DUNIT: Detection-based Unsupervised Image-to-Image Translation

Deblina Bhattacharjee, Seungryong Kim, Guillaume Vizier, Mathieu Salzmann
School of Computer and Communication Sciences, EPFL, Switzerland
{deblina.bhattacharjee, seungryong.kim, guillaume.vizier, mathieu.salzmann}@epfl.ch

Abstract

Image-to-image translation has made great strides in recent years, with current techniques being able to handle unpaired training images and to account for the multimodality of the translation problem. Despite this, most methods treat the image as a whole, which makes the results they produce for content-rich scenes less realistic. In this paper, we introduce a Detection-based Unsupervised Image-to-image Translation (DUNIT) approach that explicitly accounts for the object instances in the translation process. To this end, we extract separate representations for the global image and for the instances, which we then fuse into a common representation from which we generate the translated image. This allows us to preserve the detailed content of object instances, while still modeling the fact that we aim to produce an image of a single consistent scene. We introduce an instance consistency loss to maintain the coherence between the detections. Furthermore, by incorporating a detector into our architecture, we can still exploit object instances at test time. As evidenced by our experiments, this allows us to outperform the state-of-the-art unsupervised image-to-image translation methods. Furthermore, our approach can also be used as an unsupervised domain adaptation strategy for object detection, and it also achieves state-of-the-art performance on this task.

1. Introduction

Image-to-image translation (I2I) has recently gained significant traction, to the point of being deployed in diverse applications, such as super-resolution [20], photo-realistic image synthesis [37, 29], colorization [19, 41] and domain adaptation [2]. This trend was initiated by the pioneering Pix2Pix work [15] that used conditional generative adversarial networks (GANs) [27] on paired training images. Since then, great progress has been made in this field, first by removing the requirement for the training images to be paired, leading to cycleGAN [45] and UNIT [25], and then by accounting for the inherent multimodality of the I2I task, both with paired [44] and unpaired [12, 22] images.

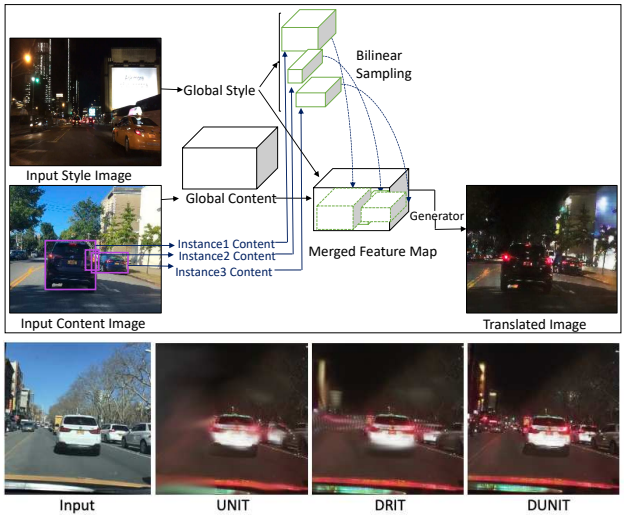

Figure 1: Overview of DUNIT. (Top) We combine the style of one domain with the image-level and instance-level content of the other. Instance-level and image-level features are extracted separately and fused to generate a single consistent image. (Bottom) By accounting for the instances, our method produces more realistic results than image-level translation techniques such as UNIT [25] and DRIT [22].

While these methods have demonstrated promising results, they all consider the I2I task as a global translation problem over the whole image. Consequently, they remain limited in their ability to translate content-rich images with many disparate object instances. INIT [34] and InstaGAN [28] address this issue by treating the object instances and the global image/background separately. While InstaGAN aims to preserve the background style when translating the instances, INIT targets the same goal as us, and as the previous methods, of translating the entire image. To achieve this, INIT independently translates the global image and the instances, using separate reconstruction losses on these different elements. At test time, INIT then uses the global image translation module only, thus discarding the instance-level information. Further, INIT has not used the instance-boosted feature representation which is shown by

