Multimodal Structure-Consistent Image-to-Image Translation

Che-Tsung Lin,[1,2] **Yen-Yi Wu,**[1] **Po-Hao Hsu,**[1] **Shang-Hong Lai**[1,3]
[1]Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
[2]Intelligent Mobility Division, Mechanical and Mechatronics Systems Research Laboratories,
Industrial Technology Research Institute, Hsinchu, Taiwan
[3]Microsoft AI R&D Center, Taipei, Taiwan
AlexLin@itri.org.tw, u106062542@m106.nthu.edu.tw, sheephow@gapp.nthu.edu.tw, lai@cs.nthu.edu.tw

Abstract

Unpaired image-to-image translation is proven quite effective in boosting a CNN-based object detector for a different domain by means of data augmentation that can well preserve the image-objects in the translated images. Recently, multimodal GAN (Generative Adversarial Network) models have been proposed and were expected to further boost the detector accuracy by generating a diverse collection of images in the target domain, given only a single/labelled image in the source domain. However, images generated by multimodal GANs would achieve even worse detection accuracy than the ones by a unimodal GAN with better object preservation. In this work, we introduce cycle-structure consistency for generating diverse and structure-preserved translated images across complex domains, such as between day and night, for object detector training. Qualitative results show that our model, Multimodal AugGAN, can generate diverse and realistic images for the target domain. For quantitative comparisons, we evaluate other competing methods and ours by using the generated images to train YOLO, Faster R-CNN and FCN models and prove that our model achieves significant improvement and outperforms other methods on the detection accuracies and the FCN scores. Also, we demonstrate that our model could provide more diverse object appearances in the target domain through comparison on the perceptual distance metric.

Introduction

Recent advances of object detection are driven by the success of two-stage detectors, (Girshick et al. 2014; Girshick 2015; Ren et al. 2015). While two-stage detectors are generally more accurate but slower, one-stage detectors (Redmon et al. 2016; Redmon and Farhadi 2017; Liu et al. 2016) have been proposed for real-time performance. These detectors keep pushing the limits of object detection on benchmark datasets, such as PASCAL VOC (Everingham et al. 2010) and MSCOCO (Lin et al. 2014).

The generalization capability of CNN-based detectors is way better than traditional machine learning approaches. However, performance still drops significantly when the trained model is deployed in a new domain different from those of the training images. Taking the on-road object detection for example, one of the most complex domain shift is between day and night because the object appearances such as vehicles at daytime are very different from their counterparts at nighttime. As indicated by (Braun et al. 2018) in pedestrian detection, training on daytime and testing on night-time gives significantly worse results than training and testing on the same time-of-day. The reason that nighttime vehicle datasets in real-driving scenario are scarce in public domain is that the labeling cost at nighttime is significantly more expensive than that at daytime.

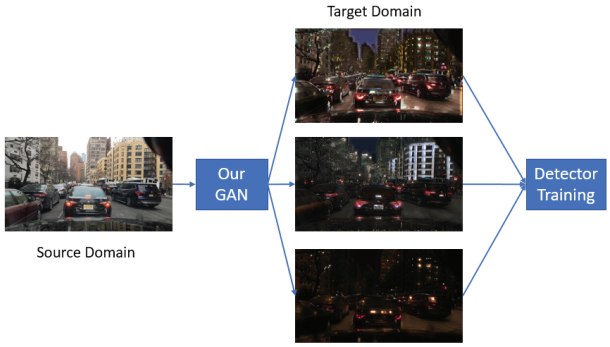

Figure 1: How our multimodal GAN helps an object detector adapt to a target domain: Our GAN model could provide structure-consistent translated images across complex domains.

The successes of Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) on image generation have made rapid progress in this area. The generators in GANs can map from noise vectors to realistic images. However, to apply GAN for improving the robustness of an on-road object detection in terms of data augmentation, it is more practical to perform image-to-image translation for the labeled data in the source domain to the target domain. This way, the tedious data annotation could be significantly mitigated. Pix2Pix (Isola et al. 2017) could provide visually-plausible images in the target domain if paired training data is available, which is not possible for on-road object detection.

Recently, unpaired image-to-image translation methods,

such as CycleGAN (Zhu et al. 2017a), have achieved astonishing results by introducing the cycle consistency constraint. UNIT (Liu, Breuel, and Kautz 2017) further applied weight-sharing constraint to increase the translation consistency. AugGAN (Huang et al. 2018a) presented a weighting-shared multi-tasking generator consisting of an image-translation subtask and a segmentation one to pursue better translation results in the target domain while the image-objects are well-preserved. CyCADA (Hoffman et al. 2018) proposed to improve domain adaptation translation by (1) applying the segmentation task loss in the forward cycle and the semantic consistency loss in the backward cycle, and (2) aligning representations at both pixel-level and feature-level. (Inoue et al. 2018) presented a cross-domain weakly supervised object detection framework by fine-tuning a detector well-trained in the source domain with (1) domain-transferred images and (2) target-domain images with image-level annotations. (Chen et al. 2018) tried to improve the cross-domain robustness in the image-level and the instance-level. The two domain adaptation components of their work are implemented by learning domain classifiers on different levels in adversarial training manner. The image translation made above were all one-to-one mapping, which limits the diversity of the generated images, given the labelled data in the source domain. BicycleGAN (Zhu et al. 2017b) is a multimodal image-to-image translation model and could provide diverse output. However, it requires paired data and is thus not suitable for many applications, such as on-road image transformation.

