Towards Instance-level Image-to-Image Translation

Zhiqiang Shen[1,3*], Mingyang Huang[2], Jianping Shi[2], Xiangyang Xue[3], Thomas Huang[1]
[1]University of Illinois at Urbana-Champaign, [2]SenseTime Research, [3]Fudan University
zhiqiangshen0214@gmail.com {huangmingyang, shijianping}@sensetime.com
xyxue@fudan.edu.cn t-huang1@illinois.edu

Abstract

Unpaired Image-to-image Translation is a new rising and challenging vision problem that aims to learn a mapping between unaligned image pairs in diverse domains. Recent advances in this field like MUNIT [11] and DRIT [17] mainly focus on disentangling content and style/attribute from a given image first, then directly adopting the global style to guide the model to synthesize new domain images. However, this kind of approaches severely incurs contradiction if the target domain images are content-rich with multiple discrepant objects. In this paper, we present a simple yet effective instance-aware image-to-image translation approach (INIT), which employs the fine-grained local (instance) and global styles to the target image spatially. The proposed INIT exhibits three import advantages: (1) the instance-level objective loss can help learn a more accurate reconstruction and incorporate diverse attributes of objects; (2) the styles used for target domain of local/global areas are from corresponding spatial regions in source domain, which intuitively is a more reasonable mapping; (3) the joint training process can benefit both fine and coarse granularity and incorporates instance information to improve the quality of global translation. We also collect a large-scale benchmark[1] for the new instance-level translation task. We observe that our synthetic images can even benefit real-world vision tasks like generic object detection.

1. Introduction

In the recent years, Image-to-Image (I2I) translation has received significant attention in computer vision community, since many vision and graphics problems can be formulated as an I2I translation problem like super-resolution, neural style transfer, colorization, etc. This technique has also been adapted to the relevant fields such as medical image processing [40] to further improve the medical volumes segmentation performance. In general, Pix2pix [13] is regarded as the first unified framework for I2I translation which adopts conditional generative adversarial networks [26] for image generation, while it requires the paired examples during training process. A more general and challenging setting is the unpaired I2I translation, where the paired data is unavailable.

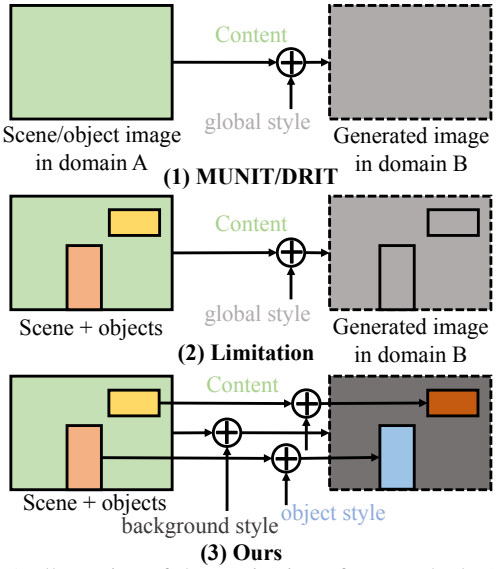


Figure 1. Illustration of the motivation of our method. (1) MUNIT [11]/DRIT [17] methods; (2) their limitation; and (3) our solution for instance-level translation. More details can be referred to the text.

Several recent efforts [42, 21, 11, 17, 1] have been made on this direction and achieved very promising results. For instance, CycleGAN [42] proposed the cycle consistency loss to enforce the learning process that if an image is translated to the target domain by learning a mapping and translated back with an inverse mapping, the output should be the original image. Furthermore, CycleGAN assumes the latent spaces are separate of the two mappings. In contrast, UNIT [21] assumes two domain images can be mapped onto

*Work done during internship at SenseTime.
[1]contains 155,529 high-resolution natural images across four different modalities with object bounding box annotations. A summary of the entire dataset is provided in the following sections.

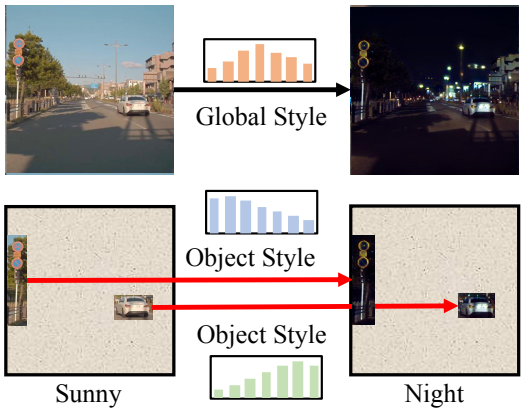


Figure 2. A natural image example of our I2I translation.

a shared latent space. MUNIT [11] and DRIT [17] further postulate that the latent spaces can be disentangled to a shared content space and a domain-specific attribute space.