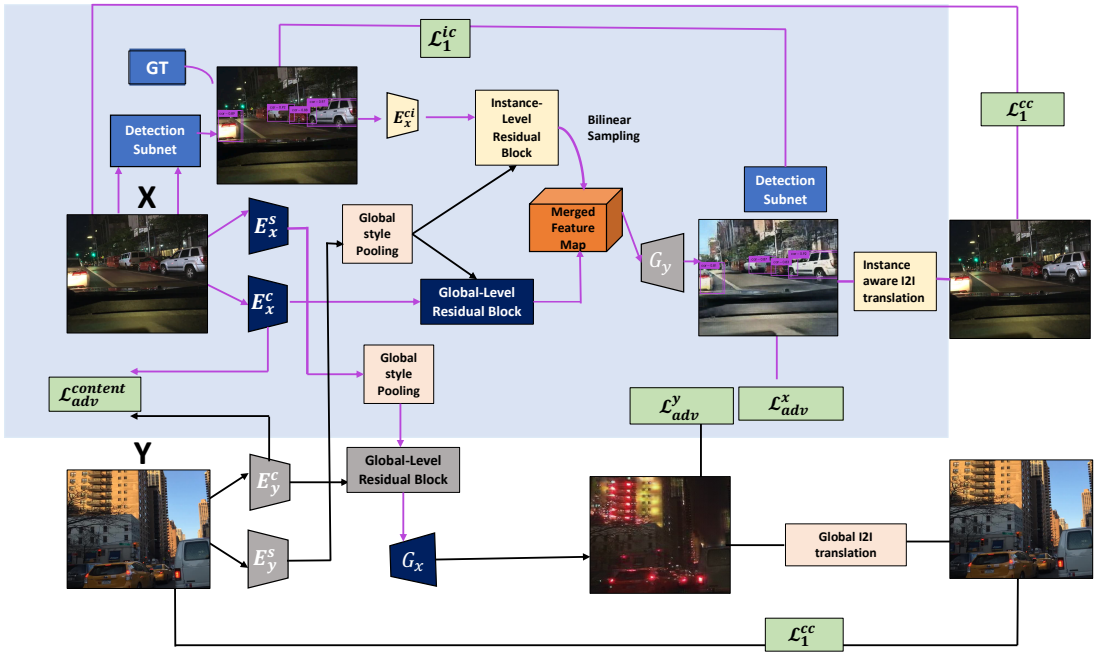

Figure 2: Overall DUNIT architecture. The instance aware I2I translation block on the right is the exact replica of the operations taking place between the night image in domain X and the corresponding translated day image. Similarly, the global I2I translation block mirrors the operations between the day image in domain Y and its translated night image.The blue background separates our contribution from the DRIT [22] backbone on which our work is built. The pink lines correspond to domain X and the black lines to domain Y. The global-level residual blocks have different features in domain X and domain Y and hence are color coded differently. The global features in domain X are shown in dark blue, those in domain Y in dark grey, the losses are in green, the global operations are in light orange, the instance features in domain X are in yellow, the detection subnetwork in light blue and the merged features in dark orange.

the merged feature map in figure 2.

In this paper, we introduce a framework that, while accounting for the object instances, unifies their translation with that of the global image as seen in figure 1, so that instances can also be leveraged at test time. To this end, we incorporate an object detector within the I2I architecture, process the instance features and the image ones separately, and reassemble the resulting representations in a joint one, which we decode into a single image. Processing the features separately allows us to account for the detailed and diverse content of the different objects and of the background, yet fusing the instance-level and image-level representation models the fact that we aim to translate one consistent scene. To further exploit the detections during training, we introduce an instance-consistency loss, which compares the detections in the original image with that in the translated one. At test time, we follow the same process as during training, consisting of detecting the object instances, processing them independently of the global image and fusing the resulting representations to generate the final image.

Our main contributions therefore are as follows:

- We improve unsupervised I2I translation by introducing a detection-based translation method that, while processing the object instances and global image separately, fuses their representations so as to produce a single consistent image.
- We introduce an instance-consistency loss, which further exploits the instances during training by modeling the intuition that the detections in the original and translated images should be consistent.
- By incorporating the detector into the architecture, we explicitly reason about instances not only during training but also at test time.
- During training, we only need access to ground-truth detections in a single domain. Therefore, our method can also be thought of as performing unsupervised domain adaptation for object detection.

Our experiments on the standard INIT, Pascal-VOC classes [7], Comic2k [14], cityscapes [4] and KITTI [9] benchmark show that our approach outperforms the state-of-the-art global I2I methods, as well as the instance-level INIT one. Furthermore, we demonstrate that our approach also outperforms the state-of-the-art unsupervised domain adaptation detection algorithms.

2. Related Work

2.1. Image to Image Translation

The advent of I2I translation methods began with the invention of conditional GAN[27], which were first applied to learn a mapping between a source and a target domain in Pix2Pix [15]. Since then, conditional GANs have been applied to a multitude of tasks, such as scene translation [13], season transfer [25], and sketch-to-photo translation [35]. While conditional GANs yield impressive results, they require paired images during training. Unfortunately, in many I2I translation scenarios, such paired training data is difficult and expensive to collect. To overcome this, cycleGAN [45] introduces a cycle consistency loss between the source and target domains, following the intuition that translating an image from the source domain to the target one and then back to the source one should yield consistent images. This idea was further extended by UNIT [25], which replaced the domain-specific latent spaces of cycleGAN with a single latent space shared by the domains.