More recently, instead of one-to-one mapping in unpaired image-to-image translation, multimodal image-translation models, such as AugCGAN (Almahairi et al. 2018), DRIT (Lee et al. 2018) and MUNIT (Huang et al. 2018b), have been proposed to provide diverse outputs using a single model. Therefore, given unpaired labelled data in different domains, the GAN models would learn to produce various outputs in terms of different styles, appearances, or observations. However, in complex domain shift, such as day-to-night, the image-objects may not be well-preserved because specific objects' appearances would normally not dominate the model training in learning the image translation.

In this paper, we propose, Multimodal AugGAN, a multimodal structure-consistent image-to-image translation network, which directly benefits object detection by translating each existing detection training image from its original domain to diverse results, each of which possesses different degrees of transformation at the target domain. The contribution of this work is three-fold: (1) we design a multimodal structure-consistent image-to-image translation network which learns from unpaired images to generate diverse images in the target domain while artifact in the transformed images is greatly reduced; (2) we quantitatively prove that the domain adaptation capability of a vehicle detector could be further boosted by incorporating multimodal transformed images in detector training; (3) our multimodal GAN model provides significant performance gain in the difficult day-to-night case in terms of vehicle detection and semantic segmentation.

Proposed Model

Our goal is to learn a multimodal structure-consistent image-translation network between two visual domains $X \subset \mathbb{R}^{H\times W\times 3}$ and $Y \subset \mathbb{R}^{H\times W\times 3}$ where the 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 in the learning stage, and the transformation between two domains is learned in the unpaired fashion.

Our network, as depicted in Fig. 2, consists of image-translation encoders $\{E_x, E_y\}$, parsing net encoders $\{E_x^p, E_y^p\}$, image-translation generators $\{G_x, G_y\}$, parsing nets $\{P_x, P_y\}$, and discriminators $\{D_x, D_y\}$ for the two image domains, respectively.

While unpaired image-to-image translation tasks have achieved success in generating realistic images, they are primarily limited to generating a deterministic output $\bar{y}$, given the input image x. In order to learn the mapping that could sample the output $\bar{y}$ from true conditional distribution given x and produce results which are both diverse and realistic, we would like to learn a low-dimensional latent vector $z \subset \mathbb{R}^Z$, which encapsulates the ambiguous aspects of the output mode. For example, a vehicle in a daytime image could map to their nighttime counterparts with different ambient light levels and rear lamp conditions (on/off) but with the same vehicle type, color and locations. In this work, we aim at learning a deterministic mapping $G_x(x, z) \rightarrow y$ to the output. To enable stochastic sampling, we desire the latent code vector z to be drawn from some prior distribution $p(z)$. We use a standard Gaussian distribution $N(0, 1)$ in this work.

As shown in Fig. 2, generator G_x generates images conditioned on the random vector drawn from a prior Gaussian distribution $N(0, 1)$ in an attempt to fool discriminator D_x and be structure-consistent in image translation by fulfilling the structure required to perform segmentation subtask done by P_y. Then, to further reduce the space of possible mappings, we claim that the mappings should be cycle-structure consistent. Therefore, given the translated $\bar{y}$, G_y will also try to generate a reconstructed image x_{rec} which is not capable of being discriminated by D_y but still preserves the structure in a way that it could be segmented by P_x to produce $\hat{x}_{rec}$ closer to $\hat{x}$. The backward cycle learns to translate image starting from domain Y basically in the same way. Therefore, our network learns the multimodal image translation to the target domain according to its structure in the source domain.

Detailed architecture of our network is given in Table 1. We use the same encoder structure for the subsequent generators and parsing networks. There are some design choices for modeling stochastic mapping. Injecting random noise by concatenation was used in (Zhu et al. 2017b) and (Radford, Metz, and Chintala 2015) while the former indicates that there is no significant difference between input-injection and all-layer-injection. However, similar to (Almahairi et al. 2018), our model also achieves more high-level variations by using Conditional Instance Normalization (CIN). For the discriminators, we follow the design of PatchGAN (Isola et al. 2017) because it is flexible to work on arbitrarily-sized images in a fully convolutional fashion.