However, all of these methods thus far have focused on migrating styles or attributes onto the entire images. As shown in Fig. 1 (1), they work well on the unified-style scenes or relatively content-simple scenarios due to the consistent pattern across various spatial areas in an image, while this is not true for the complex structure images with multiple objects since the stylistic vision disparity between objects and background in an image is always huge or even totally different, as in Fig. 1 (2).

To address the aforementioned limitation, in this paper we present a method that can translate objects and background/global areas separately with different style codes as in Fig. 1 (3), and still training in an end-to-end manner. The motivation of our method is illustrated in Fig. 2. Instead of using the global style, we use instance-level style vectors that can provide more accurate guidance for visually related object generation in target domain. We argue that styles should be diverse for different objects, background or global image, meaning that the style codes should not be identical for the entire image. More specifically, a car from "sunny" to the "night" domain should have different style codes comparing to the global image translation between these two domains. Our method achieves this goal by involving the instance-level styles. Given a pair of unaligned images and object locations, we first apply our encoders to obtain the intermediate global and instance level content and style vectors separately. Then we utilize the cross-domain mapping to obtain the target domain images by swapping the style/attribute vectors. Our swapping strategy is introduced with more details in Sec. 3. The main advantage of our method is the exploration and usage of object level styles, which affects and guides the generation of target domain objects directly. Certainly, we can also apply the global style for target objects to enforce the model to learn more diverse results.

In summary, our contributions are three fold:

- We propel I2I translation problem step forward to instance-level such that the constraints could be exploited on both instance and global-level attributes by adopting the proposed compound loss.
- We conduct extensive qualitative and quantitative experiments to demonstrate that our approach can surpass against the baseline I2I translation methods. Our synthetic images can be even beneficial to other vision tasks such as generic object detection, and further improve the performance.
- We introduce a large-scale, multimodal, highly varied I2I translation dataset, containing ∼155k streetscape images across four domains. Our dataset not only includes the domain category labels, but also provides the detailed object bounding box annotations, which will benefit the instance-level I2I translation problem.

2. Related Work

Image-to-Image Translation. The goal of I2I translation is to learn the mapping between two different domains. Pix2pix [13] first proposes to use conditional generative adversarial networks [26] to model the mapping function from input to output images. Inspired by Pix2pix, some works further adapt it to a variety of relevant tasks, such as semantic layouts → scenes [14], sketches → photographs [33], etc. Despite popular usage, the major weaknesses of these methods are that they require the paired training examples and the outputs are single-modal. In order to produce multimodal and more diverse images, BicycleGAN [43] encourages the bijective consistency between the latent and target spaces to avoid the mode collapse problem. A generator learns to map the given source image, combined with a low-dimensional latent code, to the output during training. While this method still needs the paired training data.

Recently, CycleGAN [42] is proposed to tackle the unpaired I2I translation problem by using the cycle consistency loss. UNIT [21] further makes a share-latent assumption and adopts Coupled GAN in their method. To address the multimodal problem, MUNIT [11], DRIT [17], Augmented CycleGAN [1], etc. adopt a disentangled representation to further learn diverse I2I translation from unpaired training data.

Instance-level Image-to-Image Translation. To the best of our knowledge, there are so far very few efforts on the instance-level I2I translation problem. Perhaps the most similar to our work is the recently proposed *Insta*GAN [27], which utilizes the object segmentation masks to translate both an image and the corresponding set of instance attributes while maintaining the permutation invariance property of instances. A context preserving loss is designed to encourage model to learn the identity function outside of target instances. The main difference with ours is that *in-*

Datasets	Paired	Resolution	Bbox annotations	Modalities	# images
edge↔shoes [13]	✓	low	-	{edge, shoes}	50,000
edge↔handbags [13]	✓	low	-	{edge, handbags}	137,000
CMP Facades [31]	✓	HD	-	{facade, semantic map}	606
Yosemite (summer↔winter) [42]	✗	HD	-	{summer, winter}	2,127
Yosemite* (MUNIT) [11]	✗	HD	-	{summer, winter}	5,638
Cityscapes [4]	✓	HD	✓	{ semantic, realistic}	3,475
Transient Attributes [16]	✓	HD	✗	{40 transient attributes}	8,571
Ours	✗	HD†	✓	{sunny, night, cloudy, rainy}	**155,529**

Table 1. **Feature-by-feature comparison of popular I2I translation datasets.** Our dataset contains four relevant but visually-different domains: sunny, night, cloudy and rainy. †The images in our dataset contain two types of resolutions: 1208×1920 and 3000×4000.

*sta*GAN cannot translate different domains for an entire image sufficiently. They focus on translating instances and maintain the outside areas, in contrast, our method can translate instances and outside areas simultaneously and make global images more realistic. Furthermore, *Insta*GAN is built on the CycleGAN [42], which is single modal, while we choose to leverage the MUNIT [11] and DRIT [17] to build our INIT, thus our method inherits multimodal and unsupervised properties, meanwhile, produces more diverse and higher quality images.