Nevertheless, neither conditional GANs, nor cycleGAN, nor UNIT account for the multi-modality of I2I translation; in general, a single image in one domain can be translated to another in many different, yet equally realistic ways. This was the task addressed by BicycleGAN [44], yet by leveraging paired images during training. More recently, MUNIT [12] and DRIT [22] introduced solutions to the multi-modal, unpaired scenario by learning a disentangled representation with a domain-invariant content space and a domain-specific attribute/style space.

While effective, all the above-mentioned methods perform image-level translation, without considering the object instances. As such, they tend to yield less realistic results when translating complex scenes with many objects, such as traffic scenes. InstaGAN [28] was the first work to tackle instance-level translation. To this end, InstaGAN takes as input objects' segmentation masks in the two domains of interest, and performs translation between the object instances only, while maintaining the background unchanged. Here, by contrast, we aim to translate the entire image, object instances and background included. This is also the task addressed by INIT [34], which proposed to define a style bank to translate the instances and the global image separately. During training, however, INIT treats the object instances and the global image completely independently, each one having its own reconstruction loss. As such, at test time, it does not exploit the object instances at all, thus going back to image-level translation. Here, we propose to unify the translation of the image and its instances, thus allowing us to leverage the object instances at test time. As will be shown in our experiments, this results in more realistic translations and further allows us to perform unsupervised domain adaptation.

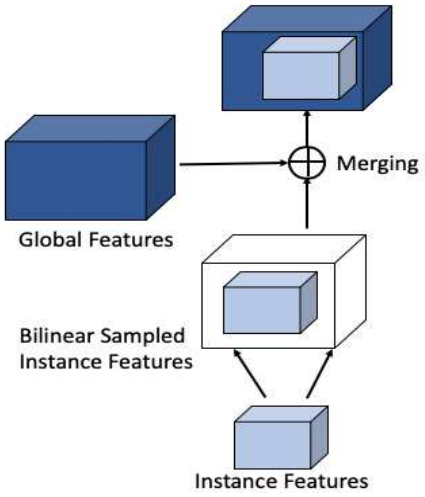

Figure 3: Merging the global and instance-level features. We use bilinear sampling [16] to position the instance-level features of each object at the corresponding location in the global feature map. We then merge this representation with the global one by, in essence, overwriting the global features at the object locations.

2.2. Domain Adaptation

Domain adaptation has been extensively studied in computer vision, using, e.g., multiple kernel learning [5, 6], subspace alignment [8] and covariance matrix alignment [38]. More recently, I2I translation has also been exploited for this task [17, 40, 39, 43, 11, 3, 23]. By contrast, the literature on domain adaptation for object detection remains sparser [26, 1], as the task is inherently more challenging due to its additional localization property. In [39] a deformable part-based model based on adaptive support vector machines was introduced to overcome domain shift for object detection. More recently, [30] used subspace alignment to transform the features extracted from two domains with a region-based convolutional neural network (RCNN) [10] to a common representation. These methods, however, are not end-to-end trainable. This was remedied in [2] via a domain adaptive Faster-RCNN object detection network reasoning jointly at image and instance level. Also, the weakly supervised cross domain adaptive paper [14] introduces a domain transfer stage where the detections on cycleGAN generated images are fine tuned. Furthermore, in [18], a domain adaptive representation learning technique is introduced which diversifies the shifted domain. All these methods use Faster-RCNN [31] as the detector to adapt between domains on a global level. While we jointly leverage the global image and the object instances, we do so in an I2I translation formalism, which we will show yields better results.

3. Methodology

3.1. Problem Formulation and Overview

We aim to learn a multi-modal mapping between two visual domains $X \subset \mathbb{R}^{H\times W\times 3}$ and $Y \subset \mathbb{R}^{H\times W\times 3}$ that jointly accounts for the global image and the object instances. To this end, we build our approach on the DRIT backbone [22], which handles multi-modality but does not reason about instances. Our architecture is depicted in Figure 2. We assume that, during training, we have access to the ground-truth bounding boxes in a *single* domain, X, i.e., `Night` in Figure 2. At test time, however, we do *not* require access to ground-truth object bounding boxes; they will be predicted by our network. We now explain the components of our network in more detail.

3.2. Image-to-Image Translation Module

Our architecture comprises 2 style encoders $\{E^s_x, E^s_y\}$, one for each domain, and 3 content encoders $\{E^c_x, E^{ci}_x, E^c_y\}$, two to process the global images in each domain (E^c_x, E^c_y), and an additional one (E^{ci}_x) for the instances in domain X. Let us now consider domain Y for which we do not have instances. We use E^c_y to extract global content features from an image I_y in domain Y, which is then merged with style features extracted from an image I_x in domain X using E^s_x. The resulting representation is then passed through a decoder G_x that generates a translation of I_y to domain X. In the standard DRIT framework, this process is mirrored to go from domain X to domain Y.

