Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Cycle-object consistency for image-to-image domain adaptation

Che-Tsung Lin[a], Jie-Long Kew[b], Chee Seng Chan[b,*], Shang-Hong Lai[c,d], Christopher Zach[a]

[a] Department of Electrical Engineering, Chalmers University of Technology, Sweden
[b] CISiP, Faculty of Comp Sci. and Info. Tech., Universiti Malaya, Malaysia
[c] Microsoft AI R&D Center, Taiwan
[d] Department of Computer Science, National Tsing Hua University, Taiwan

ARTICLE INFO

Article history:
Received 15 December 2021
Revised 10 October 2022
Accepted 10 February 2023
Available online 13 February 2023

Keywords:
Generative adversarial networks
Instance-aware image-translation
Domain adaptation
Cross-domain object detection

ABSTRACT

Recent advances in generative adversarial networks (GANs) have been proven effective in performing domain adaptation for object detectors through data augmentation. While GANs are exceptionally successful, those methods that can preserve objects well in the image-to-image translation task usually require an auxiliary task, such as semantic segmentation to prevent the image content from being too distorted. However, pixel-level annotations are difficult to obtain in practice. Alternatively, instance-aware image-translation model treats object instances and background separately. Yet, it requires object detectors at test time, assuming that off-the-shelf detectors work well in both domains. In this work, we present AugGAN-Det, which introduces Cycle-object Consistency (CoCo) loss to generate instance-aware translated images across complex domains. The object detector of the target domain is directly leveraged in generator training and guides the preserved objects in the translated images to carry target-domain appearances. Compared to previous models, which e.g., require pixel-level semantic segmentation to force the latent distribution to be object-preserving, this work only needs bounding box annotations which are significantly easier to acquire. Next, as to the instance-aware GAN models, our model, AugGAN-Det, internalizes global and object style-transfer without explicitly aligning the instance features. Most importantly, a detector is not required at test time. Experimental results demonstrate that our model outperforms recent object-preserving and instance-level models and achieves state-of-the-art detection accuracy and visual perceptual quality.

1. Introduction

Recent progress in the domain of object detection has led to a remarkable performance improvement, particularly for one-stage object detectors, which provide a good balance between detection speed and accuracy. This is achieved with sophisticated training strategies such as data augmentation [1] to increase the variability of the input images, so that the object detector has better robustness on e.g., those images obtained in different environments. However, as shown by Braun et al. [2], Yu et al. [3], the overall detection performance still drops significantly when the trained detector model is deployed in a new domain different from the (augmented) training set. A natural solution to this limitation is to perform image-to-image translation for the labeled data in a source domain (e.g., daytime images) to a target domain (e.g., nighttime images).

A popular solution is CycleGAN [4] that performed unpaired image-to-image translation with the introduction of cycle consistency in Generative Adversarial Networks (GANs) [5]. It encourages bi-directional image translation with regularized structural output. Since then, various works [6–9] have been proposed and achieved impressive results in image translation tasks, such as horse ↔ zebra, vangogh ↔ photo, and cat ↔ dog.

However, these existing methods are prone to fail at preserving the objects, as illustrated in Fig. 1. That is, existing solutions [10–13] with explicit object preservation may retain the objects, but their appearance might not be able to adapt sufficiently to the target domain. Recently, instance-aware image translation models [14,15] aim to improve this issue by aligning instance features using either detection labels or an off-the-shelf object detector for the generators. For instance, INIT [14] employed both the instance and global styles to guide the generation of the target domain objects. Unfortunately, their model neglects the instance-level information at the test time and only utilizes the global information. DUNIT [15] applied an off-the-shelf general object detector and an instance-level encoder to extract instance-boosted features during

[*] Corresponding author.
E-mail addresses: cs.chan@um.edu.my (C.S. Chan), shlai@microsoft.com (S.-H. Lai).

https://doi.org/10.1016/j.patcog.2023.109416

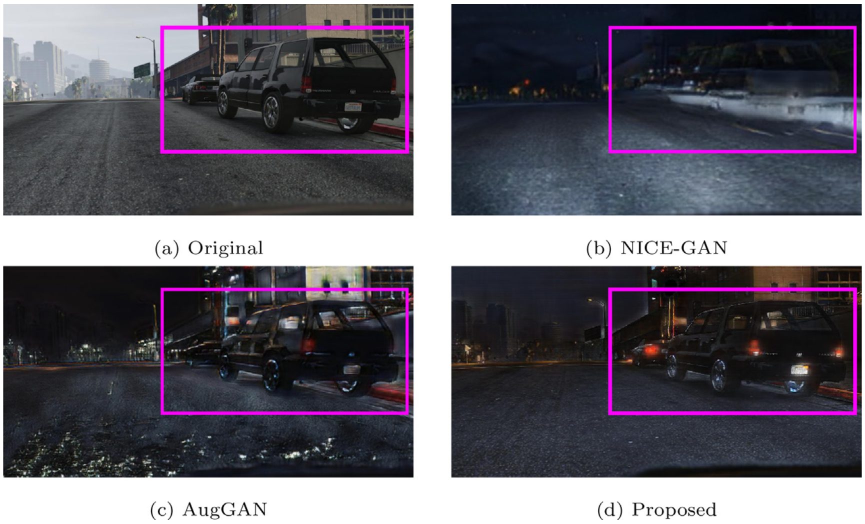


Fig. 1. Day-to-night image translation results of a sample image from GTA dataset [16]: (a) Original daytime image; Results of models (b) without and (c) with object preservation; and (d) our proposed with instance-aware image translation learning from the target-domain detector.

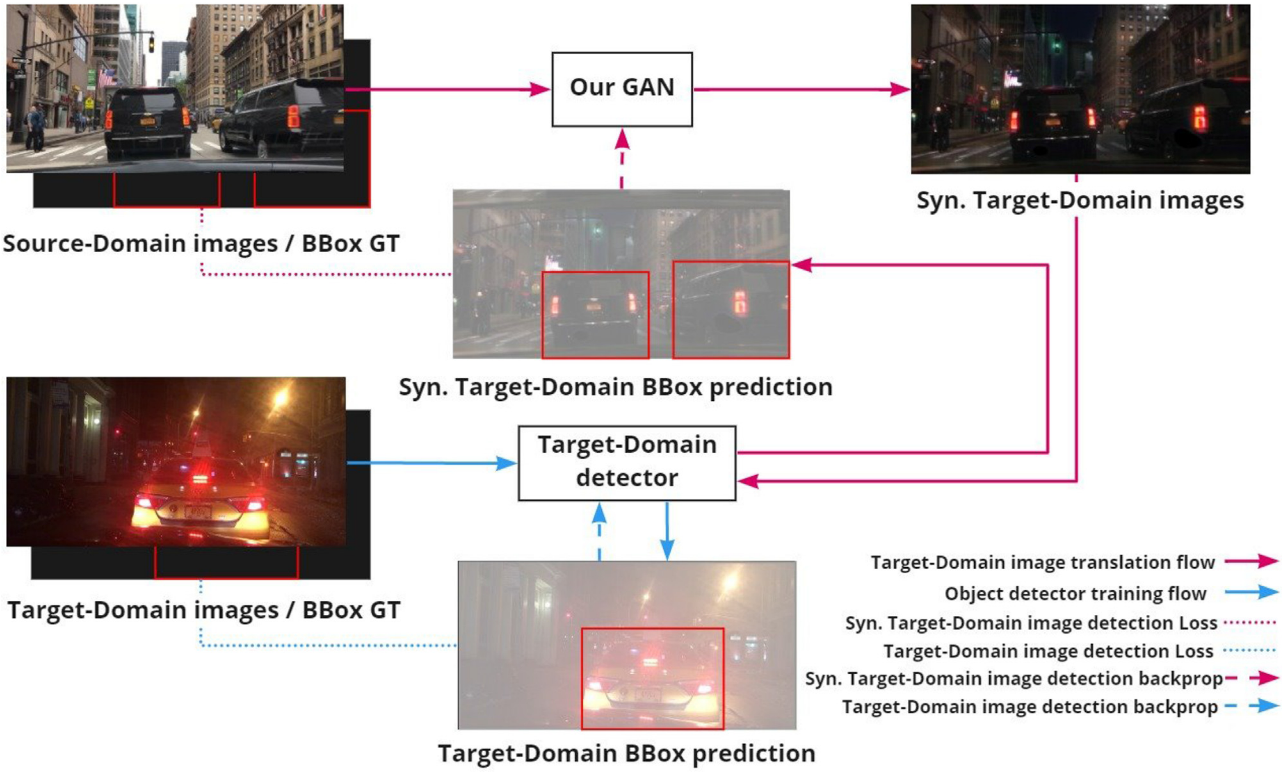

Our GAN
Source-Domain images / BBox GT
Syn. Target-Domain images
Syn. Target-Domain BBox prediction
Target-Domain detector
Target-Domain images / BBox GT
Target-Domain BBox prediction
Target-Domain image translation flow
Object detector training flow
Syn. Target-Domain image detection Loss
Target-Domain image detection Loss
Syn. Target-Domain image detection backprop
Target-Domain image detection backprop

Fig. 2. How a target-domain detector can help train a GAN to perform instance-aware image-translation.

learning, and aligned the instance features between the original and the (day-to-night) transformed images. Yet, in the test time, the object detector is still required to improve the performance.

In this paper, for the first time, we introduce an instance-aware GAN framework named as AugGAN-Det to jointly train a generator with an object detector (for image-object style) and a discriminator (for global style) as shown in Fig. 2. With this, a novel **C**ycle-**o**bject **Co**nsistency (CoCo) loss is proposed to preserve the instance level characteristic during image-to-image translation. That is, the object detector (i.e., bounding box) of the target domain will be directly involved in training the generator and resulting in guiding the image-objects in the translated images to carry realistic target-domain appearances across complex domains. Most importantly, the object detector is not required at the test time in contrast to Bhattacharjee et al. [15].

