ [32] | INIT [51] | HGAN [7] | StarGAN v2 [10] |
|---|---|---|---|---|---|
| Model Size (MB) | 179 × 4 | 620 | 179 × 4 | 220 | 236 |
| Method | CoMoGAN [47] | Kim et al. [25] | Xie et al. [56] | Ours | |
| Model Size (MB) | 741 | 924 | 57 × 4 | **125** | |

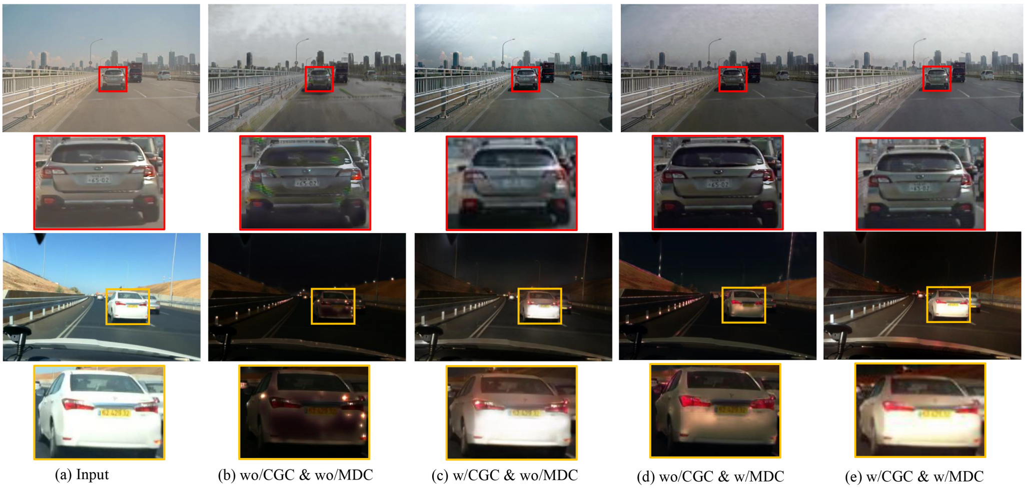


Fig. 10. Qualitative ablation study on different settings. “wo/w" means “without/with". (a) Input, (b) The method without CGC and MDC generates unpleasant artifacts and loses local details, (c) The method with CGC generates inauthentic local information, (d) The method with MDC generates unreasonable style on cars, (e) Ours. We show the results for sunny→rainy in the first column and sunny→night in the second column.

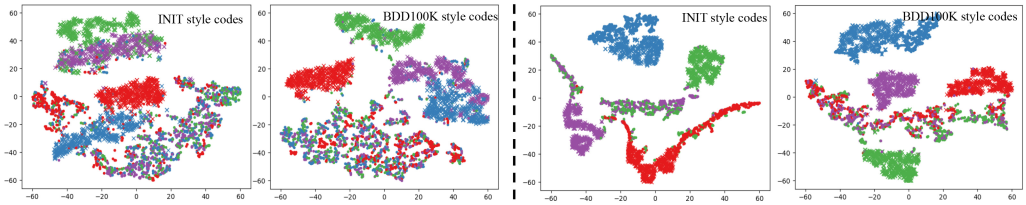
## 4.8 Ablation Studies

We evaluate the performance of our method from different perspectives. First, we present the results of three metrics obtained with the proposed cross-granularity consistency loss and the multilayer dual contrastive loss in Table 2 and Table 1. The average FID scores are improved by 27.8% on INIT dataset and 37.1% on BDD100K dataset, and the inception scores increase 30.4% and 34.4%. Nevertheless, a minor drop in LPIPS scores shows that cross-granularity learning brings the improvement of image quality at the expense of reducing the diversity of generated images due to the restrictive consistency.

Second, for the qualitative evaluations in the last two lines of Figure 10, unlike the results generated by the method w/ CGC or w/ MDC with unreasonable artifacts or styles, our current method yields more realistic global images as well as distinctive and natural instances. More importantly, the examples generated by our method are clearer and retain relatively complete semantic information. In summary, the network structure and cross-granularity contrastive strategy effectively strengthen the representation ability of the network by ensuring style and content consistency.

## 4.9 Analysis

*4.9.1 Results in Real-world Scenarios.* To verify the generalization of the proposed method, we create a dataset of 20 high-resolution and moderately bright images. We download images from the Internet. The visual results are shown in Figure 9 and the proposed method achieves considerable



(a) Baseline

(b) Ours

Fig. 11. t-SNE [29] visualization for the global and instance style features. The same colored points indicate the style features addressed to the same domains. The first (second) row are the style codes extracted by our model on INIT (BDD100K) dataset. (a) is from our baseline method (wo/ CGC and MDC) while (b) is from our method with cross-granularity contrastive learning scheme.