Some other existing works [23, 18] are more or less related to this paper. For instance, DA-GAN [23] learns a deep attention encoder to enable the instance-level translation, which is unable to handle the multi-instance and complex circumstance. BeautyGAN [18] focuses on facial makeup transfer by employing histogram loss with face parsing mask.

A New Benchmark for Unpaired Image-to-Image Translation. We introduce a new large-scale street scene centric dataset that addresses three core research problems in I2I translation: (1) unsupervised learning paradigm, meaning that there is no specific one-to-one mapping in the dataset; (2) multimodal domains incorporation. Most existing I2I translation datasets provide only two different domains, which limit the potential to explore more challenging task like multi-domain incorporation circumstance. Our dataset contains four domains: sunny, night, cloudy and rainy[2] in a unified street scene; and (3) multi-granularity (global and instance-level) information. Our dataset provides instance-level bounding box annotations, which can utilize more details for learning a translation model. Tab. 1 shows a feature-by-feature comparison among various I2I translation datasets. We also visualize some examples of the dataset in Fig. 6. For instance category, we annotate three common objects in street scenes including: car, person, traffic sign (speed limited sign). The detailed statistics (# images) of the entire dataset are shown in Sec. 4.

[2]For safety, we collect the rainy images after the rain, so this category looks more like overcast weather with wet road.

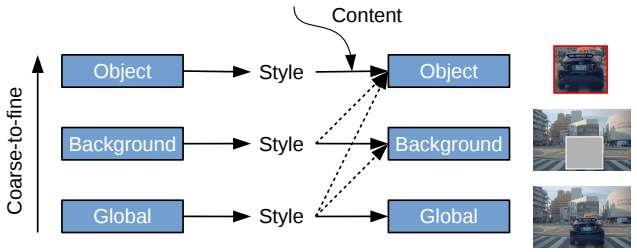


Figure 3. Our content-style pair association strategy. Only coarse styles can be applied to fine contents, the reversal of processing flow is not allowed during training.

3. Instance-aware Image-to-Image Translation

Our goal is to realize the instance-aware I2I translation between two different domains without paired training examples. We build our framework by leveraging the MUNIT [11] and DRIT [17] methods. To avoid repetition, we omit some innocuous details. Similar to MUNIT [11] and DRIT [17], our method is straight-forward and simple to implement. As illustrated in Fig. 5, our translation model consists of two encoders E_g, E_o (g and o denote the global and instance image regions respectively), and two decoders G_g, G_o in each domain $\mathcal{X}$ or $\mathcal{Y}$. A more detailed illustration is shown in Fig. 4. Since we have the object coordinates, we can crop the object areas and feed them into the instance-level encoder to extra the content/style vectors. An alternative method for object content vectors is to adopt RoI pooling [5] from the global image content features. Here we use image crop (object region) and share the parameters for the two encoders, which is more easier to implement.

Disentangle content and style on object and entire image. As [3, 25, 11, 17], our method also decomposes input images/objects into a shared content space and a domain-specific style space. Take global image as an example, each encode E_g can decompose the input to a content code c_g and a style code s_g, where $E_g = (E_g^c, E_g^s)$, $c_g = E_g^c(I)$, $s_g = E_g^s(I)$, I denotes the input image representation. c_g and s_g are global-level content/style features.

Generate style code bank. We generate the style codes from objects, background and entire images, which form