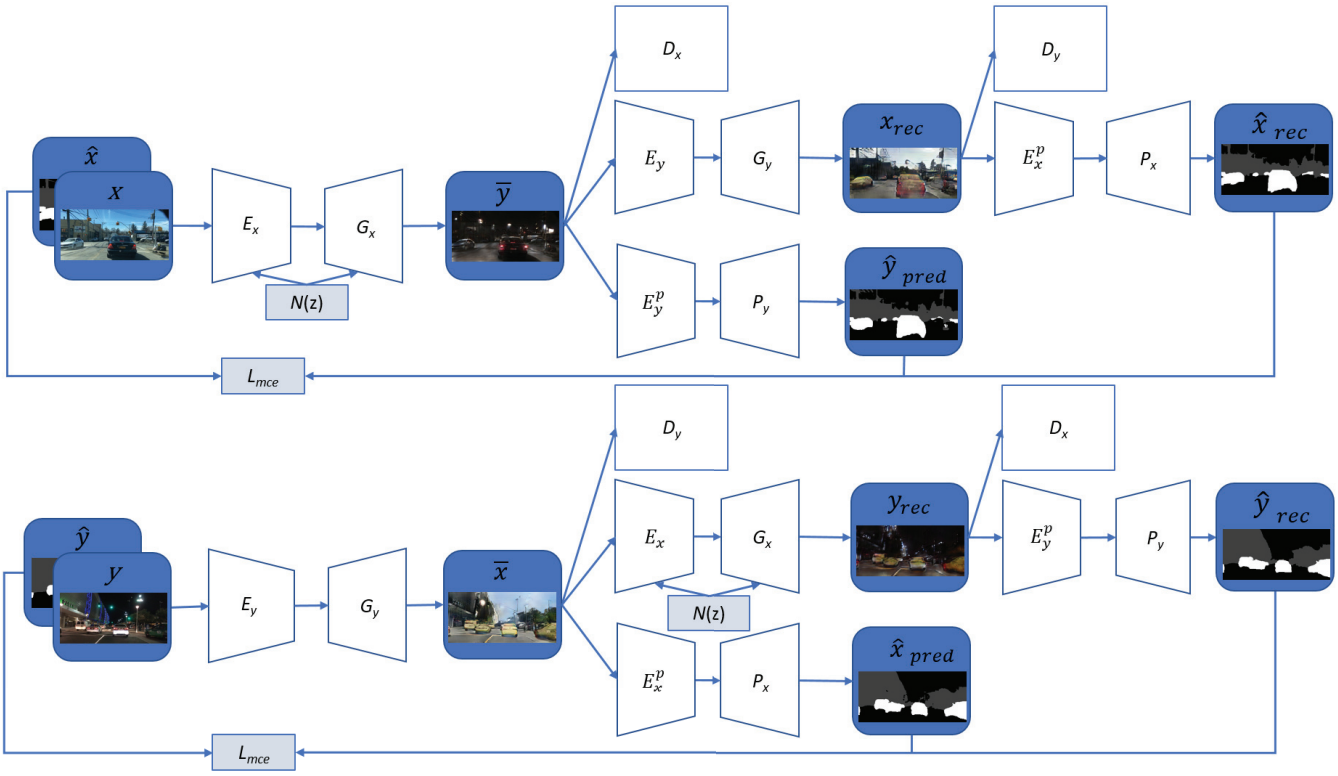

Figure 2: Overall structure of the proposed multimodal structure-consistent image-to-image translation network: x, y: sampled images from domain X and Y; $\hat{x}$, $\hat{y}$: segmentation Ground-Truth of x and y; $\bar{x}$, $\bar{y}$: translated results; $\hat{x}_{pred}$, $\hat{y}_{pred}$: predicted segmentation masks, given $\bar{x}$ and $\bar{y}$; x_{rec}, y_{rec}: reconstructed images corresponding to x and y; $\hat{x}_{rec}$, $\hat{y}_{rec}$: predicted segmentation masks, given reconstructed images x_{rec}, y_{rec}.

Adversarial learning

We first apply adversarial losses to the mapping functions. Let Z be an auxiliary noise with a standard Gaussian prior. The mapping function between two domains is firstly learned with a deterministic mapping from X domain to the structure-aware latent space Z_x and then translated to Y domain while both processes are conditioned on the same auxiliary noise. i.e., $E_x : \mathrm{X} \times \mathrm{Z} \rightarrow \mathrm{Z_x}$, $G_x : \mathrm{Z_x} \times \mathrm{Z} \rightarrow \mathrm{Y}$. We define the first adversarial loss function as:

$$\mathcal{L}_{GAN_1}(E_x, G_x, D_x, X, Y) = E_{y\sim p_{data(y)}}[\log D_x(y)] + E_{x\sim p_{data(x)}, z\sim p_{(z)}}[\log(1 - D_x(G_x(E_x(x, z), z)))], \quad (1)$$

where E_x and G_x try to generate transformed images $G_x(E_x(x))$ that look similar to images from domain Y, while D_x aims to distinguish between translated samples $G_x(E_x(x))$ and real samples y in terms of style. The same mapping function could be applied to the reconstruction phase of the backward cycle and the only difference is $E_x : \bar{\mathrm{X}} \times \mathrm{Z} \rightarrow \mathrm{Z_x}$, $G_x : \mathrm{Z_x} \times \mathrm{Z} \rightarrow \mathrm{Y}$ with the objective $\mathcal{L}_{GAN_1}(E_x, G_x, D_x, \bar{X}, Y)$. In the testing stage for stochastic sampling, we desire the auxiliary noise to be drawn from some prior distribution $p(z)$, which is still a standard Gaussian distribution $N(0, 1)$ throughout this work.

Similarly, for the image-translation phase of the backward cycle, the mapping function $E_y : \mathrm{Y} \rightarrow \mathrm{Z_y}$, $G_y : \mathrm{Z_y} \rightarrow \mathrm{X}$ and its discriminator D_y are related in the following adversarial loss:

$$\mathcal{L}_{GAN_2}(E_y, G_y, D_y, Y, X) = E_{x\sim p_{data(x)}}[\log D_y(x)] + E_{y\sim p_{data(y)}}[\log(1 - D_y(G_y(E_y(y))))], \quad (2)$$

and the reconstruction phase of the forward cycle is modeled as $\mathcal{L}_{GAN_2}(E_y, G_y, D_y, \bar{Y}, X)$.