3.3. Object Detection Module

We aim to take the object instances into account. To this end, we first detect object instances in I_x using a detection subnetwork, and then process the global image and the object instances in parallel, before fusing their representations and achieving translation to domain Y using a decoder G_y. Specifically, at the image level, we extract global content features from I_x using E^c_x and merge these features with the style features extracted from I_y using E^s_y. This is achieved using the global-level residual block depicted in Figure 2, whose architecture follows that of the residual block used in DRIT [21]. Let us denote by F_x the global feature map resulting from this operation.

In parallel as we extract F_x, we also process each individual object detection. To this end, for each instance i, we first crop the corresponding image region, which can be done in an ROI pooling manner, and extract instance level content features using E^{ci}_x. We then merge these features with the globally-pooled style features extracted from I_y using the instance-level residual block shown in Figure 2, whose architecture is the same as the global-level one. Let $\{F^i_x\}$ denote the set of feature maps obtained in this manner, with one feature map F^i_x for each object instance.

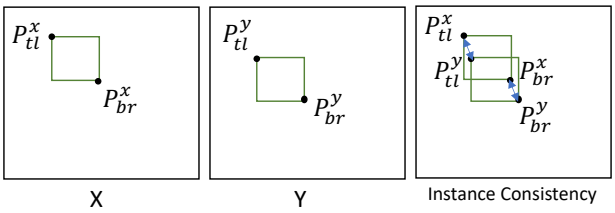

Figure 4: Visualization of the Instance Consistency loss, which is the summation of the l_1 distance between the top-left pixels of the detected bounding boxes in domains X and Y with the l_1 distance between the bottom-right pixels of the detected bounding boxes in domains X and Y.

Instead of decoding separately the image-level features F_x and the instance-level ones $\{F^i_x\}$, in a similar manner as in INIT, here we propose to fuse them so as to obtain a single representation that we can decode into one consistent image using G_y. To achieve this, we need to reintroduce the instance features at their respective locations in the global feature map. This process is illustrated in Figure 3. In essence, we make use of the bilinear sampling strategy introduced in [16] to place the instance features in a map of the same size as F_x, and then merge this map with F_x by simply replacing the features in F_x at non-empty locations in the map with those in the map.

3.4. Training

To be able to handle unpaired training images, we follow the cycle-consistency approach and translate the generated images back to their original domains. In essence, this process mirrors that described above; it uses the instances to go back to domain X and acts at the global image level to generate the domain Y image. Below, we detail the loss function and training procedure for the resulting Detection-based Unsupervised Image-to-image Translation (DUNIT) model.

Image-to-image translation module. Our method is built on the DRIT backbone which embeds the input images onto a shared style space and a domain specific content space. As such, we use the same weight-sharing strategy as DRIT for the two style encoders (E^s_x, E^s_y) and exploit the same loss terms. They include:

- A content adversarial loss $\mathcal{L}^{content}_{adv}(E^c_x, E^c_y, D^c)$ relying on a content discriminator D^c, whose goal is to distinguish the content features of both domains;
- Domain adversarial losses $\mathcal{L}^x_{adv}(E^c_y, E^s_x, G_x, D^x)$ and $\mathcal{L}^y_{adv}(E^c_x, E^{ci}_x, E^s_y, G_y, D^y)$, one for each domain, with corresponding domain classifiers D^x and D^y;

- A cross-cycle consistency loss $\mathcal{L}_1^{cc}(G_x, G_y, E_x^c, E_x^{ci}, E_y^c, E_x^s, E_y^s)$ that exploits the disentangled content and style representations for cyclic reconstruction [36];
- Self-reconstruction losses $\mathcal{L}_{rec}^x(E_x^c, E_x^{ci}, E_x^s, G_x)$ and $\mathcal{L}_{rec}^y(E_y^c, E_y^s, G_y)$ ensuring that the generators can reconstruct samples from their own domain;
- KL losses $\mathcal{L}_{KL}^x(E_x^s)$ and $\mathcal{L}_{KL}^y(E_y^s)$ encouraging the distribution of the style representations to be close to a standard normal distribution;
- Latent regression losses $\mathcal{L}_{lat}^x(E_x^c, E_x^{ci}, E_x^s, G_x)$ and $\mathcal{L}_{lat}^y(E_y^c, E_y^s, G_y)$ encouraging the mappings between the latent style representation and the image to be invertible.

Note that, in contrast to DRIT, here, several of these loss terms exploit the instance content encoder E_x^{ci}.