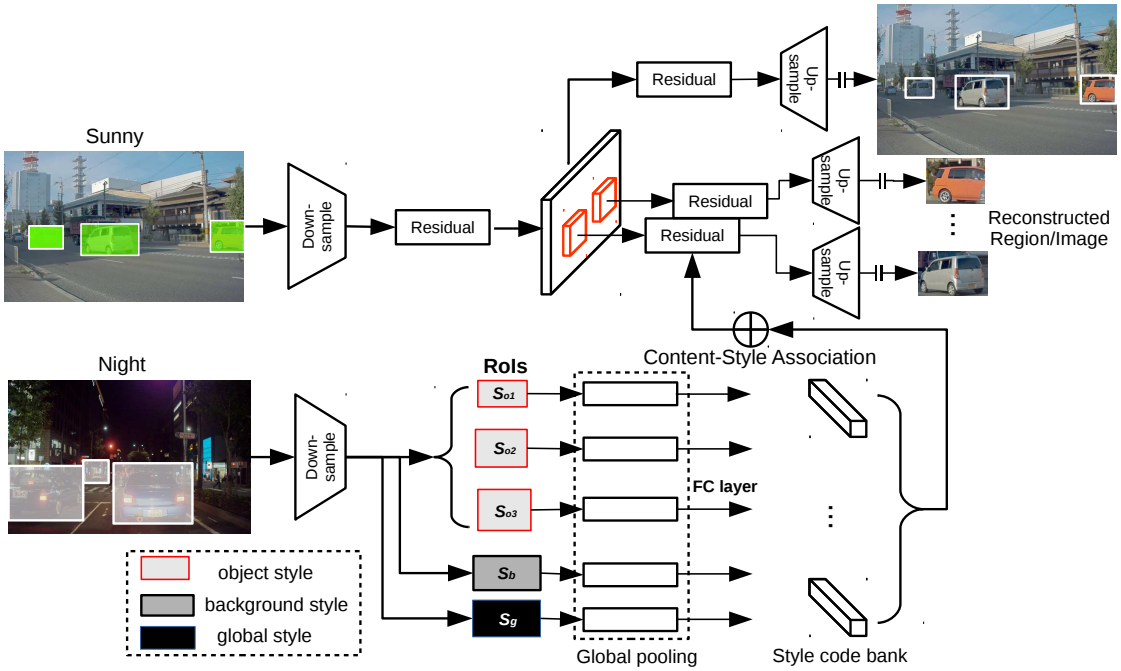


Figure 4. Overview of our instance-aware cross-domain I2I translation. The whole framework is based on the MUNIT method [11], while we further extend it to realize the instance-level translation purpose. Note that after content-style association, the generated images will place in the target domain, so a translation back process will be employed before self-reconstruction, which is not illustrated here.

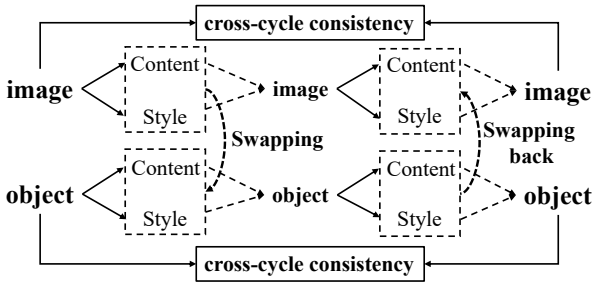


Figure 5. Illustration of our cross-cycle consistency process. We only show cross-granularity (image $\leftrightarrow$ object), the cross-domain consistency ($\mathcal{X} \leftrightarrow \mathcal{Y}$) is similar to the above paradigm.

our style code bank for the following swapping operation and translation. In contrast, MUNIT [11] and DRIT [17] use only the entire image style or attribute, which is struggling to model and cover the rich image spatial representation.

Associate content-style pairs for cyclic reconstruction. Our cross-cycle consistency is performed by swapping encoder-decoder pairs (dashed arc lines in Fig. 5). The cross-cycle includes two modes: cross-domain ($\mathcal{X} \leftrightarrow \mathcal{Y}$) and cross-granularity (entire image $\leftrightarrow$ object). We illustrate cross-granularity (image $\leftrightarrow$ object) in Fig. 5, the cross-domain consistency ($\mathcal{X} \leftrightarrow \mathcal{Y}$) is similar to MUNIT [11] and DRIT [17]. As shown in Fig. 3, the swapping or content-style association strategy is a hierarchical structure across multi-granularity areas. Intuitively, the coarse (global) style can affect fine content and be adopted to local areas, while it's not true if the process is reversed. Following [11], we also use AdaIN [10] to combine the content and style vectors.

Incorporate Multi-Scale. It's technically easy to incorporate multi-scale advantage into the framework. We simply replace the object branch in Fig. 5 with resolution-reduced images. In our experiments, we use 1/2 scale and original size images as pairs to perform scale-augmented training. Specifically, styles from small size and original size images can be performed to each other, and the generator needs to learn multi-scale reconstruction for both of them, which leads to more accurate results.

Reconstruction loss. We use self-reconstruction and cross-cycle consistency loss [17] for both entire image and object that encourage reconstruction of them. With encoded c and s, the decoders should decode them back to original input,

$$\hat{I} = G_g(E_g^c(I),\ E_g^s(I)),\ \hat{o} = G_o(E_o^c(o),\ E_o^s(o)) \quad (1)$$

We can also reconstruct the latent distribution (i.e. content and style vectors) as [11].

$$\hat{c}_o = E_o^c(G_o(c_o, s_g)),\ \hat{s}_o = E_o^s(G_o(c_o, s_g)) \quad (2)$$

where c_o and s_g are instance-level content and global-level style features. Then, we can use the following formation to learn a reconstruction of them:

$$\mathcal{L}_{recon}^k = \mathbb{E}_{k\sim p(k)} \left[\left\| \hat{k} - k \right\|_1 \right] \quad (3)$$

Figure 6. Image samples from our benchmark grouped by their domain categories (sunny, night, cloudy and rainy). In each group, left are original images and right are images with corresponding bounding box annotations.

where k can be I, o, c or s. $p(k)$ denotes the distribution of data k. The formation of cross-cycle consistency is similar to this process and more details can be referred to [17].