Image-translation-structure consistency

Our model actively guides the encoder-generator network to extract structure-aware features conditioned on the injected random vector so that the image is transformed to the style of the target domain while its structure is consistent with the structure of the corresponding image in the source domain. Here, we use the multi-class cross-entropy loss for the segmentation subtask in preserving structure for the image-translation phase of the forward cycle and it could be formulated as:

$$\mathcal{L}_{seg1}(E_x, G_x, E_y^p, P_y, X, \hat{X}) = \mathbb{E}_{x,\hat{x}\sim p_{data(x,\hat{x})}, z\sim p_{(z)}}[\ell_{mce}(\hat{x}_{pred}, \hat{x})], \quad (3)$$

Table 1: Network architecture of encoders, generators, discriminators and parsing nets in our multimodal structure-consistent image-to-image translation model: N, K, S, Nseg denote the number of convolution filters, kernel size, stride, and task-specific number of segmentation classes, respectively. Every Convolution and Deconvolution layer is followed by a Batch Normalization layer when there is no latent vector injection or a Conditional Instance Normalization layer when there is, except the last layer of each generator, parsing net and discriminator.

Layer	Encoders	Layer Info
1	CONV, ReLU	N48,K7,S1
2	CONV, ReLU	N96,K3,S1
3	CONV, ReLU	N192,K3,S2
4	CONV, ReLU	N192,K3,S2
5	RESBLK, ReLU	N192,K3,S1
Layer	Generators	Layer Info
1	RESBLK, ReLU	N192,K3,S1
2	RESBLK, ReLU	N192,K3,S1
3	DCONV, ReLU	N192,K3,S2
4	DCONV, ReLU	N96,K3,S2
5	CONV, Tanh	N3,K7,S1
Layer	Parsing Networks	Layer Info
1	RESBLK, ReLU	N192,K3,S1
2	RESBLK, ReLU	N192,K3,S1
3	DCONV, ReLU	N192,K3,S2
4	DCONV, ReLU	N96,K3,S2
5	CONV, ELU	Nseg,K7,S1
6	CONV, Softmax	Nseg,K1,S1
Layer	Discriminator	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

where $\hat{x}_{pred} = P_y(E_y^p(G_x(E_x(x,z),z)))$ and $\ell_{mce}(\hat{x}_{pred},\hat{x})$ denotes the multi-class cross-entropy loss which could be formulated as $-\frac{1}{H\times W}\sum_{i=1}^{H\times W}\sum_{c=1}^{C}\hat{x}_c(i)log(\hat{x}_{pred,c}(i))$.

For the backward cycle, the image-translation-structure consistency is modeled without the random vector:

$$\mathcal{L}_{seg2}(E_y,G_y,E_x^p,P_x,Y,\hat{Y}) = \mathbb{E}_{y,\hat{y}\sim p_{data(y,\hat{y})}}[\ell_{mce}P_x(E_x^p(G_y(E_y(y)))),\hat{y})]. \quad (4)$$

Cycle-structure Consistency

Cycle consistency was proposed in CycleGAN for producing images in another domain without the use of pairing information. However, as pointed out by (Almahairi et al. 2018), enforcing the learned mapping to be cycle-consistent is the fundamental flaw in modeling multimodal conditionals for the reconstruction error would encourage the mappings to ignore the random vector. Here, we propose cycle-structure consistency which does not force X and X_{rec} to be exactly the same when cycling $X \to \bar{Y} \to X_{rec}$. Instead, our cycle-structure consistency only encourages $\hat{X}$ and $\hat{X}_{rec}$ to be as close as possible. This way, only structure of the reconstructed image is preserved while the transferred-style is not constrained. The associated loss functions are given by

$$\mathcal{L}_{cycle1}(E_x,G_x,E_y,G_y,E_x^p,P_x,X,\hat{X}) = \mathbb{E}_{x,\hat{x}\sim p_{data(x,\hat{x})},z\sim p_{(z)}}[\ell_{mce}(\hat{x}_{rec},\hat{x})], \quad (5)$$

$$\mathcal{L}_{cycle2}(E_y,G_y,E_x,G_x,E_y^p,P_y,Y,\hat{Y}) = \mathbb{E}_{y,\hat{y}\sim p_{data(y,\hat{y})},z\sim p_{(z)}}[\ell_{mce}(\hat{y}_{rec},\hat{y})], \quad (6)$$

where $\hat{x}_{rec} = P_x(E_x^p(G_y(E_y(G_x(E_x(x,z),z))))$, $\hat{y}_{rec} = P_y(E_y^p(G_x(E_x(G_y(E_y(y)),z),z)))$, and we still apply the multi-class cross-entropy loss here.