Object detection module. In addition to all the DRIT losses listed above, we introduce a new instance consistency loss that explicitly reasons about the detected objects. The intuition behind this loss is that the same object instances should be detected in I_x and in the corresponding image after translation. Enforcing consistency between the detections in two images raises the question of how to match these detections. To overcome this, we exploit a modern detector that relies on fixed detection anchors. Since the set of anchors is the same in both images, the detections are naturally paired. We then aim for the positive detections at the same anchors in both images to correspond to the same object, and thus have the same bounding boxes. To the end, we define the instance consistency loss

$$\mathcal{L}_1^{ic} = \sum_{i|\hat{y}_i^x=1 \wedge \hat{y}_i^y=1} |P_{tl}^{xi} - P_{tl}^{yi}|_1 + \sum_{i|\hat{y}_i^x=1 \wedge \hat{y}_i^y=1} |P_{br}^{xi} - P_{br}^{yi}|_1 \,, \tag{1}$$

where $\hat{y}_i^x = 1$ indicates that anchor i is predicted as positive in domain X and similarly $\hat{y}_i^y$ in domain Y, P_{tl}^{xi} and P_{tl}^{yi} are the bounding box top-left pixels for anchor i in domain X and domain Y, respectively, and P_{br}^{xi} and P_{br}^{y} the corresponding the bottom-right pixels.

In our experiments, we make use of the RetinaNet detector [24]. We employ the focal loss of [24] for the instance detections. The focal loss further introduces parameters α to offset the class imbalance. In our experiments, we use $\alpha = [0.25, 0.5]$ as in [24] and $\gamma = 2$. We will study the influence of these parameters in our experiments. Note that we apply the RetinaNet on both I_x and its translated version in domain Y. We therefore use the focal loss in both cases, which further encourages the translated image to contain object instances at the same locations as the input image, since, during training, we compare all detections to the ground-truth bounding boxes.

4. Experiments and Results

4.1. Unsupervised Image to Image Translation

To validate our method, we conduct experiments on the INIT [34] dataset which is a very diverse and challenging dataset. Following DRIT, we resize the images in the dataset to 216×216 to consider GPU limitations. We used 4 GPUs to train our model with a batch-size of 16. To evaluate our method, we compare it with the following five state-of-the-art unpaired I2I translation approaches.

- CycleGAN [45], which comprises forward and backward translation functions between the source and target domains along with an adversarial loss.
- UNIT [25], which improves upon CycleGAN by using a shared latent space and comprises two VAE-GANs and a cycle-consistency loss.
- MUNIT [12], which assumes that an image representation can be disentangled into a domain specific style representation and a domain invariant content representation and swaps these disentangled content/style latent features to generate the translations.
- DRIT [22], which is very similar to MUNIT except that it contains generators and domain discriminators in both the domains, along with the two content encoders and two style encoders.
- INIT [34], which is built on the MUNIT backbone and considers instance level style translations along with the global translation. It uses the cross-cycle consistency loss, a global and instance level GAN loss and a global and instance level reconstruction loss.

We implemented our method in PyTorch and will make our code available at https://github.com/IVRL/DUNIT.

To evaluate these methods, we use the following three standard performance metrics.

- Inception Score (IS) [33], which encodes the diversity across all translated outputs.
- Conditional Inception Score (CIS) [12], which encodes the diversity of the translated output conditioned on a single input image, and is typically used for multimodal methods.
- LPIPS distance [42], which measures the diversity of the translated images, and has been shown to be strongly correlated with human perception. To compute this metric, following the setting used in [34], we randomly sample 19 pairs of translated outputs from the 100 input test images.

	CycleGAN [45]		UNIT [25]		MUNIT [12]		DRIT [22]		INIT [34]		DUNIT	
	IS	**CIS**	**IS**	**CIS**	**IS**	**CIS**	**IS**	**CIS**	**IS**	**CIS**	**IS**	**CIS**
Sunny →Night	1.026	0.014	1.030	0.082	1.278	1.159	1.224	1.058	1.118	1.060	1.259	1.166
Night →Sunny	1.023	0.012	1.024	0.027	1.051	1.036	1.099	1.024	1.080	1.045	1.108	1.083
Sunny →Rainy	1.073	0.011	1.075	0.097	1.146	1.012	1.207	1.007	1.152	1.036	1.225	1.029
Rainy →Sunny	1.090	0.010	1.023	0.014	1.102	1.055	1.103	1.028	1.119	1.060	1.125	1.083
Sunny →Cloudy	1.097	0.014	1.134	0.081	1.095	1.008	1.104	1.025	1.142	1.040	1.149	1.033
Cloudy →Sunny	1.033	0.090	1.046	0.219	1.321	1.026	1.249	1.046	1.460	1.016	1.472	1.077
Average	1.057	0.025	1.055	0.087	1.166	1.032	1.164	1.031	1.179	1.043	**1.223**	**1.079**

Table 1: Quantitative comparison of our approach with the state of the art on the INIT dataset. We report the Inception Score (IS) and Conditional Inception Score (CIS) (higher is better). DUNIT with the Instance Consistency Loss (IC) gives the best results.