Adversarial loss. Generative adversarial learning [6] has been adapted to many visual tasks, e.g., detection [28, 2], inpainting [30, 38, 12, 37], ensemble [34], etc. We adopt adversarial loss $\mathcal{L}_{adv}$ where $D^g_{\mathcal{X}}$, $D^o_{\mathcal{X}}$, $D^g_{\mathcal{Y}}$ and $D^o_{\mathcal{Y}}$ attempt to discriminate between real and synthetic images/objects in each domain. We explore two designs for the discriminators: weight-sharing or weight-independent for global and instance images in each domain. The ablation experimental results are shown in Tab. 3 and Tab. 4, we observe that shared discriminator is a better choice in our experiments.

Full objective function. The full objective function of our framework is:

$$\begin{aligned}
&\min_{E_{\mathcal{X}},E_{\mathcal{Y}},G_{\mathcal{X}},G_{\mathcal{Y}}} \max_{D_{\mathcal{X}},D_{\mathcal{Y}}} \mathcal{L}(E_{\mathcal{X}},E_{\mathcal{Y}},G_{\mathcal{X}},G_{\mathcal{Y}},D_{\mathcal{X}},D_{\mathcal{Y}}) \\
&= \underbrace{\lambda_g(\mathcal{L}^{g_{\mathcal{X}}} + \mathcal{L}^{g_{\mathcal{Y}}}) + \lambda_{c_g}(\mathcal{L}_g^{c_{\mathcal{X}}} + \mathcal{L}_g^{c_{\mathcal{Y}}}) + \lambda_{s_g}(\mathcal{L}_g^{s_{\mathcal{X}}} + \mathcal{L}_g^{s_{\mathcal{Y}}})}_{global-level\ reconstruction\ loss} \\
&\quad + \underbrace{\lambda_o(\mathcal{L}^{o_{\mathcal{X}}} + \mathcal{L}^{o_{\mathcal{Y}}}) + \lambda_{c_o}(\mathcal{L}_o^{c_{\mathcal{X}}} + \mathcal{L}_o^{c_{\mathcal{Y}}}) + \lambda_{s_o}(\mathcal{L}_o^{s_{\mathcal{X}}} + \mathcal{L}_o^{s_{\mathcal{Y}}})}_{instance-level\ reconstruction\ loss} \\
&\quad + \underbrace{\mathcal{L}_{adv}^{\mathcal{X}_g} + \mathcal{L}_{adv}^{\mathcal{Y}_g}}_{global-level\ GAN\ loss} + \underbrace{\mathcal{L}_{adv}^{\mathcal{X}_o} + \mathcal{L}_{adv}^{\mathcal{Y}_o}}_{instance-level\ GAN\ loss}
\end{aligned} \tag{4}$$

During inference time, we simply use the global branch to generate the target domain images (See Fig. 4 upper-right part) so that it's not necessary to use bounding box annotations at this stage, and this strategy can also guarantee that the generated images are harmonious.

Domain	Training (85%)	Testing (15%)	Total (100%)
Sunny	49,663	8,764	58,427
Night	24,559	4,333	28,892
Rainy	6,041	1,066	7,107
Cloudy	51,938	9,165	61,103
Total	132,201	23,328	155,529

Table 2. Statistics (# images) of the entire dataset across four domains: sunny, night, rainy and cloudy. The data is divided into two subsets: 85% for training and 15% for testing.

4. Experiments and Analysis

We conduct experiments on our collected dataset (INIT). We also use COCO dataset [20] to verify the effectiveness of data augmentation.

INIT Dataset. INIT dataset consists of 132,201 images for training and 23,328 images for testing. The detailed statistics are shown in Tab. 2. All the data are collected in Tokyo, Japan with SEKONIX AR0231 camera. The whole collection process lasted about three months.

Implementation Details. Our implementation is based on MUNIT[3] with PyTorch [29]. For I2I translation, we resize the short side of images to 360 pixels due to the limitation of GPU memory. For COCO image synthesis, since the training images (INIT dataset) and target images (COCO) are in different distributions, we keep the original size of our training image and crop 360×360 pixels to train our model, in order to learn more details of images and objects, meanwhile, ignore the global information. In this circumstance, we build our object part as an independent branch and each

[3]https://github.com/NVlabs/MUNIT

Method	Diversity			
	sunny → night	sunny→rainy	sunny→cloudy	Average
UNIT [21]	0.067	0.062	0.068	0.066
CycleGAN [42]	0.016	0.008	0.011	0.012
MUNIT [11]	0.292	0.239	0.211	0.247
DRIT [17]	0.231	0.173	0.166	0.190
INIT w/ D_s	**0.330**	**0.267**	**0.224**	**0.274**
INIT w/o D_s	0.324	0.238	0.177	0.246
Real Images	0.573	0.489	0.465	0.509

Table 3. **Diversity scores on our dataset.** We use the average LPIPS distance [39] to measure the diversity of generated images.