Network Learning

We jointly solve the learning problems for the image-translation subtask:$\{E_x,G_x,D_x\}$ and $\{E_y,G_y,D_y\}$, the image parsing subtask: $\{E_x,G_x,E_y^p,P_y\}$ and $\{E_y,G_y,E_x^p,P_x\}$, to be image-translation-structure-consistent and cycle-structure-consistent. The full objective function is given as follows:

$$\begin{aligned}\mathcal{L}_{full}(E_x,G_x,E_y,G_y,E_x^p,P_x,E_y^p,P_y,D_x,D_y) = \\ \mathcal{L}_{GAN_1}(E_x,G_x,D_x,X,Y) \\ +\mathcal{L}_{GAN_2}(E_y,G_y,D_y,Y,X) \\ +\mathcal{L}_{GAN_1}(E_x,G_x,D_x,\bar{X},Y) \\ +\mathcal{L}_{GAN_2}(E_y,G_y,D_y,\bar{Y},X) \\ +\mathcal{L}_{seg1}(E_x,G_x,E_y^p,P_y,X,\hat{X}) \\ +\mathcal{L}_{seg2}(E_y,G_y,E_x^p,P_x,Y,\hat{Y}) \\ +\lambda_{cyc}*(\mathcal{L}_{cycle1}(E_x,G_x,E_y,G_y,E_x^p,P_x,X,\hat{X}) \\ +\mathcal{L}_{cycle2}(E_y,G_y,E_x,G_x,E_y^p,P_y,Y,\hat{Y}),\end{aligned} \quad (7)$$

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

$$\min_{\substack{E_x,G_xD_x,D_y\\E_y,G_y,\\E_x^p,P_x,\\E_y^p,P_y}}\max \mathcal{L}_{full}(E_x,G_x,E_y,G_y,E_x^p,P_x,E_y^p,P_y,D_x,D_y). \quad (8)$$

Experimental Results

Advanced driver assistance systems (ADAS) or autonomous vehicles are expected to function well at both daytime and nighttime. However, most vehicle-related datasets (Yang et al. 2015; Zhou et al. 2016; Sivaraman and Trivedi 2010; Geiger, Lenz, and Urtasun 2012) in public domain were

captured at daytime. Synthetic datasets, such as SYNTHIA (Ros et al. 2016) and GTA (Richter et al. 2016), and real-driving datasets, such as BDD100k (Yu et al. 2018), provide detection and segmentation Ground-Truth in driving scenarios including different weather and time-of-day. Since our network includes segmentation subtask for better preserving the image-structure in image translation, we conduct GAN model training and detector testing in SYNTHIA, GTA and BDD100k, respectively. Also, each GAN model learning from SYNTHIA, GTA, or a mixture of GTA and BDD100k is also verified on BDD100k.

In SYNTHIA, only images in stereo-left are adopted and the day-to-night GAN training is performed with the spring and night images in sequences other than sequence-1. The data used for detector training and testing come from seq-1-spring transformed by GANs and seq-1-night images. In GTA, all the daytime and the nighttime images in training sets are used in GAN training and the daytime images in validation set are transformed by GANs to train detectors which would be later assessed by the nighttime validation images. In BDD100k, the GAN training is done by using BDD100k-seg-train. The detectors are trained with day-to-night-transformed BDD100k-val-day and tested on BDD100k-val-night.

We applied both one-stage YOLO (Redmon et al. 2016) and two-stage Faster R-CNN (VGG-16) (Ren et al. 2015) detectors in assessing how well the day-to-night transformation is done by each GAN model in terms of vehicle detection. Except that both detectors are revised to perform single-class vehicle detection, all hyper-parameters and evaluation metrics follow the same setting in their PASCAL VOC results. Besides, we found the best strategy to boost the detection accuracy is to train the vehicle detector with unimodal images first and then fine-tune the detector with multimodal data. The learning rate of both detectors is 1e-4 in the fine-tuning stage. It is worth mentioning that each unimodal image generated by a multimodal GAN model is done by simply feeding zero vector to the network throughout this work. Several other detector training strategies will also be discussed in this section.

Our major competitors in unimodal GANs which also utilize segmentation subtask include CyCADA and AugGAN, while the former originally only applies the segmentation task loss in the forward cycle only, and the final version of the latter has proved that using segmentation subtasks in both cycles is quantitatively beneficial. Since the segmentation Ground-Truth at both daytime and nighttime is available in our target datasets, we revise CyCADA to perform segmentation subtasks in both domains for a fairer comparison with AugGAN and our model.

This work is implemented in PyTorch (Paszke et al. 2017). We use the input image size of 256x152 for SYNTHIA and 320x152 for both GTA and BDD100k datasets. We set the size of the random vector $Z \in R^{16}$ throughout this work. For training, we use the Adam optimizer (Kingma and Ba 2015) with a batch size of 4, a learning rate of 0.0002, exponential decay rates $(\beta_1, \beta_2) = (0.5, 0.999)$. In all the experiments, we set the weightings related to structure consistency in the multi-task loss to be $\mathcal{L}_{seg1} = \mathcal{L}_{seg2} = \mathcal{L}_{cycle1} = \mathcal{L}_{cycle2} = 5$; others are all set to 1.

Synthetic Datasets

We first evaluate nighttime detector training using day-to-night-transformed images in synthetic datasets. The images transformed by AugGAN and CyCADA are quantitatively better than the ones by other unimodal GAN models including CycleGAN and UNIT. AugCGAN, MUNIT and our model, Multimodal AugGAN, all provide both unimodal and multimodal day-to-night transformed images. As shown in Table 2 and Table 3, Multimodal AugGAN outperforms competing methods in terms of nighttime detection accuracy.