Our contributions are as follows: (i) We design an image-to-image translation network which jointly trains a generator with an object detector (for object-style) and a discriminator (for global-style) by leveraging a novel cycle consistency loss dubbed CoCo. Most importantly, an object detector is not required at test time; (ii) We quantitatively demonstrate that solely using the object labels (i.e., bounding box) for learning object-preserving image translation can achieve better results than leveraging pixel-level semantic segmentation into GAN training [10–12] (see Table 3); and (iii) Extensive experiments are conducted. Our method achieves better quantitative and qualitative results

Table 1
Network architecture of encoders, generators, discriminators and detectors in instance-aware image-to-image translation model: N, K, S denote the number of convolution filters, kernel size, and stride; n is the number of neurons of the fully-connected layers and the last layer of detector assumes $C = 1$. Please refer to Fig. 3 for knowing where E, G, D, B, and H are used in the overall network structure.

Layer	Encoders (E)	Layer info
1	CONV, ReLU	N64,K7,S1
2	CONV, ReLU	N128,K3,S1
3	CONV, ReLU	N256,K3,S2
4	CONV, ReLU	N512,K3,S2
5	RESBLK, ReLU	N512,K3,S1
6	RESBLK, ReLU	N512,K3,S1
Layer	Generators (G)	Layer info
1	RESBLK, ReLU	N512,K3,S1
2	RESBLK, ReLU	N512,K3,S1
3	DCONV, ReLU	N128,K3,S2
4	DCONV, ReLU	N64,K3,S2
5	CONV, Tanh	N3,K7,S1
Layer	Discriminator (D)	Layer info
1	CONV, LeakyReLU	N64, K4, S2
2	CONV, LeakyReLU	N128, K4, S2
3	CONV, LeakyReLU	N256, K4, S2
4	CONV, LeakyReLU	N512, K4, S2
5	CONV, LeakyReLU	N512, K4, S1
6	CONV, Sigmoid	N1, K4, S1
Layer	Detectors (B + H)	Layer info
1	CONV, ReLU	N64,K7,S2
2	Maxpool	K3,S2
3	CONV, ReLU	N64,K3,S2
4	CONV, ReLU	N128,K3,S2
5	CONV, ReLU	N256,K3,S2
6	CONV, ReLU	N512,K3,S2
7	Avgpool	K12x6
8	Fully-connected, ReLU	n4096
9	Fully-connected	n792

mainly on three popular benchmarks namely **INIT**, **GTA** and **BDD100k**.

2. Related work

2.1. Object detection

In the past few years, object detectors have achieved remarkable performance with the advent of CNNs. A modern detector is usually composed of two parts, a pre-trained CNN backbone and a detection head to predict the classes and bounding boxes of objects. In general, object detectors can be categorized into two camps, i.e., one-stage object detectors [1,17–20] and two-stage object detectors [21–26]. One-stage object detectors recently received more attention, since real-time applicability is of great and practical interest in many applications.

2.2. Data augmentation

Data augmentation is an essential technique to increase the robustness and to achieve higher detection accuracy of an object detection model. For example, Random Erasing [27] and CutOut [28], tried to simulate object occlusion in the hope that the detector learns to visually understand the essence of objects' appearances even though only part of an object can be seen. Works such as DropOut [29], DropConnect [30] and DropBlock [31] apply a similar concept to feature maps. More recently, MixUp [32], CutMix [33], GridMix [34] and Mosaic [1] were proposed to combine multiple images for additional data augmentation. However, the above-introduced methods are often not specifically designed to enhance a model's robustness across domains. As pointed out in Braun et al. [2], the (pedestrian) detector is recommended to be trained using the data of the domain for the detector to be deployed to achieve the highest accuracy. Thus, standard data augmentation strategies are not sufficient across domains.

2.3. Generative adversarial networks

Due to the recent success of GANs [5], many approaches adopted GANs for the image translation task. For example, Pix2Pix [35] provides visually plausible images in the target domain given paired training data. By introducing the cycle consistency constraint to encourage bidirectional image translation with regularized structural output, CycleGAN [36] achieved astonishing image translation results when only unpaired data is available. UNIT [6] further applied weight-sharing constraints to increase the translation consistency. Usually, GAN models abandon the discriminator once the training process is completed. However, NICE-GAN [37] demonstrated that the encoder trained by the adversary loss in the discriminator is still informative. Therefore, reusing discriminators for encoding in generating images is quantitatively beneficial.

To enforce the structure-consistency between the source and the generated images, CyCADA [10] tried to incorporate a downstream segmentation model in the forward cycle and a semantic consistency loss in the backward cycle. AugGAN [11] proposed to utilize auxiliary segmentation tasks in a multi-tasking fashion in both cycles to prevent content distortion. The major difference between CyCADA and AugGAN is that the former only involves a downstream segmentation task in the forward cycle. AugGAN designed multi-tasking generators that learn to perform image-translation and segmentation simultaneously in both cycles. BicycleGAN [38] is a multimodal image-to-image translation model, but requires paired data that cannot be easily acquired in real-driving scenarios. Both DRIT [39] and MUNIT [8] are multimodal GANs able to work with unpaired images. However, the overall image style and particular objects appearances cannot be transformed individually. Multimodal AugGAN [12], a multimodal structure-consistent image-to-image translation network, integrates semantic segmentation models for both domains with a multimodal image-translation network. Compared to AugGAN, Multimodal AugGAN can provide diverse and visually compelling results in the target domain with better object preservation due to the multimodal behavior. However, the necessity of pixel annotations limits the method's applicability. Both INIT [14] and DUNIT [15] are instance-aware GAN models. The former method employs the instance and the global style to guide the generation of target-domain objects. However, the model discards the instance-level information at test time, and only the global module is used. The latter work applies an off-the-shelf object detector (trained by MSCOCO [40]) and an instance-level encoder to extract instance-boosted features during learning, and align the instance features between the original and transformed images. As shown in Braun et al. [2], for pedestrian detection, the highest detection accuracy is obtained when the training and test data are from the same domain (i.e., time-of-day). However, MSCOCO only contains less than 1% low-light images, and DUNIT still requires an object detector at test time to leverage the object instance features.

3. Proposed model

In the image translation problem, the goal is to learn a network between two visual domains $X \subset \mathbb{R}^{H \times W \times 3}$ and $Y \subset \mathbb{R}^{H \times W \times 3}$. Previous methods sometimes expect that an n-class segmentation ground-truth, i.e., $\hat{X} \subset \mathbb{R}^{H \times W}$ and $\hat{Y} \subset \mathbb{R}^{H \times W}$ is available so that image-structure of a transformed image is consistent with its counterpart in the original domain. However, obtaining pixel-wise annotation is very expensive. Therefore, in this work, we only assume that bounding boxes with associated object labels from two

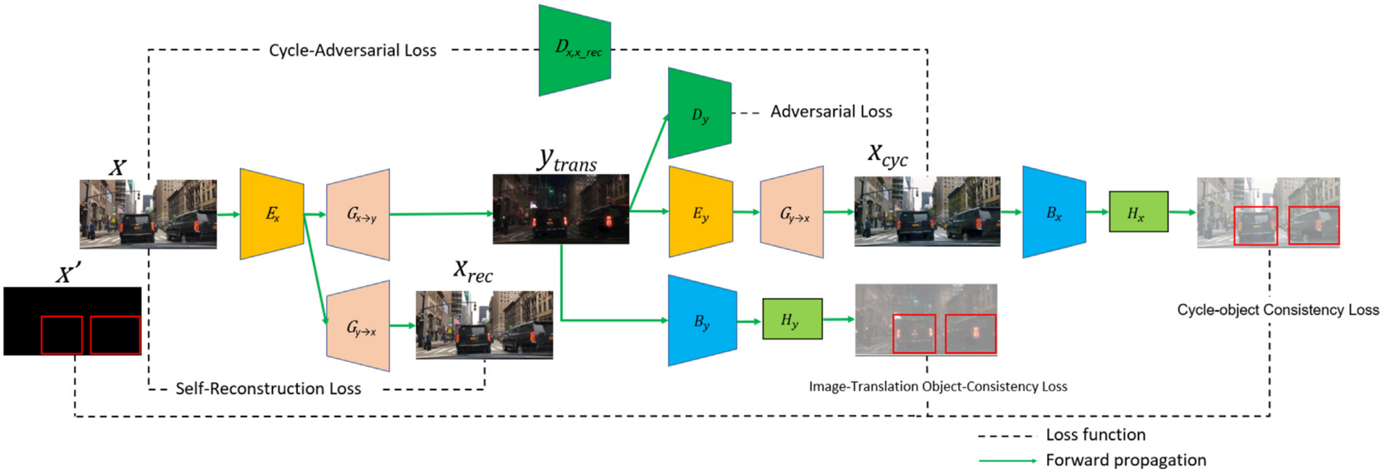


Fig. 3. Overall structure of the proposed cycle-object-consistent image-to-image translation network: x: sampled image from domain X; x': bounding box Ground-Truth of x; y_{trans}: translated result; x_{rec}: self-reconstructed image given x; x_{cyc}: cycle-reconstructed image corresponding to x.

visual domains, $X' \subset \mathbb{R}^{M\times N(C+5k)}$ and $Y' \subset \mathbb{R}^{M\times N(C+5k)}$, i.e., k objects with C classes inside $M \times N$ grid cells are available. Our objective is to learn the mapping $G_{x\to y}$ and $G_{y\to x}$ conditioned on X' and Y', given X and Y.