	CycleGAN [42]		UNIT [21]		MUNIT [11]		DRIT [17]		INIT w/ D_s		INIT w/o D_s	
	CIS	IS	CIS	IS	CIS	IS	CIS	IS	CIS	IS	CIS	IS
sunny→night	0.014	1.026	0.082	1.030	1.159	1.278	1.058	1.224	1.060	1.118	1.083	1.120
night→sunny	0.012	1.023	0.027	1.024	1.036	1.051	1.024	1.099	1.045	1.080	1.024	1.104
sunny→rainy	0.011	1.073	0.097	1.075	1.012	1.146	1.007	1.207	1.036	1.152	1.034	1.146
rainy→sunny	0.010	1.090	0.014	1.023	1.055	1.102	1.028	1.103	1.060	1.119	1.059	1.124
sunny→cloudy	0.014	1.097	0.081	1.134	1.008	1.095	1.025	1.104	1.040	1.142	1.025	1.147
cloudy→sunny	0.090	1.033	0.219	1.046	1.026	1.321	1.046	1.249	1.016	1.460	1.006	1.363
Average	0.025	1.057	0.087	1.055	1.032	1.166	1.031	1.164	**1.043**	**1.179**	1.039	1.167

Table 4. **Comparison of Conditional Inception Score (CIS) and Inception Score (IS).** To obtain high CIS and IS scores, a model is required to synthesis images that are more realistic, diverse with high-quality.

object is resized to 120×120 pixels during training.

4.1. Baselines

We perform our evaluation on the following four recent proposed state-of-the-art unpaired I2I translation methods:

- CycleGAN [42]: CycleGAN contains two translation functions ($\mathcal{X} \rightarrow \mathcal{Y}$ and $\mathcal{X} \leftarrow \mathcal{Y}$), and the corresponding adversarial loss. It assumes that the input images can be translated to another domain and then can be mapped back with a cycle consistency loss.
- UNIT [21]: The UNIT method is an extension of CycleGAN [42] that is based on the shared latent space assumption. It contains two VAE-GANs and also uses cycle-consistency loss [42] for learning models.
- MUNIT [11]: MUNIT consists of an encoder and a decoder for each domain. It assumes that the image representation can be decomposed into a domain-invariant content space and a domain-specific style space. The latent vectors of each encoder are disentangled to a content vector and a style vector. I2I translation is performed by swapping content-style pairs.
- DRIT [17]: The motivation of DRIT is similar to MUNIT. It consists of content encoders, attribute encoders, generators and domain discriminators for both domains. The content encoder maps images into a shared content space and the attribute encoder maps images into a domain-specific attribute space. A cross-cycle consistency loss is adopted for performing I2I translation.

Figure 7. **Visualization of our synthetic images.** The left group images are from COCO and the right are from Cityscapes.

4.2. Evaluation

We adopt the same evaluation protocol from previous unsupervised I2I translation works and evaluate our method with the LPIPS Metric [39], Inception Score (IS) [32] and Conditional Inception Score (CIS) [11].

LPIPS Metric. Zhang et al. proposed LPIPS distance [39] to measure the translation diversity, which has been verified to correlate well with human perceptual psychophysical similarity. Following [11], we calculate the average LPIPS distance between 19 pairs of randomly sampled translation outputs from 100 input images of our test set. Following [11] and recommended by [39], we also use the pre-trained AlexNet [15] to extract deep features.

Results are summarized in Tab. 3, "INIT w/ D_s" denotes we train our model with shared discriminator between entire

COCO 2017 training		COCO 2017 validation		object detection (%)			instance segmentation (%)		
Real	Synthetic	Real	Synthetic	Avg. Precision, IoU: 0.5:0.95	0.5	0.75	Avg. Precision, mask: 0.5:0.95	0.5	0.75
✓		✓		37.7	59.2	40.8	34.3	56.0	36.2
✓			✓	30.4	49.7	32.6	27.8	46.6	29.2
	✓	✓		30.0	50.0	31.6	27.2	46.5	28.0
	✓		✓	30.5	49.7	32.7	27.8	46.4	29.0
✓	✓		✓	$32.6^{\uparrow 2.1}$	$52.6^{\uparrow 2.9}$	$34.2^{\uparrow 1.5}$	$29.0^{\uparrow 1.2}$	$49.0^{\uparrow 2.6}$	$29.8^{\uparrow 0.8}$
✓	✓	✓		$38.8^{\uparrow 1.1}$	$60.2^{\uparrow 1.0}$	$42.5^{\uparrow 1.7}$	$35.2^{\uparrow 0.9}$	$57.0^{\uparrow 1.0}$	$37.4^{\uparrow 1.2}$

Table 5. **Mask-RCNN with ResNet-50-FPN [19] detection and segmentation results on MS COCO 2017 `val` set.**

Figure 8. **Visualization of multimodal results.** We use randomly sampled style codes to generate these images and the darkness are slightly different across them.