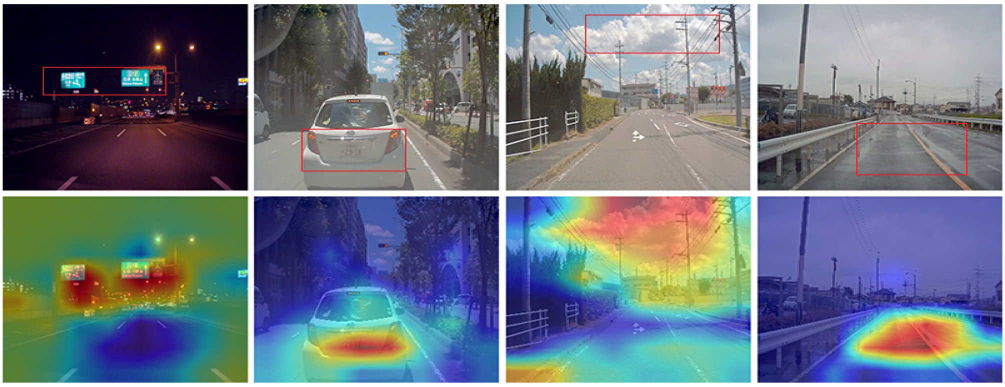


Fig. 12. Visualization of discriminator for domain classification. We use grad class activation mapping (Grad-CAM) in [49]. Best viewed in color.

visual quality in the listed examples which can synthesis accurate stylized images and instances while keeping texture and structure details.

*4.9.2 Disentangle the Global and Instance Style.* We project the embedded style codes from the images into two-dimensional space by t-SNE tools [29]. Specifically, we randomly sample 500 images and instances in the test set of each domain and visualize global (cross) and instance (dot) style codes in multiple domains. As shown in Figure 11, there are remarkable margins (In this paper, the margin refers to the distance between the local and global features) between global and instance style codes extracted from the same domain images and instances. It demonstrates the effectiveness of our model.

*4.9.3 Visualize Results with CAM.* In this part, we provide in-depth analyses for our model besides performance. we deploy the technique of **Grad Class Activation Mapping (Grad-CAM)** [49] to understand what the discriminator has learnt for multi-domain image translation tasks. Grad-CAM helps to visualize the predicted domain class scores on the global images, highlighting the discriminative instance parts. As shown in Figure 12, our model pays attention to the areas with prominent styles in the image, e.g., traffic signs at night, clouds in the sky in cloudy weather, and wet roads on rainy days. These observations qualitatively show that our method has remarkable performance for multi-domain image translation tasks.

## 5 CONCLUSION

In this paper, we propose an effective multi-domain I2I translation model with cross-granularity contrastive learning. We first present the cross-granularity consistency to guide the learning of the relationship between the image style and the instance style. Additionally, we propose a dual contrastive learning module for preserving local details of the original images or instances. A multilayer contrastive loss is proposed to encourage content preservation in unpaired I2I translations. Finally, we conduct extensive experiments on multi-domain I2I translation datasets to demonstrate the superiority of our proposed framework compared with state-of-the-art approaches.

## REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*. PMLR, Sydney, NSW, Australia, 214–223.

[2] Deblina Bhattacharjee, Seungryong Kim, Guillaume Vizier, and Mathieu Salzmann. 2020. DUNIT: Detection-based unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* Computer Vision Foundation/IEEE, Seattle, WA, USA, 4787–4796.

[3] Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. 2019. Learning disentangled semantic representation for domain adaptation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Vol. 2019. NIH Public Access, ijcai.org, Macao, China, 2060.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2017. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 4 (2017), 834–848.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*. PMLR, Virtual Event, 1597–1607.

[6] Xinlei Chen and Kaiming He. 2021. Exploring simple Siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* Computer Vision Foundation/IEEE, virtual, 15750–15758.

[7] Ying-Cong Chen, Xiaogang Xu, Zhuotao Tian, and Jiaya Jia. 2019. Homomorphic latent space interpolation for unpaired image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* Computer Vision Foundation/IEEE, Long Beach, CA, USA, 2408–2416.

[8] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* Computer Vision Foundation/IEEE, Salt Lake City, UT, USA, 6306–6314.

[9] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* Computer Vision Foundation/IEEE, Salt Lake City, UT, USA, 8789–8797.