Visually, the transformation results of Multimodal AugGAN are clearly better than the competing methods in terms of realism, diversity, and image-object preservation as shown in Fig. 3 and Fig. 4. It is worth mentioning that other multimodal models would potentially produce a number of nearly black pixels even on vehicle body inside image for achieving more diversity, which is harmful for training a vehicle detector because the detector would be struggled to learn something from nothing. However, our network would try to transform images while maintaining the structure-consistency in both the image-translation phase and the image-reconstruction one. Therefore, every day-to-night transformed image is beneficial for the detector to gradually learn different appearances of vehicles under different levels of ambient light at nighttime.

Table 2: Detection accuracy comparison (AP) - detectors trained with (SYNTHIA-seq-1-spring) images day-to-night-transformed by GANs (trained with SYNTHIA dataset sequences other than seq-1), and tested with night sequence (SYNTHIA-seq-1-night).

CyCADA	AugGAN	AugCGAN	MUNIT	Ours	Detector
39.5	39.0	28.2	33.7	42.6	YOLO
72.6	72.2	55.1	68.5	73.5	FRCN

Table 3: Detection accuracy comparison (AP) - detectors trained with (GTA-train-day) images day-to-night-transformed by GANs (trained with training set in GTA dataset), and tested with nighttime images in GTA validation set.

CyCADA	AugGAN	AugCGAN	MUNIT	Ours	Detector
25.1	25.3	24.8	22.8	33.0	YOLO
66.2	67.4	61.3	62.0	68.5	FRCN

BDD100k datasets

Real-driving BDD100k dataset provides on-road object detection labels and segmentation ones. However, there are no day/night attributes in BDD100k-seg-train. We later found that the nighttime labelled images are quite limited after manually classifying them to either day or night. Therefore, we also try to train the GAN models using synthetic, real, and synthetic + real datasets.

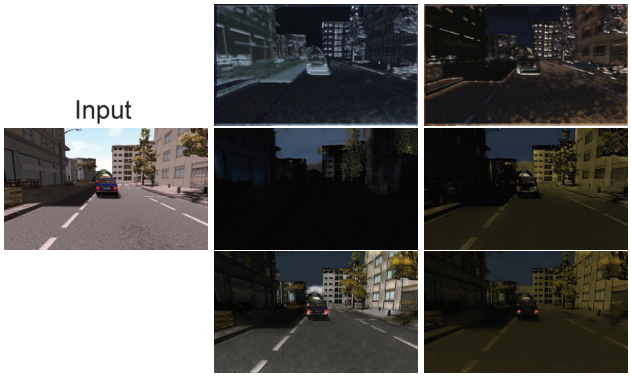

Figure 3: SYNTHIA day-to-night transformation results generated by GANs trained with SYNTHIA given different random vectors: 1st row: results of AugCGAN; 2nd row: results of MUNIT; 3rd row: results of this work.

Figure 4: GTA-val day-to-night transformation results generated by GANs trained with GTA-train given different random vectors: 1st row: results of AugCGAN; 2nd row: results of MUNIT; 3rd row: results of this work.

Our model consistently outperforms others in every dataset. As shown in Table 4, the highest detection accuracy is reported by using the training data generated by our GAN model learning from GTA and BDD100k and the transformed results are shown in Fig. 5

The detection results corresponding to the detector trained by using both datasets are shown in Fig. 6. A nighttime vehicle detector trained by using day-to-night-transformed data from AugCGAN or MUNIT would easily encounter difficulties in determining the exact boundary of vehicles especially in the images captured under low light. Our model could provide visually-appealing results in terms of realism under different ambient light levels at nighttime. Therefore, the nighttime vehicle detector would learn to better understand diverse appearances of vehicles at nighttime.

Semantic Segmentation Across Domains

This work has been proven effective in boosting the nighttime vehicle detection accuracy. However, the analysis is done in vehicles only. In order to evaluate the quality of the entire transformed image, we also adopt the popular FCN8s

Figure 5: BDD100k-val-day day-to-night transformation results generated by GANs trained with BDD100k-seg-train + GTA-train given different random vectors: 1st row: results of AugCGAN; 2nd row: results of MUNIT; 3rd row: results of this work.

Table 4: Detection accuracy comparison (AP) - YOLO & Faster R-CNN trained with (BDD100k-val-day) images day-to-night-transformed by GANs (trained with S: SYNTHIA sequences other than seq-1, G: GTA-train, B: BDD100k-seg-train, B+G: BDD100k-seg-train + GTA-train and tested on BDD100k-val-night.