The detailed architecture of our network is given in Table 1. Our detector comprises a backbone, pretrained ResNet-18, and a detection head. It is worth mentioning that the grid cell size is 12×6 and each grid cell could predict two objects (described by u, v, w, h and an objectness score) of the same class. In this work, when C classes of objects are considered, the neurons of the last layer become $12 \times 6 \times (5 \times 2 + C)$ for a 384×192 image. For the discriminators, we follow the design of PatchGAN [35] because it is flexible to work on arbitrarily-sized images in a fully convolutional fashion.

3.1. Detection loss

3.1.1. Cycle-object consistency (CoCo) loss

Our model utilizes a target-domain detector instead of an off-the-shelf object detector [15] to guide the generator. Generally, we expect that given the encoded latent vector generated by E_x, the generator $G_{x\to y}$ learns to generate images in an attempt to fool the discriminator D_x while the object consistency is kept. As pointed out by Chu et al. [36], a strong cycle-consistency will enforce the reconstructed information to be hidden in the translated image, y_{trans}. Therefore, in our cycle-reconstruction phase, another discriminator is added to constrain the cycle-adversarial consistency between x and x_{rec} which is produced by E_y, the generator $G_{y\to x}$. Most importantly, we propose Cycle-object Consistency (CoCo) loss as shown in Fig. 3. It keeps the objects in both y_{trans} and x_{cyc} detectable at the same time in the forward cycle and in both x_{trans} and y_{cyc} in the backward cycle. i.e., to encourage detection result $H_x(B_x(x_{\text{cyc}}))$, $H_y(B_y(y_{\text{cyc}}))$ predicted by the detection backbones, B_x and B_y, and the prediction heads, H_x and H_y, to be similar to the detection Ground-Truth, x', y', respectively. That is, it enforces object(s) reservation in the translated images.

Technically, we incorporate a one-stage object detector into the training of our generators. In the forward cycle, the loss of an object whose bounding box center is located inside grid cell (i, j) is defined by

$$\mathcal{L}_{\text{obj}}(E_x, G_{x\to y}, E_y, G_{y\to x}, B_x, H_x, X, X') =$$

$$\mathbb{E}_{x,x'\sim p_{X,X'}}\Bigg[\sum_{n=1}^{k}\Big(\sum_{p\in\{u,v\}}\alpha_p(G^p_{ij,n}(x_{\text{cyc}}) - G^p_{ij}(x'))^2$$

$$+\sum_{q\in\{w,h\}}\alpha_q(\sqrt{G^q_{ij,n}(x_{\text{cyc}})} - \sqrt{G^q_{ij,n}(x')})^2$$

$$+(G^o_{ij,n}(x_{\text{cyc}}) - G^o_{ij,n}(x'))^2) + \sum_{c=1}^{C}(G^c_{ij,n}(x_{\text{cyc}}) - G^c_{ij,n}(x'))^2)\Bigg], \quad (1)$$

where $x_{\text{cyc}} = G_{y\to x}(E_y(G_{x\to y}(E_x(x))))$. $G^c_{ij,n}(x_{\text{cyc}})$ is the class score of grid cell (i, j). $G^o_{ij,n}(x_{\text{cyc}})$, $G^u_{ij,n}(x_{\text{cyc}})$, $G^v_{ij,n}(x_{\text{cyc}})$, $G^w_{ij,n}(x_{\text{cyc}})$, and $G^h_{ij,n}(x_{\text{cyc}})$ are the objectness score, the coordinate, the width, and the height of the nth predicted output, given the input image x. The corresponding ground-truth values are $G^c_{ij,n}(x')$, $G^o_{ij,n}(x')$, $G^u_{ij,n}(x')$, $G^v_{ij,n}(x')$, $G^w_{ij,n}(x')$, and $G^h_{ij,n}(x')$, respectively. In this work, we set $\alpha_u = \alpha_v = \alpha_w = \alpha_h = 5$.

As to a non-object grid, i.e., no object whose center of object window locates inside grid cell (i, j), its loss function is given by

$$\mathcal{L}_{\neg\text{obj}}(E_x, G_{x\to y}, E_y, G_{y\to x}, B_x, H_x, X, X') =$$

$$\mathbb{E}_{x,x'\sim p_{X,X'}}\left[\sum_{n=1}^{k}\alpha_{\neg\text{obj}}(G^o_{ij,n}(x_{\text{cyc}}) - G^o_{ij,n}(x'))^2\right], \quad (2)$$

where $\alpha_{\neg\text{obj}} = 0.5$. As to the backward cycle, the object-grid and non-object grid loss is stated as $\mathcal{L}_{\text{obj}}(E_y, G_{y\to x}, E_x, G_{x\to y}, B_y, H_y, Y, Y')$ and $\mathcal{L}_{\neg\text{obj}}(E_y, G_{y\to x}, E_x, G_{x\to y}, B_y, H_y, Y, Y')$.

3.1.2. Image-translation object-consistency loss

Aside from CoCo loss, we found that the interaction between the object detector and generator is very similar to the one between the discriminator and generator, i.e., both the object detector and the discriminator can overpower the generator. Inspired by label smoothing [41] that proposed to prevent the discriminator from being overconfident, we also try to attain a balance between the generator and the object detector by controlling the detector's training convergence. For this purpose, we experimented with GTA dataset. As seen in Table 6, an unconverged and frozen object detector can only guide the generator to yield objects with very-limited target-domain-style appearances, thus leading to a very low AP (Average Precision) [42]. Meanwhile, a converged and frozen detector tends to provide a non-informative back-propagation signal to the generator and has a better AP than an unconverged and frozen object detector. Finally, when the object detector is jointly trained with the generator, it achieves the highest AP as the generator continuously learns to generate realistic and target-domain-detector-detectable objects.

Technically, given the translated images y_{trans}, the image-translation object-consistency losses are modeled via $\mathcal{L}_{\text{obj}}(E_x, G_{x\to y}, B_y, H_y, X, X')$ and $\mathcal{L}_{\neg\text{obj}}(E_x, G_{x\to y}, B_y, H_y, X, X')$. Analogously, the additional losses utilized in the backward cycle are $\mathcal{L}_{\text{obj}}(E_y, G_{y\to x}, B_x, H_x, Y, Y')$ and $\mathcal{L}_{\neg\text{obj}}(E_y, G_{y\to x}, B_x, H_x, Y, Y')$, respectively.

3.2. Other losses

3.2.1. Adversarial loss

There are two kinds of adversarial losses in our model. The first one is designed for leading x and y to be properly translated to y and x, respectively, in terms of style. In this work, we apply least-squares adversarial loss [43] because it yields better image-translation results in our experiments. The first adversarial loss function is given as

$$\mathcal{L}_{\text{GAN}}(E_x, G_{x\to y}, D_y, X, Y) = \mathbb{E}_{y\sim p_Y}\left[(D_y(y))^2\right] + \mathbb{E}_{x\sim p_X}\left[(1 - D_y(G_{x\to y}(E_x(x))))^2\right], \tag{3}$$

where E_x and $G_{x\to y}$ try to generate transformed images $G_{x\to y}(E_x(x))$ that look similar to images from domain Y, while D_y aims to distinguish between translated samples $G_{x\to y}(E_x(x))$ and real samples y in terms of style. In the image-translation phase of the backward cycle, the adversarial loss is $\mathcal{L}_{\text{GAN}}(E_y, G_{y\to x}, D_x, X, Y)$.

In order to encourage x_{rec} and y_{rec} to be close to the original x and y, the second adversarial loss uses two additional discriminators, $D_{x,x_{\text{rec}}}$ and $D_{y,y_{\text{rec}}}$, respectively. The cycle-adversarial loss in the forward cycle is modeled as follows,

$$\mathcal{L}_{\text{GAN}}(E_x, G_{x\to y}, E_y, G_{y\to x}, D_{x,x_{\text{rec}}}, X) = \mathbb{E}_{x\sim p_X}\left[(D_{x,x_{\text{rec}}}(x))^2\right] + \mathbb{E}_{x\sim p_X}\left[(1 - D_{x,x_{\text{rec}}}(G_{y\to x}(E_y(G_{x\to y}(E_x(x))))))^2\right]. \tag{4}$$

The backward cycle is modeled via an analogous loss, $\mathcal{L}_{\text{GAN}}(E_y, G_{y\to x}, E_x, G_{x\to y}, D_{y,y_{\text{rec}}}, Y)$.

3.2.2. Self-reconstruction loss

CycleGAN adopted a technique proposed by Taigman et al. [44] to regularize the generator to be close to an identity mapping when real samples of the target domain are provided as the input to the generator. The reconstruction loss in this work is based on the shared-latent space assumption [6]. It is done by regularizing the translation to approximate the identity mapping when the latent vectors of the source images are provided as the input to $G_{y\to x}$. It is modeled via an auto-encoder type loss,

$$\mathcal{L}_{\text{AE}}(E_x, G_{y\to x}, X) = \mathbb{E}_{x\sim p_X}[|x - G_{y\to x}(E_x(x))|_1]. \tag{5}$$

The same reconstruction loss is applied to the backward cycle as $\mathcal{L}_{\text{AE}}(E_y, G_{x\to y}, Y)$.