Method	**Diversity of results (LPIPS Distance)**			
	Sunny →Night	Sunny →Rainy	Sunny →Cloudy	Average
UNIT [25]	0.067	0.062	0.068	0.066
CycleGAN [45]	0.016	0.008	0.011	0.012
MUNIT [12]	0.292	0.239	0.211	0.247
DRIT [22]	0.231	0.173	0.166	0.190
INIT [34]	0.330	0.267	0.224	0.274
DUNIT	**0.338**	**0.298**	**0.225**	**0.287**
Real Images	0.573	0.489	0.465	0.509

Table 2: Quantitative comparison of our approach with the state of the art in terms of image diversity. Following [21], we report the LPIPS metric. Note that DUNIT yields the highest diversity score.

Figure 5: Qualitative comparison on Sunny to Night. We show, from left to right, the input image in the source domain, the result of cycleGAN [45] and UNIT [25], and random outputs from MUNIT [12], DRIT [22] and DUNIT (ours), respectively.

We provide the IS and CIS in Table 1 and the LPIPS in Table 2. DUNIT outperforms the baselines on all pairs of domains and according to all metrics, with the exception of IS on Sunny→Night, where MUNIT yields a slightly higher score. On average, we outperform the baselines by a comfortable margin, including INIT which also exploits object instances. In Fig. 5, we compare qualitatively the different methods. Note that DUNIT yields sharper and more realistic images than the baselines. We do not include INIT in this qualitative comparison because its code is not publicly available.

Ablation study. We now evaluate different aspects of our method. First, we study the influence of the instance consistency loss. To this end, we compare the results obtained using our method with the instance consistency loss (DUNIT w/ IC) and without it (DUNIT w/o IC). We report the LPIPS distance and the IS and CIS metrics over three pairs of domains from the INIT dataset in Table 3. Note that the IC loss consistently improves the results of DUNIT on all pairs. This demonstrates the benefits of constraining the content in the translated image to preserve the instances in the input image. We believe this to be an instance of a more general phenomenon, where auxiliary tasks can help to improve the translation process. We then turn to exploring the choice of detection subnetwork in our architecture. In addition to the RetinaNet used in our previous experiments, we also evaluate the Faster-RCNN. Note that, in this case, the detections are not paired, and we cannot use the

Figure 6: Qualitative comparisons conditioned on one image style for Sunny to Cloudy (first row) and Sunny to Rainy (second row). We show, from left to right, the input image in the source domain, the style image for translation, followed by outputs from MUNIT [12], DRIT [22] and DUNIT (ours), respectively. We only show multimodal results as they perform better than unimodal methods like CycleGAN [45] and UNIT [25] as seen in figure 5.

	Effect of the IC loss					
	w/ L_i^{ic}	w/o L_i^{ic}	w/ L_i^{ic}		w/o L_i^{ic}	
	LPIPS	LPIPS	IS	CIS	IS	CIS
Sunny to Night	0.338	0.322	1.259	1.166	1.216	1.049
Sunny to Rainy	0.298	0.252	1.225	1.029	1.138	1.002
Sunny to Cloudy	0.225	0.203	1.149	1.033	1.108	1.009
Average	**0.287**	**0.259**	**1.211**	**1.076**	**1.154**	**1.020**

Table 3: Ablation Study: We compare our method with the instance consistency loss (DUNIT w/ L_i^{ic}) and without it (DUNIT w/o L_i^{ic}). We report the LPIPS distance, inception score (IS) and conditional inception score (CIS).

Figure 7: Predicted detections for an input day image (left) using RetinaNet (middle), Faster-RCNN (right). Note that the higher mAP of RetinaNet yields better instance translations.

IC loss. Furthermore, with RetinaNet, we evaluate different hyperparameter settings for $\alpha \in [0, 1]$ and $\gamma \in \{1, 2\}$. From Table 4, we can see that RetinaNet with $\alpha = 0.25$ and $\gamma = 2$ gives the best IS and CIS across the tested domains. Changing α while keeping $\gamma = 2$ yields very close results, but decreasing γ to 1 results in lower scores. Nevertheless, RetinaNet consistently outperforms Faster-RCNN. In Fig. 7, we compare qualitatively the detections obtained with RetinaNet and Faster-RCNN, and their impact on the translated image; the higher detection accuracy of RetinaNet translates to higher quality images. Note that the relatively low performance of our approach with a Faster-RCNN detector is due to the detector itself. In particular, the fact that Faster-RCNN is a two-stage detector prevents a complete end-to-end training and means that the region proposals in the original and translated image may not match. This resulted in less robust detections than RetinaNet, and we observed these non-robust detections to impede the minimization of the instance consistency loss during training. Ultimately, this contributed to incorporating incorrect instance information in the global feature map, thus yielding worse results than when using a RetinaNet that can be trained end-to-end and relies on anchors that are naturally paired. This shows that the results of our approach depend on the detector, but that state-of-the-art, single-stage detectors already achieve sufficient accuracy for us to outperform global image translation.