Data	CyCADA	AugGAN	AugCGAN	MUNIT	Ours	Detector
S	39.5	39.2	37.1	39.1	41.0	YOLO
G	38.2	38.3	37.9	37.2	39.2	YOLO
B	39.0	37.9	32.0	36.7	40.4	YOLO
B+G	39.3	38.1	34.1	35.9	41.9	YOLO
S	62.0	61.7	51.2	60.7	64.5	FRCN
G	63.0	63.3	56.9	56.7	64.7	FRCN
B	62.2	62.1	51.8	58.1	64.5	FRCN
B+G	64.2	64.4	53.5	59.9	67.0	FRCN

(VGG16-based) (Long, Shelhamer, and Darrell 2015) to report the FCN score as Pix2Pix and CycleGAN did. The intuition is that if the day-to-night transformed images are realistic, then FCN8s could be trained by them to achieve better segmentation results on real nighttime images. The analysis is done in SYNTHIA dataset. We follow the same protocol as we did in the detector analysis, i.e., the model is firstly trained by unimodal data and then fine-tuned by multimodal data. In our experiments, the images are all re-sized to 600×600. For unimodal models, CycleGAN, UNIT, CyCADA and AugGAN, the FCNs are trained for 100k iterations. For multimodal models, AugCGAN, MUNIT and this work, the FCNs are trained with their unimodal data for 90k iterations and fine-tuned with their multimodal data for 10k iterations. The learning rates for unimodal data training are set to 1e-10 and the ones for multimodal data training are 1e-11. As can be seen in Table 5, the higher per-class accuracy, mIoU and fwIoU have shown that this work, Multimodal AugGAN, consistently outperforms other models on most of the classes.

Figure 6: Faster R-CNN (trained with unimodal and multimodal BDD100k-val-day day-to-night-transformed images generated by GANs learning from BDD100k-seg-train and GTA-train) detection result comparison on BDD100k-val-night: 1st row: results using training images generated by AugCGAN; 2nd row: results using training images generated by MUNIT; 3rd row: results using training images generated by this work.

Image Quality and Diversity Evaluation

We conduct a user study to evaluate realism, level of object preservation and diversity at the same time by mean opinion score (MOS). In each given question, the same daytime image is transformed to five different nighttime-style images by giving different random vectors and they are shown in the form of a single GIF image. The observers are expected to score from one to five (very low, relatively, low, medium, relatively high, and very high) for each question according to the aforementioned factors. To further assess the diversity objectively, similar to (Zhu et al. 2017b), we use the LPIPS metric (Zhang et al. 2018) (ImageNet-pretrained AlexNet (Krizhevsky 2014)) to measure the diversity in the BDD100k-val-day case. The LPIPS distance is computed between each day-to-night image pair corresponding to the same daytime image. Besides, we only focus on the largest vehicles in every single image because they are the most salient objects for vehicle detectors.

There are 74 observers involved in the user study, and the results are summarized in Table 6. The MOS comparison indicates that our work outperforms AugCGAN and MUNIT because other works would sometimes lead to unnatural, structure-inconsistent or even locally-darkened results and they are harmful for detector training. The LPIPS distance comparison also indicates that this work could bring more diverse looking for vehicles at nighttime in terms of different ambient light levels. It is worth mentioning that unimodal models are excluded in the diversity analysis because they could only generate a single nighttime-looking image given a daytime image.

Table 5: FCN-scores from FCN8s trained with SYNTHIA-seq-1-spring day-to-night-transformed by GANs (trained with SYNTHIA spring & night sequences other than seq-1), and tested with SYNTHIA-seq-1-night sequence excluded in GAN training.

GAN model	Per-class acc.	mIoU	fwIoU
CycleGAN	62.0	53.5	82.5
UNIT	65.8	55.8	85.6
AugGAN	67.4	58.6	87.8
CyCADA	67.7	60.1	88.8
AugCGAN	52.0	39.7	69.1
MUNIT	67.8	59.8	88.6
Multimodal AugGAN (Ours)	70.1	62.6	90.1

Table 6: LPIPS score for diversity analysis and MOS of realism, diversity and level of object preservation for day-to-night-transformed images in BDD100k-val-day by using the GAN models trained from BDD100k-seg-train and GTA-train.

GAN model	MOS	LPIPS
AugCGAN	1.57	0.38 ± 0.078
MUNIT	1.63	0.31 ± 0.108
Multimodal AugGAN (Ours)	3.31	0.47 ± 0.155

Detector Training Strategy

There are different ways of applying data augmentation for training a vehicle detector. In this work, we explore four kinds of training strategies including (1) only unimodal, (2) only multimodal, (3) unimodal & multimodal, (4) unimodal then multimodal day-to-night transformed data in the real-world dataset-BDD100k. As can be seen in Table 7, using only unimodal transformed results, the detector would learn to detect the vehicles in similar nighttime looking. Multimodal transformed results would result in diverse vehicle appearance. However, its diversity would sometimes lead to slightly-inferior results because learning the essence of vehicles' appearances from scratch given highly-varying appearance is very challenging. Mixing unimodal and multimodal data would lead to better results to some extent, but we found that using multimodal data for fine-tuning the detector pretrained with unimodal data would achieve the highest detection accuracy. In the fine-tuning stage, we adopt the learning rate 1e-4 which is the one used in the final stage of training both YOLO and Faster R-CNN. It is worth mentioning that, for a fair comparison, every detector trained with multimodal data also follows the official iterations in the original settings of PASCAL VOC results.

Training with Generated Night Images v.s Real Night Images

To label objects in nighttime images is expensive. Although using day-to-night-transformed images are beneficial, it is necessary to compare a detector trained with generated nighttime images and real ones. We train YOLO with images randomly sampled from BDD100k-train-night and all the BDD100k-val-day day-to-night-transformed images (5222

Table 7: Detection accuracy comparison (AP) using different training strategies - detectors trained with (BDD100k-val-day) images day-to-night-transformed by GANs (trained with BDD100k-seg-train) and tested on BDD100k-val-night.