3.3. Network learning

The goal of our network is that both generators learn to transform the style of the overall image and the particular object appearances individually. The entire objective is given as follows,

$$\begin{aligned}
\mathcal{L} = {} & \mathcal{L}_{\text{GAN}}(E_x, G_{x\to y}, D_y, X, Y) \\
& + \mathcal{L}_{\text{GAN}}(E_y, G_{y\to x}, D_x, X, Y) \\
& + \mathcal{L}_{\text{AE}}(E_x, G_{y\to x}, X) + \mathcal{L}_{\text{AE}}(E_y, G_{x\to y}, Y) \\
& + \lambda_{\text{img-obj}}\big(\mathcal{L}_{\text{obj}}(E_x, G_{x\to y}, B_y, H_y, X, X') \\
& \quad + \mathcal{L}_{\neg\text{obj}}(E_x, G_{x\to y}, B_y, H_y, X, X') \\
& \quad + \mathcal{L}_{\text{obj}}(E_y, G_{y\to x}, B_x, H_x, Y, Y') \\
& \quad + \mathcal{L}_{\neg\text{obj}}(E_y, G_{y\to x}, B_x, H_x, Y, Y')\big) \\
& + \lambda_{\text{cyc-obj}} \times \\
& \big(\mathcal{L}_{\text{obj}}(E_x, G_{x\to y}, E_y, G_{y\to x}, B_x, H_x, X, X') \\
& \quad + \mathcal{L}_{\neg\text{obj}}(E_x, G_{x\to y}, E_y, G_{y\to x}, B_x, H_x, X, X') \\
& \quad + \mathcal{L}_{\text{obj}}(E_y, G_{y\to x}, E_x, G_{x\to y}, B_y, H_y, Y, Y') \\
& \quad + \mathcal{L}_{\neg\text{obj}}(E_y, G_{y\to x}, E_x, G_{x\to y}, B_y, H_y, Y, Y')\big) \\
& + \lambda_{\text{cyc-adv}} \times \\
& \big(\mathcal{L}_{\text{GAN}}(E_x, G_{x\to y}, E_y, G_{y\to x}, D_{x,x_{\text{rec}}}, X) \\
& \quad + \mathcal{L}_{\text{GAN}}(E_y, G_{y\to x}, E_x, G_{x\to y}, D_{y,y_{\text{rec}}}, Y)\big),
\end{aligned} \tag{6}$$

and we aim to solve the following optimization problem during model training:

$$\min_{\substack{E_x, G_{x\to y},\\ E_y, G_{y\to x},\\ B_x, H_x,\\ B_y, H_y}} \max_{\substack{D_x, D_{x,x_{\text{rec}}},\\ D_y, D_{y,y_{\text{rec}}}}} \mathcal{L}. \tag{7}$$

4. Experimental results

Datasets Generally, most of the existing freely available datasets [45–47] were collected during the day. In this work, we have mainly tested our model, AugGAN-Det, on three datasets: (i) INIT dataset [14] is proposed for on-road image translation in four driving scenarios where object detection labels are provided. All the data (132,201 images for training and 23,328 images for testing) were collected in Tokyo, Japan; (ii) GTA dataset [16] - one of the most famous synthetic datasets that contain both low-level and high-level annotations including optical flow, semantic segmentation, instance segmentation, object detection, and tracking. The dataset is split into 134 K, 50 K, and 70 K frames for training, validation, and testing, respectively; and (iii) BDD100k dataset [3] is collected in many cities and regions in the US and contains 100k driving videos recorded in diverse weather conditions at different time-of-day. The videos are split into training (70k), validation (10k) and testing (20k). Each video's frame at the 10th second is annotated for image tasks, including detection and segmentation. Recently, DarkFace dataset [48] provides face annotation in poor visibility situations such as challenging lighting conditions at nighttime. Since the faces are tiny to guide the GAN model, we trained our model using the pedestrian labels from BDD100k. Then the trained model performed day-to-night image-translation on FEEDS (Face pEdestrain dEtection DataSet) dataset [49] and sampled LTFT (Long-Term Face Tracking) dataset [50] to train a better face detector to be assessed on DarkFace dataset. More specifically, 4138 face-containing images from the training set of FEEDS dataset and sampled 1000 images from the street and Bengal sequences of LTFT dataset were involved. There are 6000 labeled images in DarkFace dataset and the training and validation split was done by randomly sampling 4000 and 2000 images, respectively. Finally, we also evaluated our model on the task of cross-dataset domain adaptation from KITTI [45] to cityscape [46]. The former is captured by driving around the mid-sized city of Karlsruhe in Germany and consists of 7481 training images and 7518 testing images; while the latter is collected in 50 different cities in Europe and is composed of 2975 training, 500 validation, and 1525 testing images. Both datasets are designed for a suite of vision tasks including object detection.

Previous works [10–12] that are designed explicitly for object-preserving image transformation need semantic segmentation labels to prevent significant image distortion. Since this work tries to make a fairer comparison, we also conduct our GAN training on both GTA and BDD100k datasets, but only the detection labels of the same images are used. We consider car, bus, truck, and van for GTA; car, bus, and truck for BDD100k; car for INIT. For the KITTI-to-Cityscape cross-dataset domain adaptation, pedestrian, car, truck, and cyclist are involved in the experiment.

Table 2
IS of different models in INIT dataset.

Scenario	CycleGAN	UNIT	MUNIT	DRIT	INIT	DUNIT	Ours
Sunny-to-night	1.026	1.030	1.278	1.224	1.118	1.259	**1.344**
Night-to-sunny	1.023	1.024	1.051	1.099	1.080	1.108	**1.184**
Sunny-to-rainy	1.073	1.075	1.146	1.207	1.152	1.225	**1.287**
Rainy-to-sunny	1.090	1.023	1.102	1.103	1.119	1.125	**1.271**
Sunny-to-cloudy	1.097	1.134	1.095	1.104	1.142	1.149	**1.214**
Cloudy-to-sunny	1.033	1.046	1.321	1.249	1.460	1.472	**1.509**
Average	1.057	1.055	1.166	1.164	1.179	1.223	**1.301**

Table 3
Detection accuracy comparison - YOLOv4 trained with day-to-night-transformed images generated by GANs. G-D2N: GTA-val-day-to-night; B-D2N: BDD100k-det-val-day-to-night; G-N: GTA-val-night; B-N: BDD100k-det-val-night.

Train	Test	NICE-GAN	CyCADA	MUNIT	AugGAN	M-AugGAN	AugGAN-Det
G-D2N	G-N	0.485	0.530	0.475	0.537	0.545	**0.568**
G-D2N	B-N	0.457	0.475	0.453	0.481	0.486	**0.510**
(B + G)-D2N	B-N	0.485	0.501	0.489	0.505	0.510	**0.559**

Architecture Both [11,12] evaluated the performance of day-to-night image transformation with YOLOv1 and Faster R-CNN [23], which are already somehow outdated object detectors. Moreover, most detection applications in advanced driver assistance systems (ADAS) or autonomous vehicles have moved to YOLOv3 or even YOLOv4. Therefore, in this work, other than the domain adaptive detection experiment analyzed by Faster R-CNN between KITTI and Cityscape datasets, we conduct most of the analysis using YOLOv4. We follow the same protocol of the previous methods in assessing the images transformed by different GAN models.

Implementation Our proposed model is implemented in PyTorch [51]. Due to the GPU memory limitation, we use an input image resolution of 384×192 pixels. In our detector implementation, we adopted lightweight ResNet-18 as the backbone and a YOLO-like (12×6 grid cells) head stacked on top of it. In all of our experiments, we train source- and target-domain detectors separately and they will later be jointly trained with the generators. Both detectors are trained for 30 epochs using SGD with a batch size of 32, a momentum of 0.9, a learning rate of 0.0001 and a decay of 0.0005. Finally, when the training of both detectors is finished, they are integrated with our GAN model to represent the CoCo loss for the generators to produce instance-aware image-translation results. It is worth mentioning that, both detectors are still trained simultaneously with our GAN model, but they only learn from real images and the corresponding detection labels. Finally, the four discriminators, two generators and two detectors are jointly trained using the Adam optimizer [52] with a batch size of 2, a learning rate of 0.0002, exponential decay rates $(\beta_1, \beta_2) = (0.5, 0.999)$, epochs of 200. We set the weightings related to the multi-task loss to be $\lambda_{\text{img-obj}} = 1$, $\lambda_{\text{cyc-obj}} = 2$, and $\lambda_{\text{cyc-adv}} = 5$.

4.1. INIT dataset

As shown in Table 2, our results are consistently better than other models, particularly for INIT and DUNIT, in a totally 6 scenarios in terms of Inception Score (IS) [53] which is an important metric for assessing the performance of image-translation. Qualitatively, in the day-to-night (Fig. 4) scenario, it shows that our model has a more balanced color contrast in terms of global-style and object-style as compared to other competitors, especially to MUNIT and DRIT. Then, in terms of day-to-cloudy (Fig. 5) scenario, yet, it shows that our model can display more object-preserving results against other competitors as indicated in the zoom-in bounding box.

4.2. GTA dataset

Next, we evaluate our model on the synthetic dataset, GTA. Our model clearly surpasses Multimodal-AugGAN and only needs bounding box GT annotations. As shown in the 1st row of Table 3, our model outperforms competing methods in terms of nighttime detection accuracy. Qualitatively, as seen in Fig. 6, it is quite evident that our model could yield visually-plausible instance-aware target-domain-looking results.

4.3. BDD100k dataset

In order to achieve better results in the real-driving BDD100k dataset, we not only perform day-to-night image-translation for BDD100k using GANs learning from GTA but also try to use both of them. As shown in Table 3, our model consistently outperforms other competing methods. In Fig. 7, we could easily observe that NICE-GAN and MUNIT would sometimes turn off the front or the rear lamps of the vehicles inside images. Even though the lamps are turned on in some cases, the location and the color might be wrong. AugGAN and Multimodal AugGAN could achieve better results considering both factors. However, even with semantic segmentation subtasking network, the appearance of nighttime vehicles is still not realistic enough because the style-translation of particular objects is not instance-aware and might be compromised by the overall style of the image.