4.2. Unsupervised Domain Adaptive Detection

We further test our method on the task of unsupervised domain adaptation for object detection. We use as baselines the state-of-the-art methods that tackle this task, namely, domain adaptive Faster RCNN [2], the domain transfer stage of [14], the shifted domain stage of [18] and the style transfer with feature consistency stage of [32]. We conduct experiments on diverse datasets including Pascal VOC classes [7] as source domain with Comics2K [14] as target domain, and the Kitty object detection benchmark [9] as source domain with Cityscapes [4] as target domain. We follow the same data preparation and the same experimental setup as in [18]. Note that the inter-class variance and the large difference in data distribution between the source and target domains make these dataset very challenging for synthesis. In Table 5, we report the mean average pre-

	Object Detection Methods							
	Faster-RCNN		**RetinaNet** $\alpha = 0.25$, $\gamma = 2$		**RetinaNet** $\alpha = 0.50$, $\gamma = 2$		**RetinaNet** $\alpha = 1.0$, $\gamma = 1.0$	
	IS	CIS	IS	CIS	IS	CIS	IS	CIS
Sunny →Night	1.223	1.058	1.259	1.166	1.255	1.163	1.230	1.104
Sunny →Rainy	1.208	1.008	1.225	1.029	1.223	1.025	1.213	1.017
Sunny →Cloudy	1.104	1.025	1.149	1.033	1.144	1.031	1.113	1.027
Average	**1.178**	**1.030**	**1.211**	**1.076**	**1.207**	**1.073**	**1.185**	**1.049**

Table 4: Ablation Study: We compare our method (DUNIT) used in conjunction with different object detection subnetworks (Faster R-CNN or RetinaNet). For RetinaNet, we report the results obtained with different values of the hyperparameters α and γ. We report the inception score (IS) and conditional inception score (CIS). We use $\alpha = 0.25$ and $\gamma = 2$ for our best model.

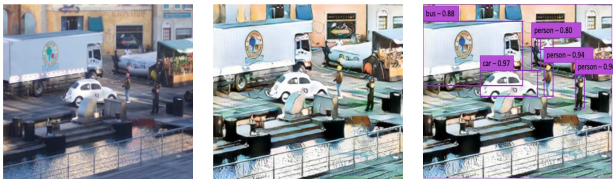

Figure 8: Qualitative domain adaptation results. We translate Pascal VOC images to the Comics2k domain using DUNIT and apply a detector trained on the original Comics2k data to the translated images. (Left) Input image, (Middle) translated image and (Right) detections.

Methods	VOC →Comics	KITTI →Citysc.
DT [14]	23.5	31.2
DAF [2]	23.2	38.5
DARL [18]	34.5	45.3
DAOD [32]	36.4	46.1
DUNIT w/L_i^{ic}	**40.2**	**54.1**
DUNIT wo/L_i^{ic}	**39.4**	**47.2**

Table 5: Quantitative comparison on the task of domain adaptation. We report the mAP for two pairs of domains. DT is the domain transfer stage in [14], DAF is the domain adaptive faster-rcnn method [2] and DARL is the domain adaptive representation learning method [18].

cision (mAP) for the detected objects across the different domains and, in Table 6, we detail the per-class average precisions (AP) for the KITTI→Cityscapes case. Furthermore, in Fig. 8, we show qualitative detection examples the VOC→Comics case. In this set of experiments, we used Faster-RCNN [31] as detector on the translated outputs of our approach because all the above-mentioned baselines rely on this detector. Note that our model significantly outperforms the baselines. Note that the robust pseudo-labeling technique proposed in the baselines above, could be incorporated in our approach and can further boost the performance. Furthermore, the gap is consistent across all classes. This suggests that translating images between domains with our approach is more effective than aiming to learn domain-invariant representations.

Methods	Pers.	Car	Truc.	Bic.	mAP
DT [14]	28.5	40.7	25.9	29.7	31.2
DAF [2]	39.2	40.2	25.7	48.9	38.5
DARL [18]	46.4	58.7	27.0	49.1	45.3
DAOD [32]	47.3	59.1	28.3	49.6	46.1
DUNIT w/L_i^{ic}	**60.7**	**65.1**	**32.7**	**57.7**	**54.1**
DUNIT wo/L_i^{ic}	**56.2**	**59.5**	**24.9**	**48.2**	**47.2**

Table 6: Quantitative comparison of the per-class Average Precision for the KITTI→Cityscapes adaptation scenario.

5. Conclusion

We have introduced an approach to account for object instances when translating images between domains. To this end, we have proposed to process the instances and the global image separately, but to fuse their respective representations so as to generate a single consistent image. This has allowed us to translate content-rich images, resulting in realistic images that quantitatively outperform the state-of-the-art I2I translation algorithms. By only requiring access to ground-truth object bounding boxes for a single domain during training, our approach has also allowed us to perform unsupervised domain adaptation for object detection, again producing state-of-the-art results.