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

[11] David Eigen and Rob Fergus. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision.* Computer Vision Foundation/IEEE, Santiago, Chile, 2650–2658.

[12] Hajar Emami, Majid Moradi Aliabadi, Ming Dong, and Ratna Babu Chinnam. 2020. SPA-GAN: Spatial attention GAN for image-to-image translation. *IEEE Transactions on Multimedia* 23 (2020), 391–401.

[13] 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.* ACM, Seattle, WA, USA, 3099–3107.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems* 27 (2014), 2672–2680.

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems* 33 (2020), 21271–21284.

[16] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C. Courville. 2017. Improved training of Wasserstein GANs. *Advances in Neural Information Processing Systems* 30 (2017), 5767–5777.

[17] Kehua Guo, Liang Chen, Xiangyuan Zhu, Xiaoyan Kui, Jian Zhang, and Heyuan Shi. 2023. Double-layer search and adaptive pooling fusion for reference-based image super-resolution. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 1 (2023), 1–23.

[18] 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. *Advances in Neural Information Processing Systems* 30 (2017), 6626–6637.

[19] Jialu Huang, Jing Liao, and Sam Kwong. 2021. Unsupervised image-to-image translation via pre-trained styleGAN2 network. *IEEE Transactions on Multimedia* 24 (2021), 1435–1448.

[20] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision*. Springer, Munich, Germany, 172–189.

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation/IEEE, Honolulu, HI, USA, 1125–1134.

[22] Somi Jeong, Youngjung Kim, Eungbean Lee, and Kwanghoon Sohn. 2021. Memory-guided unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation/IEEE, virtual, 6558–6567.

[23] Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, and Chen Change Loy. 2020. TSIT: A simple and versatile framework for image-to-image translation. In *European Conference on Computer Vision*. Springer, Glasgow, UK, 206–222.

[24] Yiting Jin, Jie Wu, Wanliang Wang, Yidong Yan, Jiawei Jiang, and Jianwei Zheng. 2023. Cascading blend network for image inpainting. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 1 (2023), 1–21.

[25] Kunhee Kim, Sanghun Park, Eunyeong Jeon, Taehun Kim, and Daijin Kim. 2022. A style-aware discriminator for controllable image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation/IEEE, New Orleans, LA, USA, 18239–18248.

[26] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. 2017. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, Vol. 70. PMLR, Sydney, NSW, Australia, 1857–1865.

[27] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*. openreview, San Diego, CA, USA, 0–0.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* 25 (2012), 1106–1114.

[29] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605.

[30] Thi-Ngoc-Hanh Le, Ya-Hsuan Chen, and Tong-Yee Lee. 2023. Structure-aware video style transfer with map art. *ACM Transactions on Multimedia Computing, Communications and Applications* 19 (2023), 1–25.

[31] 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 *Proceedings of the European Conference on Computer Vision*. Springer, Munich, Germany, 35–51.

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

[33] Chunyuan Li, Hao Liu, Changyou Chen, Yuchen Pu, Liqun Chen, Ricardo Henao, and Lawrence Carin. 2017. ALICE: Towards understanding adversarial learning for joint distribution matching. *Advances in Neural Information Processing Systems* 30 (2017), 5495–5503.

[34] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. 2017. Perceptual generative adversarial networks for small object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation/IEEE, Honolulu, HI, USA, 1951–1959.

[35] Kun Liu, Wu Liu, Chuang Gan, Mingkui Tan, and Huadong Ma. 2018. T-C3D: Temporal convolutional 3D network for real-time action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, New Orleans, Louisiana, USA, 7138–7145.

[36] Kun Liu, Wu Liu, Huadong Ma, Mingkui Tan, and Chuang Gan. 2020. A real-time action representation with temporal encoding and deep compression. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 2 (2020), 647–660.

[37] Kun Liu and Huadong Ma. 2019. Exploring background-bias for anomaly detection in surveillance videos. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, Nice, France, 1490–1499.

[38] Kun Liu, Minzhi Zhu, Huiyuan Fu, Huadong Ma, and Tat-Seng Chua. 2020. Enhancing anomaly detection in surveillance videos with transfer learning from action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, Seattle, WA, USA, 4664–4668.

[39] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*. MIT, Long Beach, CA, USA, 700–708.

[40] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. 2021. AdaAttN: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Montreal, QC, Canada, 6649–6658.

[41] Ziyang Liu, Zhengguo Li, Xingming Wu, Zhong Liu, and Weihai Chen. 2022. DSRGAN: Detail prior-assisted perceptual single image super-resolution via generative adversarial networks. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 11 (2022), 7418–7431.