4.4. DarkFace dataset

Using face detection as the downstream detector in our GAN for performing image-translation is possible in practice. However, detecting tiny objects in a shallow detector to be integrated into our GAN framework is very difficult. Therefore, we propose to train AugGAN-Det with the pedestrian labels from BDD100k dataset, perform D2N translation on sampled FEEDS and LTFT datasets which provide face labels for later training powerful YOLOv4, and then test the face detection results on the validation set of DarkFace dataset. It is worth mentioning that since DarkFace dataset is very dark and is significantly different from BDD100K, YOLOv4 was trained with the training set of DarkFace and the Day-to-Night-transformed images from FEEDS and LTFT datasets. As can be seen in Table 4, since MUNIT and NICE-GAN cannot darken the image while keeping the persons in the translated images, involving images transformed by either model will trail the detection accuracy of YOLOv4 trained only with DarkFace training set. Our AugGAN-Det not only fulfilled the two conditions above but also achieved

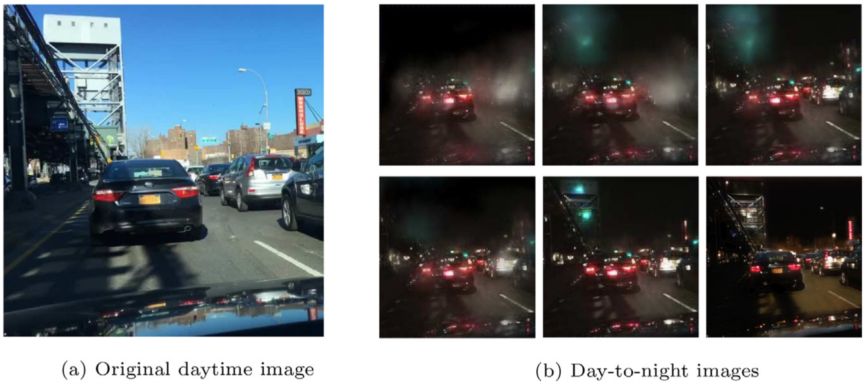

(a) Original daytime image
(b) Day-to-night images

Fig. 4. INIT Dataset (Day-to-night): A comparison of different models. From left to right - (Top) CycleGAN, UNIT and MUNIT. (Bottom) DRIT, DUNIT and Ours model, respectively.

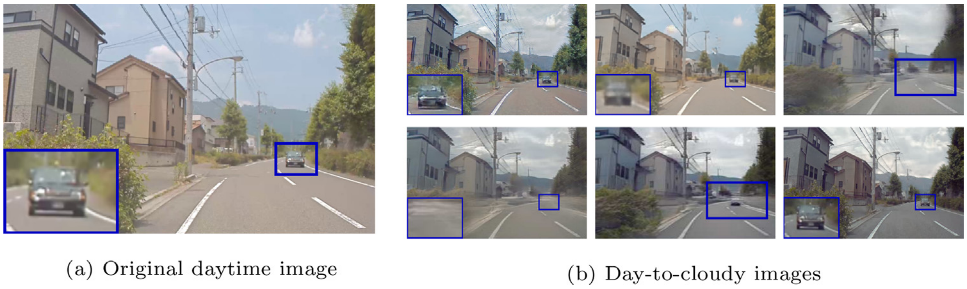

(a) Original daytime image
(b) Day-to-cloudy images

Fig. 5. INIT Dataset (Day-to-Cloudy): A comparison of different models. From left to right - (Top) CycleGAN, UNIT and MUNIT. (Bottom) DRIT, DUNIT and Ours model, respectively.

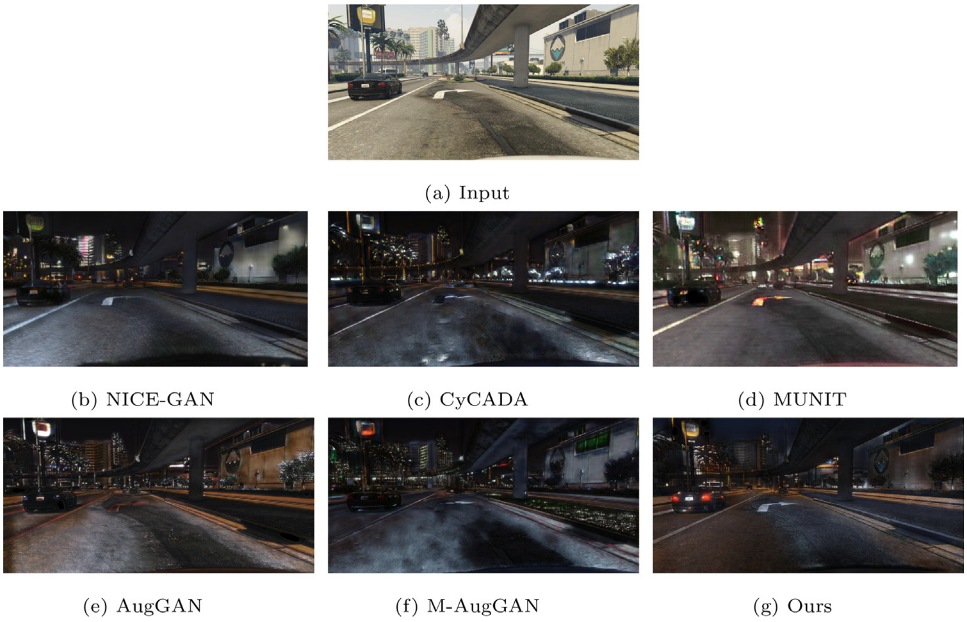

(a) Input
(b) NICE-GAN
(c) CyCADA
(d) MUNIT
(e) AugGAN
(f) M-AugGAN
(g) Ours

Fig. 6. GTA-val-day: A comparison of day-to-night transformation results done by different models.

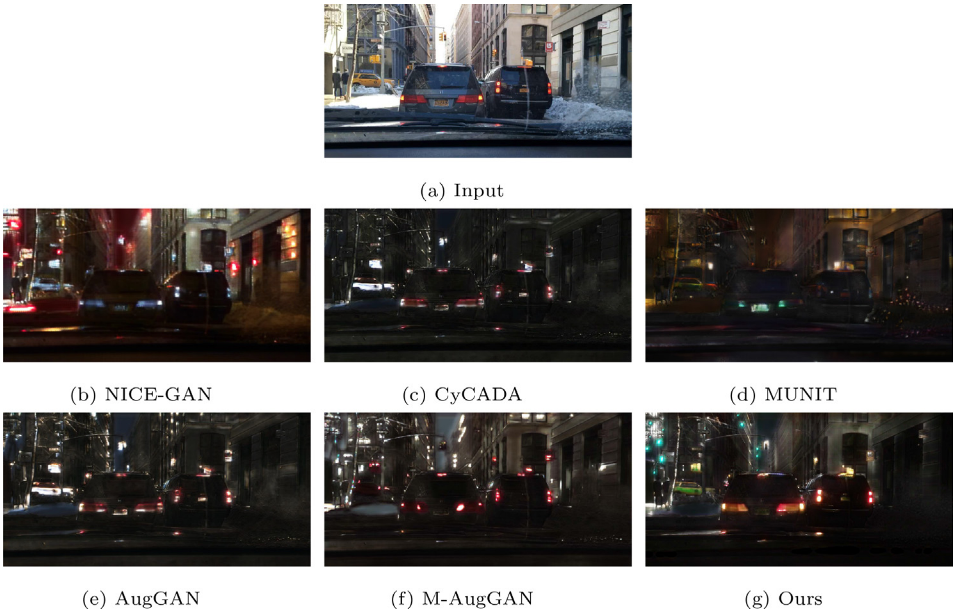
higher detection accuracy, which clearly shows that involving day-to-night-transformed images is also helpful in detecting faces in an extremely dark scenario. The translated images can be seen in Figs. 8 and 9, respectively.

4.5. Transformations of more domains

Our model is capable of learning transformation across unpaired domain pairs where either of the domain could be in different weather conditions and times-of-the-day. A more thorough demonstration is shown in Fig. 10 across the three public datasets employed in this paper. We can easily observe that the style-translation of particular objects is instance-aware without being compromised by the style of the overall image.

4.6. Object detection in real-driving scenario

Since our model can provide more visual-compelling image-translation results, nighttime detector learning from the day-to-night transformed images generated by our model could achieve

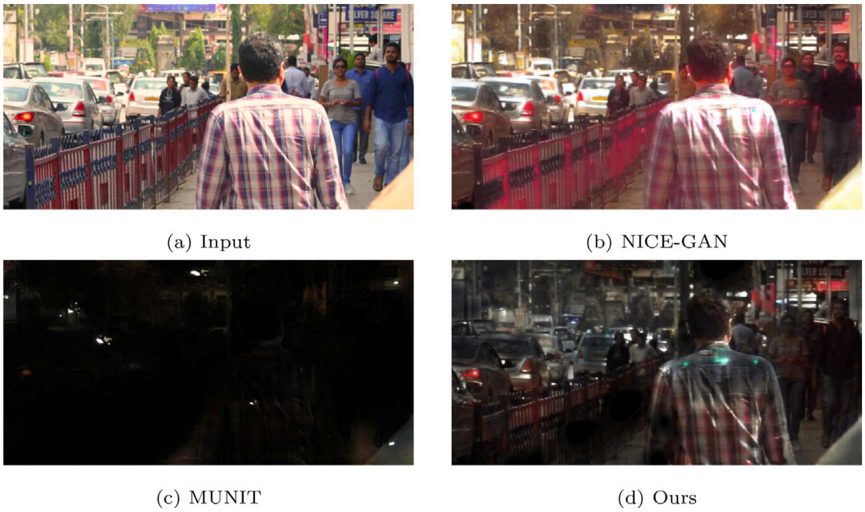


(a) Input

(b) NICE-GAN (c) CyCADA (d) MUNIT

(e) AugGAN (f) M-AugGAN (g) Ours

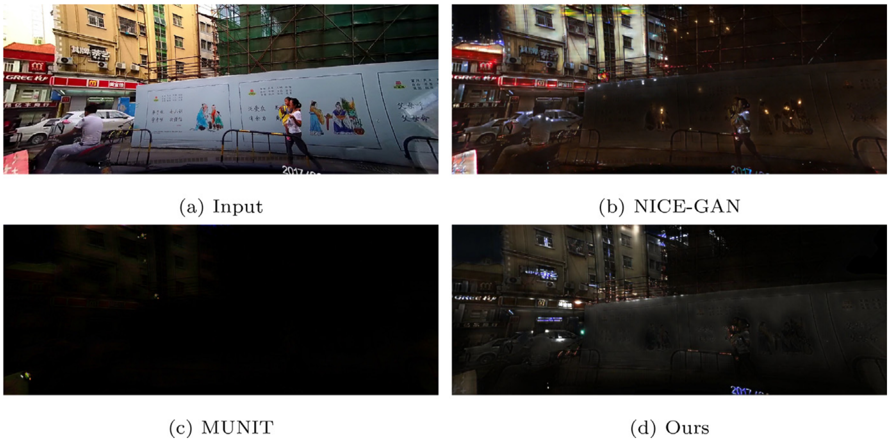
Fig. 7. BDD100k-val-day: A comparison of day-to-night transformation results done by different models.