COCO 2017 (%)	IoU	$IoU_{0.5}$	$IoU_{0.75}$
+Syn. (MUNIT [11])	+0.7	+0.4	+1.0
+Syn. (Ours)	**+1.1**	**+1.0**	**+ 1.7**

Table 6. Improvement comparison on COCO detection with different image synthetic methods.

	Metric	Percentage (%)
COCO	Det.&Seg.	↓19.1 & ↓19.0
Cityscapes	mIoU&mAcc	↓ **2.6 &** ↓**2.4**

Table 7. **Performance decline** when training and testing on real image, and comparing to results on synthetic image. We adopt PSPNet [41] with ResNet-50 [9] on Cityscapes [4] and obtain (real&real): mIoU: 76.6%, mAcc: 83.1%; (syn.&syn.): 74.6%/81.1% .

image and object. "INIT w/o D_s" denotes we build separate discriminators for image and object. Thanks to the coarse and fine styles we used, our average INIT w/ D_s score outperforms MUNIT with a notable margin. We also observe that our dataset (real image) has a very large diversity score, which indicates that the dataset is diverse and challenging.

Inception Score (IS) and Conditional Inception Score (CIS). We use the Inception Score (IS) [32] and Conditional Inception Score (CIS) [11] to evaluate our learned models. IS measures the diversity of all output images and CIS measures diversity of output conditioned on a single input image, which is a modified IS that is more suitable for evaluating multimodal I2I translation task. The detailed definition of CIS can be referred to [11]. We also employ with Inception V3 model [36] to fine-tune our classification model on four domain category labels of our dataset. Other

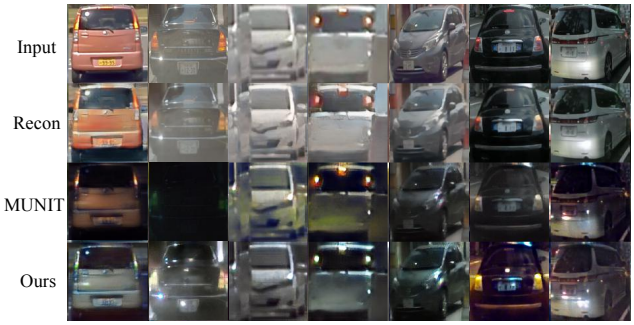


Figure 9. **Qualitative comparison on randomly selected instance level results.** The first row shows the input objects. The second row shows the self-reconstruction results. The third and fourth rows show outputs from MUNIT and ours, respectively.

settings are the same as [11]. It can be seen in Tab. 4 that our results are consistently better than the baselines MUNIT and DRIT.

Image Synthesis on Multiple Datasets The visualization of our synthetic images is shown in Fig. 7. The left group images are on COCO and the right are on Cityscapes. We observe that the most challenging problem for multiple datasets synthesis is the inter-class variance among them.

Data Augmentation for Detection & Segmentation on COCO. We use Mask RCNN [8] framework for the experiments. A synthetic copy of entire COCO dataset is generated by our sunny→night model. We employ open-source implementation of Mask RCNN[4] for training the COCO models. For training, we use the same number of training epochs and other default settings including the learning rating schedule, # batchsize, etc.

All results are summarized in Tab. 5, the first column (group) shows the training data we used, the second group shows the validation data where we tested on. The third and fourth groups are detection and segmentation results, respectively. We can observe that our real-image trained model can obtain 30.4% mAP on synthetic validation images, this indicates that the distribution differences between original COCO and our synthetic images are not very huge. It seems that our generation process is more likely to do

[4]https://github.com/facebookresearch/maskrcnn-benchmark

Figure 10. **Case-by-case comparison on sunny→night.** The first row shows the input images. The second and third rows show random outputs from MUNIT [11] and ours, respectively.

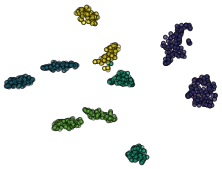
photo-metric distortions or brightness adjustment of images, which can be regarded as a data augmentation technique and has been verified the effectiveness for object detection in [22]. From the last two rows we can see that not only the synthetic images can help improve the real image testing performance, but the real image can also boost the results of synthetic images (both train and test on synthetic images). We also compare improvement with different generation methods in Tab. 6. The results show that our object branch can bring more benefits for detection task than the baseline. We also believe that the proposed data augmentation method can benefit to some limited training data scenarios like learning detectors from scratch [35, 7].

We further conduct scene parsing on Cityscapes [4]. However, we didn't see obvious improvement in this experiment. Using PSPNet [41] with ResNet-50 [9], we obtain mIoU: 76.6%, mAcc: 83.1% when training and testing on real images and 74.6%/81.1% on both synthetic images. We can see that the gaps between real and synthetic image are really small. We conjecture this case (no gain) is because the synthetic Cityscapes is too close to the original one. We compare the performance decline in Tab. 7. Since the metrics are different in COCO and Cityscapes, we use the relative percentage for comparison. The results indicate that the synthetic images may be more diverse for COCO since the decline is much smaller on Cityscapes.