In our I2I translation experiments, our instance consistency loss has proven to be important to obtain realistic results. We believe, however, that this loss is just an instance of a broader idea: One can enforce consistency between the outputs of any auxiliary task to help the I2I translation process. In the future, we will therefore investigate the use of other auxiliary tasks, such as instance segmentation, surface depth and normal prediction, and object pose estimation.

Acknowledgement. This work was supported in part by the Swiss National Science Foundation via the Sinergia grant CRSII5−180359.

References

[1] Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulò. Autodial: Automatic domain alignment layers. *CoRR*, abs/1704.08082, 2017. 3

[2] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster R-CNN for object detection in the wild. *CoRR*, abs/1803.03243, 2018. 1, 3, 7, 8

[3] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency, 2020. 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 7

[5] Lixin Duan, Ivor Tsang, and Dezhi Xu. Domain transfer multiple kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:1 – 1, 05 2011. 3

[6] Lixin Duan, Dong Xu, Ivor Wai-Hung Tsang, and Jiebo Luo. Visual event recognition in videos by learning from web data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1667–1680, 2011. 3

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 2, 7

[8] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pages 2960–2967, 2013. 3

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2, 7

[10] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013. 3

[11] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *ICML*, abs/1711.03213, 2018. 3

[12] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. *ECCV*, abs/1804.04732, 2018. 1, 3, 5, 6, 7

[13] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36:1–14, 07 2017. 3

[14] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. *CoRR*, abs/1803.11365, 2018. 2, 3, 7, 8

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016. 1, 3

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *CoRR*, abs/1506.02025, 2015. 3, 4

[17] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *CoRR*, abs/1703.05192, 2017. 3

[18] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. *CoRR*, abs/1905.05396, 2019. 3, 7, 8

[19] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. *CoRR*, abs/1603.06668, 2016. 1

[20] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016. 1

[21] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. *CoRR*, abs/1808.00948, 2018. 4, 6

[22] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision*, 2018. 1, 2, 3, 4, 5, 6, 7

[23] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation, 2019. 3

[24] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *CoRR*, abs/1708.02002, 2017. 5

[25] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *CoRR*, abs/1703.00848, 2017. 1, 3, 5, 6, 7

[26] Hao Lu, Lei Zhang, Zhiguo Cao, Wei Wei, Ke Xian, Chunhua Shen, and Anton van den Hengel. When unsupervised domain adaptation meets tensor representations. *CoRR*, abs/1707.05956, 2017. 3

[27] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. 1, 3

[28] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Instagan: Instance-aware image-to-image translation. *CoRR*, abs/1812.10889, 2018. 1, 3

[29] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. *CoRR*, abs/1903.07291, 2019. 1

[30] Anant Raj, Vinay P. Namboodiri, and Tinne Tuytelaars. Subspace alignment based domain adaptation for RCNN detector. *CoRR*, abs/1507.05578, 2015. 3

[31] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015. 3, 8

[32] Adrian Lopez Rodriguez and Krystian Mikolajczyk. Domain adaptation for object detection via style consistency, 2019. 7, 8

[33] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *CoRR*, abs/1606.03498, 2016. 5

[34] Zhiqiang Shen, Mingyang Huang, Jianping Shi, Xiangyang Xue, and Thomas S. Huang. Towards instance-level image-to-image translation. *CoRR*, abs/1905.01744, 2019. 1, 3, 5, 6

[35] Radim Tyleček and Radim Šára. Spatial pattern templates for recognition of objects with regular structure. In *German Conference on Pattern Recognition*, pages 364–374. Springer, 2013. 3

[36] Chengjia Wang, Gillian Macnaught, Giorgos Papanastasiou, Tom MacGillivray, and David E. Newby. Unsupervised learning for cross-domain medical image synthesis using deformation invariant cycle consistency networks. *CoRR*, abs/1808.03944, 2018. 5

[37] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. *CoRR*, abs/1711.11585, 2017. 1

[38] Yifei Wang, Wen Li, Dengxin Dai, and Luc Van Gool. Deep domain adaptation by geodesic distance minimization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2651–2657, 2017. 3

[39] Jiaolong Xu, Sebastian Ramos, David Vázquez, and Antonio Manuel López Peña. Domain adaptation of deformable part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:2367–2380, 2014. 3

[40] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. *CoRR*, abs/1704.02510, 2017. 3

[41] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. *CoRR*, abs/1603.08511, 2016. 1

[42] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CoRR*, abs/1801.03924, 2018. 5

[43] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017. 3

[44] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *CoRR*, abs/1711.11586, 2017. 1, 3

[45] 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, 3, 5, 6, 7