(a) Input (b) NICE-GAN

(c) MUNIT (d) Ours

Fig. 8. LTFT-day-to-night results generated by GANS learning from BDD100k.



(a) Input (b) NICE-GAN

(c) MUNIT (d) Ours

Fig. 9. FEEDS-day-to-night results generated by GANS learning from BDD100k.

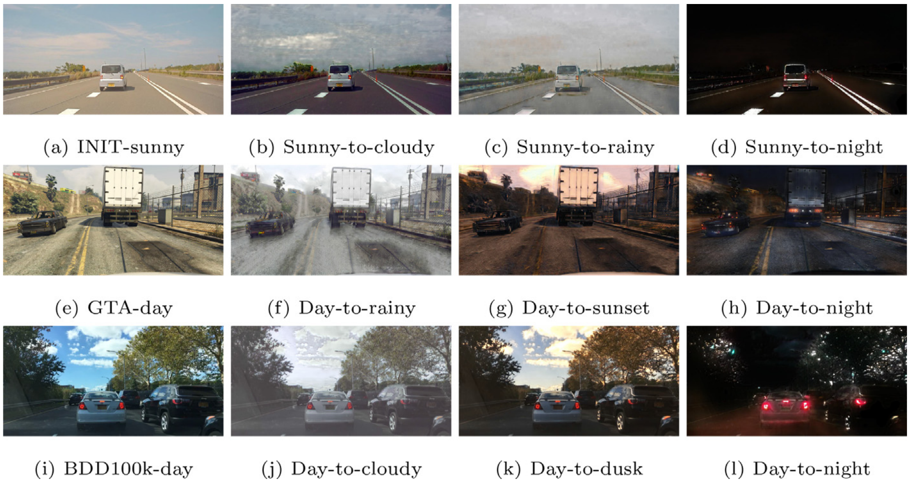


Fig. 10. More image-translations across different domains.

Table 4
Detection accuracy comparison - YOLOv4 trained with DarkFace-train (baseline) and (LTFT + FEEDS)-day-to-night-transformed images generated by GANs. Please remind that in this setting, DarkFace-train is mixed with the day-to-night images in each experiment.

Baseline	NICE-GAN	MUNIT	AugGAN-Det
0.191	0.179	0.175	**0.205**

Table 5
Detection accuracy comparison for the KITTI-to-Cityscape adaptation scenario.

Methods	Pedestrian	Car	Truck	Cyclist	mAP
DT [54]	0.285	0.407	0.259	0.297	0.312
DAF [55]	0.392	0.402	0.257	0.489	0.385
DARL [56]	0.464	0.587	0.270	0.491	0.453
DAOD [57]	0.473	0.591	0.283	0.496	0.461
DUNIT [15]	0.607	0.651	0.327	0.577	0.541
NICE-GAN [37]	0.282	0.525	0.285	0.479	0.393
MUNIT [8]	0.482	0.572	0.271	0.510	0.459
AugGAN-Det	**0.610**	**0.681**	**0.348**	**0.585**	**0.556**

Table 6
Different detector and generator training strategy in the GTA case.

Detector setting (in generator training)	AP
Unconverged & frozen	0.501
Converged & frozen	0.517
Jointly-trained	**0.568**

significantly better results in terms of nighttime vehicle detection, as seen in Fig. 11. Besides, during the training process, our model gradually learns to darken the overall image and turn on the rear lamps of the front vehicles, as shown in the bottom row of Fig. 12 (from left to right).

4.7. Face detection in extremely dark scenario

Training with DarkFace images and day-to-night-transformed images generated by our model can achieve better face detection results in poor visibility situations. The face detection result comparison can be seen in Fig. 13.

4.7.1. Other domain adaptation detection results

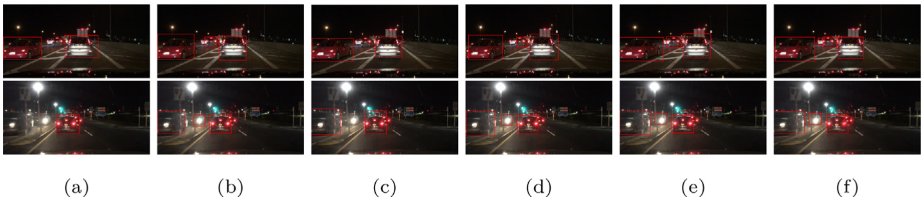
Our model is also tested on the task of cross-dataset domain adaptation. We follow the same experimental setup as Bhattacharjee et al. [15] in the KITTI-to-Cityscape domain adaptation. i.e., KITTI [45] dataset is the source domain and Cityscape [46] dataset is the target domain. In this experiment, Faster R-CNN is trained on the target-domain images and then evaluated on the source-to-target images provided by different models including DT [54], DAF [55], DARL [56], and DAOD [57]. This way, the performance of image-translation done by different models could be assessed by the detection accuracy. The detection results of translated images can be seen in Fig. 14. In the images transformed by MUNIT, the pedestrian and the car inside the red boxes are not preserved. Therefore, they cannot be detected and leads to a lower detection accuracy in Table 5. In our proposed approach, the generator learns to perform image-translation while keeping objects detectable by the target-domain detector as much as possible. DUNIT applies an off-the-shelf object detector in training GAN, which is why the detection accuracy is significantly higher than NICE-GAN and MUNIT. Our model outperforms other models in this task in terms of per-class AP and mAP because the generator is jointly trained with the object detector to learn to generate object-preserving and instance-aware translated images gradually.

5. Further model analysis

In the journey of this work to pursue visually-better and quantitatively-beneficial results, we performed some analysis targeting better architecture and training strategies.

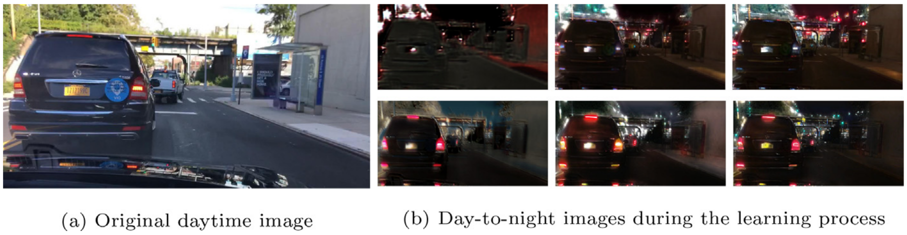
5.1. Detector and generator training

At the early stage of this work, we found that the interaction between the detector and the generator is very similar to the one between the discriminator and the generator. i.e., both the detector and the discriminator can overpower the generator. Inspired by label smoothing [41] that proposed to prevent the discriminator from being overconfident, we tried to attain the balance between the generator and the detector by controlling the detector's training convergence. We conducted three experiments to assess how an object detector is involved in generator training. As seen in Table 6, an unconverged and frozen detector can only guide the generator to yield objects with very-limited target-domain-style appearances, thus leading to lower AP. A converged and frozen detector tends to provide a non-informative back-propagation signal



(a) (b) (c) (d) (e) (f)

Fig. 11. YOLOv4 (trained with BDD100k-det-val-day day-to-night-transformed images generated by GANs learning from BDD100k-seg and GTA-train) detection result comparison on BDD100k-det-val-night: (a) YOLOv4 trained with images generated by NICE-GAN; (b) YOLOv4 trained with images generated by CyCADA; (c) YOLOv4 trained with images generated by MUNIT; (d) YOLOv4 trained with images generated by AugGAN; (e) YOLOv4 trained with images generated by Multimodal AugGAN; (f) YOLOv4 trained with images generated by our work.



(a) Original daytime image

(b) Day-to-night images during the learning process

Fig. 12. The learning process of NICE-GAN and our model. (Left image) a daytime image from BDD100k-val-day; (top) NICE-GAN; (bottom) our work.

(a) (b) (c) (d)

Fig. 13. YOLOv4 detection results on DarkFace-val set with different settings: (a) YOLOv4 trained with DarkFace training set only; (b) YOLOv4 trained with DarkFace training set and (LTFT + FEEDS)-day-to-night-transformed images generated by NICE-GAN; (c) YOLOv4 trained with DarkFace training set and (LTFT + FEEDS)-day-to-night-transformed images generated by MUNIT; (d) YOLOv4 trained with DarkFace training set and (LTFT + FEEDS)-day-to-night-transformed images generated by AugGAN-Det.