5. Analysis

Qualitative Comparison. We qualitatively compare our method with baseline MUNIT [11]. Fig. 10 shows example results on sunny→night. We randomly select one output for each method. It's obvious that our results are much more realistic, diverse with higher quality. If the object area is small, MUNIT [11] may fall into mode collapse and brings small artifacts around object area, in contrast, our method can overcome this problem through instance-level reconstruction. We also visualize the multimodal results in Fig. 8 with randomly sampled style vectors. It can be observed that the various degrees of darkness are generated across these images.

Instance Generation. The results of generated instances are shown in Fig. 9, our method can generate more diverse objects (columns 1, 2, 6), more details (columns 5, 6, 7) with even the reflection (column 7). MUNIT sometimes fails to generate desired results if the global style is not suitable for the target object (column 2).



Visualization of style distribution by t-SNE [24]. The groups with the same color are paired object and global styles of same domain.

Comparison of Local (Object) and Global Style Code Distributions. To further verify our assumption that the object and global styles are distinguishable enough to disentangle, we visualize the embedded style vectors from our w/ D_s model. The visualization is plotted by t-SNE tool [24]. We randomly sample 100 images and objects in the test set of each domain, results are shown in Fig. 5. The same color groups represent the paired global images and objects in the same domain. We can observe that the style vectors of same domain global and object images are grouped and separate with a remarkable margin, meanwhile, they are neighboring in the embedded space. This is reasonable and demonstrates the effectiveness of our learning process.

6. Conclusion

In this paper, we have presented a framework for instance-aware I2I translation with unpaired training data. Extensive qualitative and quantitative results demonstrate that the proposed method can capture the details of objects and produce realistic and diverse images. Meanwhile, we also built up a large scale dataset with bounding box annotation for the instance-level I2I translation problem.

Acknowledgements Xiangyang Xue was supported in part by NSFC under Grant (No.61572138 & No.U1611461) and STCSM Project under Grant No.16JC1420400.

References

[1] Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *ICML*, 2018. 1, 2

[2] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Finding tiny faces in the wild with generative adversarial network. In *CVPR*, 2018. 5

[3] Brian Cheung, Jesse A Livezey, Arjun K Bansal, and Bruno A Olshausen. Discovering hidden factors of variation in deep networks. In *ICLR workshop*, 2015. 3

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 3, 7, 8

[5] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 3

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 5

[7] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. *arXiv preprint arXiv:1811.08883*, 2018. 8

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 7

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7, 8

[10] Xun Huang and Serge J Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 4

[11] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 1, 2, 3, 4, 6, 7, 8

[12] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 2017. 5

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 3

[14] Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. Learning to generate images of outdoor scenes from attributes and semantic layouts. *arXiv preprint arXiv:1612.00215*, 2016. 2

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 6

[16] Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics (proceedings of SIGGRAPH)*, 33(4), 2014. 3

[17] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 1, 2, 3, 4, 5, 6

[18] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 645–653. ACM, 2018. 3

[19] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 7

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[21] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017. 1, 2, 6

[22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 8

[23] Shuang Ma, Jianlong Fu, Chang Wen Chen, and Tao Mei. Da-gan: Instance-level image translation by deep attention generative adversarial networks. In *CVPR*, 2018. 3

[24] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 8

[25] Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *NIPS*, 2016. 3

[26] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1, 2

[27] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Instance-aware image-to-image translation. In *International Conference on Learning Representations*, 2019. 2

[28] Vu Nguyen, Yago Vicente, F Tomas, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. Shadow detection with conditional generative adversarial networks. In *ICCV*, 2017. 5

[29] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS workshop*, 2017. 5

[30] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 5

[31] Radim Šára Radim Tyleček. Spatial pattern templates for recognition of objects with regular structure. In *Proc. GCPR*, Saarbrucken, Germany, 2013. 3

[32] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016. 6, 7

[33] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *CVPR*, 2017. 2

[34] Zhiqiang Shen, Zhankui He, and Xiangyang Xue. Meal: Multi-model ensemble via adversarial learning. In *AAAI*, 2019. 5

[35] Zhiqiang Shen, Zhuang Liu, Jianguo Li, Yu-Gang Jiang, Yurong Chen, and Xiangyang Xue. Dsod: Learning deeply supervised object detectors from scratch. In *ICCV*, 2017. 8

[36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 7

[37] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018. 5

[38] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 5

[39] Richard Zhang, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[40] Zizhao Zhang, Lin Yang, and Yefeng Zheng. Translating and segmenting multimodal medical volumes with cycle-and shapeconsistency generative adversarial network. In *CVPR*, 2018. 1

[41] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 7, 8

[42] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 1, 2, 3, 6

[43] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, 2017. 2