[42] Ishan Misra and Laurens van der Maaten. 2020. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation/IEEE, Seattle, WA, USA, 6707–6717.

[43] Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*. Springer, Amsterdam, The Netherlands, 69–84.

[44] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* 0 (2018), 0.

[45] Yingxue Pang, Jianxin Lin, Tao Qin, and Zhibo Chen. 2021. Image-to-image translation: Methods and applications. *IEEE Transactions on Multimedia* 24 (2021), 3859–3881.

[46] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. 2020. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, Glasgow, UK, 319–345.

[47] Fabio Pizzati, Pietro Cerri, and Raoul de Charette. 2021. CoMoGAN: Continuous model-guided image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation/IEEE, virtual, 14288–14298.

[48] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*. openreview, virtual, 0.

[49] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision* 128, (2020), 336–359.

[50] 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*. MIT, Barcelona, Spain, 2226–2234.

[51] Zhiqiang Shen, Mingyang Huang, Jianping Shi, Xiangyang Xue, and Thomas S. Huang. 2019. Towards instance-level image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation/IEEE, Long Beach, CA, USA, 3683–3692.

[52] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation/IEEE, Las Vegas, NV, USA, 2818–2826.

[53] Xiaolong Wang and Abhinav Gupta. 2016. Generative image modeling using style and structure adversarial networks. In *European Conference on Computer Vision*. Springer, Amsterdam, The Netherlands, 318–335.

[54] Yuxi Wang, Zhaoxiang Zhang, Wangli Hao, and Chunfeng Song. 2020. Multi-domain image-to-image translation via a unified circular framework. *IEEE Transactions on Image Processing* 30 (2020), 670–684.

[55] Jiayu Xiao, Liang Li, Chaofei Wang, Zheng-Jun Zha, and Qingming Huang. 2022. Few shot generative model adaption via relaxed spatial structural alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation/IEEE, New Orleans, LA, USA, 11204–11213.

[56] Shaoan Xie, Yanwu Xu, Mingming Gong, and Kun Zhang. 2023. Unpaired image-to-image translation with shortest path regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation/IEEE, Vancouver, BC, Canada, 10177–10187.

[57] Xuewen Yang, Dongliang Xie, and Xin Wang. 2018. Crossing-domain generative adversarial networks for unsupervised multi-domain image-to-image translation. In *Proceedings of the 26th ACM International Conference on Multimedia*. ACM, Seoul, Republic of Korea, 374–382.

[58] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. DualGAN: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*. Computer Vision Foundation/IEEE, Venice, Italy, 2849–2857.

[59] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. 2020. Bdd100k: A diverse driving video database with scalable annotation tooling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE/CVF, Seattle, WA, USA, 2636–2645.

[60] Masoumeh Zareapoor and Jie Yang. 2021. Equivariant adversarial network for image-to-image translation. *ACM Transactions on Multimedia Computing, Communications, and Applications* 17, 2s (2021), 1–14.

[61] Jiqing Zhang, Chengjiang Long, Yuxin Wang, Haiyin Piao, Haiyang Mei, Xin Yang, and Baocai Yin. 2021. A two-stage attentive network for single image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 3 (2021), 1020–1033.

[62] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation/IEEE, Salt Lake City, UT, USA, 586–595.

[63] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. 2019. RankSRGAN: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, Seoul, Korea (South), 3096–3105.

[64] Yuzhi Zhao, Lai-Man Po, Wing Yin Yu, Yasar Abbas Ur Rehman, Mengyang Liu, Yujia Zhang, and Weifeng Ou. 2022. VCGAN: Video colorization with hybrid generative adversarial network. *IEEE Transactions on Multimedia* 25 (2022), 3017–3032.

[65] Jianwei Zheng, Yu Liu, Yuchao Feng, Honghui Xu, and Meiyu Zhang. 2023. Contrastive attention-guided multi-level feature registration for reference-based super-resolution. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 2 (2023), 1–21.

[66] Ziqiang Zheng, Zhibin Yu, Haiyong Zheng, Yang Yang, and Heng Tao Shen. 2021. One-shot image-to-image translation via part-global learning with a multi-adversarial framework. *IEEE Transactions on Multimedia* 24 (2021), 480–491.

[67] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. 2016. Generative visual manipulation on the natural image manifold. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, Amsterdam, The Netherlands, 597–613.

[68] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE Computer Society, Venice, Italy, 2223–2232.

[69] 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 *Advances in Neural Information Processing Systems*. MIT, Long Beach, CA, USA, 465–476.

Received 9 December 2023; revised 19 February 2024; accepted 20 March 2024