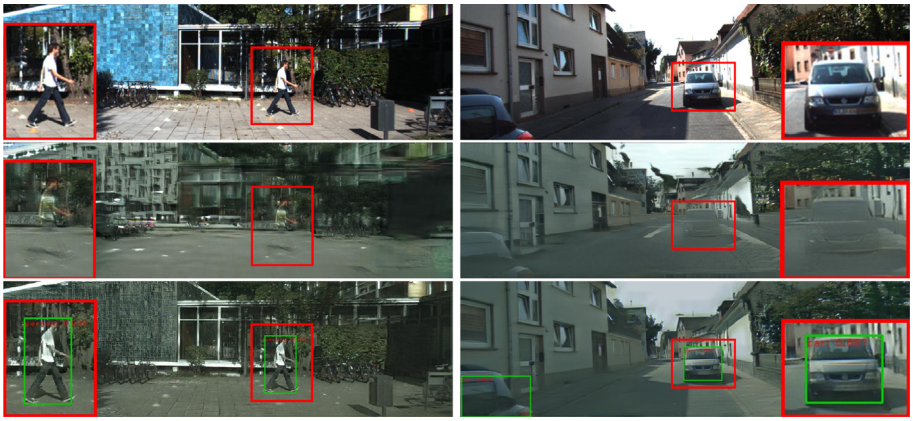
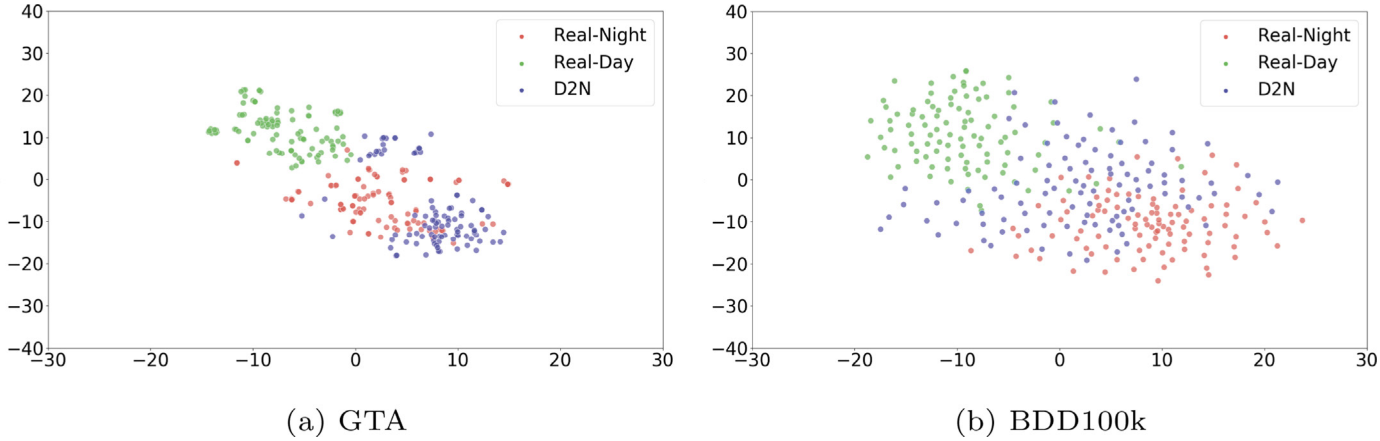


Fig. 14. KITTI-to-Cityscape results: (1st row) original images; (2nd row) detection results in MUNIT-transformed images (pedestrian and cars cannot be detected because they are not preserved during image-translation); (3rd row) detection results (green bounding boxes) of images transformed by this work. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to the generator and simply increases its loss. When the detector is jointly trained with the generator, the generator continuously learns to generate realistic and target-domain-detector-detectable objects.

5.2. Ablation study

Our network design assumes that object detectors, through CoCo loss, can serve as an auxiliary regularization for image-to-image translation. However, object consistency loss for both image-translation phases in both cycles might still be helpful to some extent in terms of detection accuracy. The FID and the detection analysis of our model variations is shown in Table 7. It is evident that CoCo loss is more important than the object consistency losses, as seen in the top two rows. In contrast, a single detector trained by using day and night images to perform object consistency loss and CoCo loss can only provide inferior results. Finally, as seen in the last row, the best results are achieved by regularizing the generators with daytime and night detectors. That is, a daytime detector for the image-reconstruction phase in the forward cycle and image-translation phase in the backward cycle; while a nighttime detector for the image-translation phase in the forward cycle and image-reconstruction phase in the backward cycle.



(a) GTA

(b) BDD100k

Fig. 15. t-SNE visualization results in GTA and BDD100k cases: Green/Red/Blue points are daytime/nighttime/day-to-night-transformed cropped vehicles, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 7
Ablation study of object-consistency and detector (trained with day, night or day + night) comparison in terms of FID and detection accuracy - detectors trained with transformed images (BDD100k-det-val-day) generated by GANs (learning from both GTA and BDD100k) and tested on BDD100k-det-val-night. For FID, lower is better.

Models	FID	AP
Two detectors in image-trans phases only	0.414	0.528
Two detectors in image-recons phases only	0.512	0.535
Only one detector in both phases and cycles	0.451	0.525
Two detectors in both phases and cycles	**0.309**	**0.559**

Table 8
Training a nighttime detector (YOLOv4) using different data: the testing data is GTA-val-night2.

Training data	AP
GTA-val-day	0.455
GTA-val-day + GTA-val-day-to-night	0.507
GTA-val-day-to-night(GAN w/ daytime detector only)	0.491
GTA-val-day-to-night	0.542
GTA-val-night1	0.563
GTA-val-day-to-night + GTA-val-night1	**0.622**

Table 9
FID scores, FCN scores, and MOS of the GTA case: note that FID is estimated between GTA-val-day-to-night transformed images and GTA-val-night images; FCN models are trained by the former and testing on the latter; MOS is evaluated by showing observers the former images of different models. For FID, lower is better.

GAN model	FID	FCN-Acc	FCN-IoU
NICE-GAN [37]	2.466	0.925	0.825
MUNIT [8]	1.066	0.939	0.831
AugGAN [11]	1.020	0.940	0.855
M-AugGAN [12]	1.023	0.941	0.860
AugGAN-Det	**0.799**	**0.947**	**0.887**

5.3. Different training data for a target-domain detector

We have conducted several experiments to know how the transformed images can help train a nighttime detector, as seen in Table 8. GTA-val-night dataset is split into equal-sized GTA-val-night1 & GTA-val-night2. We found that using GTA-val-day alone achieved the lowest accuracy. When training generators of both domains with daytime detector only, the image-translation quality degrades so using detectors of both domains is essential. Mixing day-to-night data with real-nighttime data achieved the highest accuracy, which proves that training a nighttime detector with real and synthetic data is valuable.

5.4. Additional subjective and objective evaluation

To further objectively evaluate the quality of the generated nighttime images in the GTA case, we use FID [58] for analyzing the similarity between the day-to-night-transformed images and the real nighttime ones. In theory, the day-to-night-transformed images are supposed to be also helpful in performing semantic segmentation for nighttime vehicles. We adopt the popular FCN8s (VGG16-based) [59] to report the FCN scores and completely follow the protocol mentioned by other GAN models, such as Lin et al. [12], for a fairer comparison. How much the day-to-night-transformed images help perform semantic segmentation for nighttime vehicles is also analyzed using the popular FCN8s (VGG16-based) [59] to report the FCN scores. Finally, in order to know if day-to-night-transformed images generated by our model, AugGAN-Det, are visually better, we also conducted a subjective evaluation of mean opinion score (MOS) to provide a visual rating (from one to five, the higher, the better) for our method and other competing ones. There are 51 random non-expert observers involved and the questions are designed to demonstrate three factors. The first one considers the instance-aware image fidelity. i.e., the color, and the location of the vehicle lamps in the day-to-night case. The second one is the overall style-transfer quality. The third one is the level of object preservation. These three factors are integral for determining if day-to-night-transformed images are realistic. As seen in Table 9, our model achieves the lowest FID because the day-to-night-transformed vehicles are more realistic in terms of the nighttime-looking texture of the vehicle body, brightness of vehicles' rear lamps, and the sharpness of vehicle's body at nighttime. Our model also leads to higher FCN scores because FCN is trained to better understand vehicle's nighttime appearance. In Table 10, the MOS comparison indicates that our work outperforms MUNIT and NICE-GAN in terms of the fidelity of the rear lamps and better object preservation. Multimodal AugGAN could achieve better object preservation than both MUNIT and NICE-GAN, but object appearances are not instance-aware enough.

5.5. t-SNE visualization

t-SNE [60] is a non-linear technique widely used for dimensionality reduction and can thus visualize high-dimensional data. This

Table 10
MOS of our model and other competing methods in both GTA and BDD100 cases: the former learns from GTA only and the latter learns from BDD100k + GTA.

GAN model	GTA	BDD100k
NICE-GAN [37]	3.049	2.289
MUNIT [8]	2.331	2.095
Multimodal AugGAN [12]	2.422	2.810
AugGAN-Det	**3.763**	**3.126**

powerful tool can find the patterns in the data by identifying observed clusters based on the similarity of data points with multiple features. t-SNE works well in our case because it can group objects with similar appearance together when they are mapped from high to low dimensions. Figure 15 shows the visualization results of t-SNE on GTA and BDD100k datasets. It is evident that most of the day-to-night-transformed images (blue dots) are very close to the real nighttime ones (red dots). That is, our proposed method has successfully made the transformed vehicles carry the characteristics of the real nighttime ones.

6. Conclusion and future work

In this paper, we propose CoCo loss to leverage object detectors for performing instance-aware image-translation via a GAN model. We empirically demonstrate that the generator learns to lower the detection loss for the objects to be transformed to the style of their counterparts in the target domain. Compared to previous models, which e.g., require pixel-level semantic segmentation to force the latent distribution to be object-preserving, this work only needs bounding box annotations which are significantly easier to acquire. As to the instance-aware GAN models, our model, AugGAN-Det, internalizes global and object style-transfer without explicitly aligning the instance features. Most importantly, a detector is not required at test time. Therefore, most published datasets for object detection become valuable, since labeled data in e.g., different weather conditions and times-of-the-day can be converted “for free” for an object detector to achieve better results in a designated scenario. The limitation of this work is that the objects in the image cannot be too small. Therefore, we trained the detector in our GAN model by the whole body of a person instead of his face in the DarkFace experiment. We plan to create a multimodal version of this model in the future. Therefore, a single annotation can be transformed in an instance-aware manner to multiple images in the target domain to further improve object detectors’ accuracy.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] A. Bochkovskiy, C.-Y. Wang, H.-Y. M. Liao, YOLOv4: optimal speed and accuracy of object detection(2020) arXiv e-prints, arXiv–2004.
[2] M. Braun, S. Krebs, F. Flohr, D.M. Gavrila, Eurocity persons: a novel benchmark for person detection in traffic scenes, IEEE Trans. Pattern Anal. Mach. Intell. 41 (8) (2019) 1844–1861.
[3] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, T. Darrell, BDD100K: a diverse driving video database with scalable annotation tooling, 2(5) (2018) 6. arXiv preprint arXiv:1805.04687
[4] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: ICCV, 2017, pp. 2223–2232.
[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.
[6] M.-Y. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, in: NIPS, 2017, pp. 1–9.
[7] Z. Yi, H. Zhang, P. Tan, M. Gong, DualGAN: unsupervised dual learning for image-to-image translation, in: ICCV, 2017, pp. 2849–2857.
[8] X. Huang, M.-Y. Liu, S. Belongie, J. Kautz, Multimodal unsupervised image–to-image translation, in: ECCV, 2018, pp. 172–189.
[9] W. Xu, K. Shawn, G. Wang, Toward learning a unified many-to-many mapping for diverse image translation, Pattern Recognit. 93 (2019) 570–580.
[10] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, T. Darrell, CyCADA: cycle-consistent adversarial domain adaptation, in: International Conference on Machine Learning, 2018, pp. 1989–1998.
[11] S.-W. Huang, C.-T. Lin, S.-P. Chen, Y.-Y. Wu, P.-H. Hsu, S.-H. Lai, AugGAN: cross domain adaptation with GAN-based data augmentation, in: ECCV, 2018, pp. 718–731.
[12] C.-T. Lin, Y.-Y. Wu, P.-H. Hsu, S.-H. Lai, Multimodal structure-consistent image–to-image translation, in: AAAI, 2020, pp. 11490–11498.
[13] R. Li, W. Cao, Q. Jiao, S. Wu, H.-S. Wong, Simplified unsupervised image translation for semantic segmentation adaptation, Pattern Recognit. 105 (2020) 107343.
[14] Z. Shen, M. Huang, J. Shi, X. Xue, T.S. Huang, Towards instance-level image–to-image translation, in: CVPR, 2019, pp. 3683–3692.
[15] D. Bhattacharjee, S. Kim, G. Vizier, M. Salzmann, DUNIT: detection-based unsupervised image-to-image translation, in: CVPR, 2020, pp. 4787–4796.
[16] S.R. Richter, Z. Hayder, V. Koltun, Playing for benchmarks, in: ICCV, 2017, pp. 2213–2222.
[17] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: CVPR, 2016, pp. 779–788.
[18] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: CVPR, 2017, pp. 7263–7271.
[19] J. Redmon, A. Farhadi, YOLOv3: an incremental improvement(2018). arXiv e-prints, arXiv–1804
[20] W. Ma, Y. Wu, F. Cen, G. Wang, MDFN: multi-scale deep feature learning network for object detection, Pattern Recognit. 100 (2020) 107149.
[21] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: CVPR, 2014, pp. 580–587.
[22] R. Girshick, Fast R-CNN, in: ICCV, 2015, pp. 1440–1448.
[23] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2016) 1137–1149.
[24] J. Dai, Y. Li, K. He, J. Sun, R-FCN: object detection via region-based fully convolutional networks, in: NIPS, 2016, pp. 379–387.
[25] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, D. Lin, Libra R-CNN: towards balanced learning for object detection, in: CVPR, 2019, pp. 821–830.
[26] J. Peng, H. Wang, S. Yue, Z. Zhang, Context-aware co-supervision for accurate object detection, Pattern Recognit. 121 (2022) 108199.
[27] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 13001–13008.
[28] T. DeVries, G.W. Taylor, Improved regularization of convolutional neural networks with cutout (2017) arXiv e-prints, arXiv–1708
[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.
[30] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, R. Fergus, Regularization of neural networks using dropconnect, in: ICML, 2013, pp. 1058–1066.
[31] G. Ghiasi, T.-Y. Lin, Q.V. Le, DropBlock: a regularization method for convolutional networks, in: NIPS, 2018, pp. 10750–10760.
[32] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, Mixup: beyond empirical risk minimization (2017) arXiv e-prints, arXiv–1710
[33] S. Yun, D. Han, S.J. Oh, S. Chun, J. Choe, Y. Yoo, CutMix: regularization strategy to train strong classifiers with localizable features, in: ICCV, 2019, pp. 6023–6032.
[34] K. Baek, D. Bang, H. Shim, GridMix: strong regularization through local context mapping, Pattern Recognit. 109 (2021) 107594.
[35] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: CVPR, 2017, pp. 1125–1134.
[36] C. Chu, A. Zhmoginov, M. Sandler, CycleGAN, a master of steganography (2017) arXiv e-prints, arXiv–1712
[37] R. Chen, W. Huang, B. Huang, F. Sun, B. Fang, Reusing discriminators for encoding: Towards unsupervised image-to-image translation, in: CVPR, 2020, pp. 8168–8177.
[38] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A.A. Efros, O. Wang, E. Shechtman, Toward multimodal image-to-image translation, in: NIPS, 2017, pp. 465–476.
[39] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, M.-H. Yang, Diverse image-to-image translation via disentangled representations, in: ECCV, 2018, pp. 35–51.
[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: ECCV, 2014, pp. 740–755.
[41] R. Müller, S. Kornblith, G. Hinton, When does label smoothing help? in: NIPS, 2019, pp. 4694–4703.

[42] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.
[43] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, S. Paul Smolley, Least squares generative adversarial networks, in: ICCV, 2017, pp. 2794–2802.
[44] Y. Taigman, A. Polyak, L. Wolf, Unsupervised cross-domain image generation (2016) arXiv e-prints, arXiv–1611
[45] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The KITTI vision benchmark suite, in: CVPR, 2012, pp. 3354–3361.
[46] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: CVPR, 2016, pp. 3213–3223.
[47] G.J. Brostow, J. Fauqueur, R. Cipolla, Semantic object classes in video: a high--definition ground truth database, Pattern Recognit. Lett. 30 (2) (2009) 88–97.
[48] W. Yang, Y. Yuan, W. Ren, J. Liu, W.J. Scheirer, Z. Wang, T. Zhang, Q. Zhong, D. Xie, S. Pu, et al., Advancing image understanding in poor visibility environments: a collective benchmark study, IEEE Trans. Image Process. 29 (2020) 5737–5752.
[49] FEEDS (Face pEdestrain dEtection DataSet) dataset, 2019, https://github.com/neverland7D/Face-and-Pedestrian-Detection-Dataset.
[50] G. Barquero, C. Fernández, I. Hupont, Long-term face tracking for crowded video-surveillance scenarios, in: IJCB, 2020, pp. 1–8.
[51] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in PyTorch, in: NIPS Workshops, 2017, pp. 1–4.
[52] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization (2014) arXiv preprint arXiv:1412.6980.
[53] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training GANs, in: NIPS, 2016, pp. 2234–2242.
[54] N. Inoue, R. Furuta, T. Yamasaki, K. Aizawa, Cross-domain weakly-supervised object detection through progressive domain adaptation, in: CVPR, 2018, pp. 5001–5009.
[55] Y. Chen, W. Li, C. Sakaridis, D. Dai, L. Van Gool, Domain adaptive Faster R-CNN for object detection in the wild, in: CVPR, 2018, pp. 3339–3348.
[56] T. Kim, M. Jeong, S. Kim, S. Choi, C. Kim, Diversify and match: a domain adaptive representation learning paradigm for object detection, in: CVPR, 2019, pp. 3339–3348.
[57] A.L. Rodriguez, K. Mikolajczyk, Domain adaptation for object detection via style consistency, in: BMVC, 2019, pp. 1–14.
[58] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local Nash equilibrium, in: NIPS, 2017, pp. 6629–6640.
[59] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: CVPR, 2015, pp. 3431–3440.
[60] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008) 2579–2605.

Che-Tsung Lin is currently a postdoctoral researcher at Chalmers University of Technology, Sweden. His research is mainly about object detection, semantic segmentation, domain adaptation and their applications in ADAS and autonomous vehicles.

Jie-Long Kew is currently a research assistant at the University of Malaya, Malaysia. His current research of interest focuses on computer vision especially in image translation, image super resolution and object detection.

Chee Seng Chan is currently a Full Professor with the Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia. From 2020 till 2022, he was seconded to the Ministry of Science, Technology and Innovation (MOSTI) in Putrajaya, Malaysia as an undersecretary. His research interests include computer vision and machine learning.

Shang-Hong Lai is a professor in the department of computer science at National Tsing Hua University, Hsinchu, Taiwan. Since the summer of 2018, He has been on leave to Microsoft AI R&D Center in Taipei as a principal researcher. He served as an area chair for several top computer vision conferences and an associate editor for Pattern Recognition and Journal of Signal Processing Systems.

Christopher Zach is a research professor at Chalmers University of Technology, Sweden, with research interests at the intersection of 3D computer vision, optimization methods and machine learning.