Strategy	AugCGAN	MUNIT	Ours	Detector
uni	26.9	30.0	37.1	YOLO
uni	44.6	46.2	61.2	FRCN
multi	30.0	31.0	38.9	YOLO
multi	41.0	44.0	60.0	FRCN
uni+multi	31.1	31.7	37.9	YOLO
uni+multi	43.0	49.5	62.0	FRCN
uni then multi	32.0	36.7	40.4	YOLO
uni then multi	51.8	58.1	64.5	FRCN

images). However, the training in CNN is non-deterministic in the sense that the resulted AP would be slightly different every time. Therefore, we simply perform each training for five times and report the averaged results. Quantitative results show that the AP of YOLO vehicle detector trained with day-to-night images transformed by this work is close to the AP with 2k real nighttime images. However, AugCGAN and MUNIT achieve the performance reached by using only 0.5k and 1k images. Fairly speaking, this work is nearly four times and two times better than AugCGAN and MUNIT in terms of the AP achieved with different numbers of real nighttime images.

Table 8: Average precision (on BDD100k-val-night) comparison for night-time vehicle detectors (YOLO) trained with real nighttime images (BDD100k-val-night) and BDD100k-val-day day-to-night-transformed image generated by GANs learning from BDD100k-seg-train

Training data	AP
BDD100k-train-night random 0.5k	31.1
BDD100k-train-night random 1k	36.4
BDD100k-train-night random 1.5k	38.0
BDD100k-train-night random 2k	40.5
BDD100k-val-day day-to-night by AugCGAN	32.0
BDD100k-val-day day-to-night by MUNIT	36.7
BDD100k-val-day day-to-night by this work	40.4

Conclusion and Future work

In this work, we proposed Multimodal AugGAN, a mulimodal structure-consistent image-to-image translation network for realizing domain adaptation for vehicle detection. Our method quantitatively surpasses competing methods including unimodal and multimodal GANs for providing object-preserved, multimodal training data to achieve higher nighttime vehicle detection accuracy. The robustness of the vehicle detector is significantly improved because the vehicle detector would learn to adapt to different (1) ambient light levels, (2) brightness of vehicles' rear lamps, and (3) sharpness of vehicle's body at nighttime. The quantitative results demonstrate that multimodal translated images are beneficial in boosting the detection accuracy by fine-tuning a vehicle detector trained by the corresponding unimodal training data. Besides, the semantic segmentation experiments also indicate that our model could also provide performance gain on most of the classes in the nighttime scenario. This way, most daytime on-road datasets in public domain become valuable in the development of a segmentation model or an object detector at nighttime. In the future, we plan to dig into the latent spaces for explicitly manipulating semantic features corresponding to different parts of the objects in the progress of image translation. For example, once the appearance, style, brightness and lamp condition of buildings or vehicles could be directly and separately controlled, a CNN model would learn to better perceive the environment at nighttime.

References

Almahairi, A.; Rajeswar, S.; Sordoni, A.; Bachman, P.; and Courville, A. 2018. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *ICML*.

Braun, M.; Krebs, S.; Flohr, F.; and Gavrila, D. M. 2018. The eurocity persons dataset: A novel benchmark for object detection. *arXiv preprint arXiv:1805.07193*.

Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *IJCV*.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*.

Girshick, R. 2015. Fast r-cnn. In *ICCV*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A. A.; and Darrell, T. 2018. Cycada: Cycle-consistent adversarial domain adaptation. *ICML*.

Huang, S.-W.; Lin, C.-T.; Chen, S.-P.; Wu, Y.-Y.; Hsu, P.-H.; and Lai, S.-H. 2018a. Auggan: Cross domain adaptation with gan-based data augmentation. In *ECCV*.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018b. Multimodal unsupervised image-to-image translation. In *ECCV*.

Inoue, N.; Furuta, R.; Yamasaki, T.; and Aizawa, K. 2018. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*.

Kingma, D. P., and Ba, J. 2015. A method for stochastic optimization. In *ICLR*.

Krizhevsky, A. 2014. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*.

Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; Singh, M.; and Yang, M.-H. 2018. Diverse image-to-image translation via disentangled representations. In *ECCV*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *ECCV*.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *NIPS*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch. In *NIPS workshop*.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Redmon, J., and Farhadi, A. 2017. Yolo9000: Better, faster, stronger. In *CVPR*.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *CVPR*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for data: Ground truth from computer games. In *ECCV*.

Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; and Lopez, A. M. 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*.

Sivaraman, S., and Trivedi, M. M. 2010. A general active-learning framework for on-road vehicle recognition and tracking. *IEEE Transactions on Intelligent Transportation Systems* 11(2):267–276.

Yang, L.; Luo, P.; Change Loy, C.; and Tang, X. 2015. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*.

Yu, F.; Xian, W.; Chen, Y.; Liu, F.; Liao, M.; Madhavan, V.; and Darrell, T. 2018. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Zhou, Y.; Liu, L.; Shao, L.; and Mellor, M. 2016. Dave: a unified framework for fast vehicle detection and annotation. In *ECCV*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *ICCV*.

Zhu, J.-Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A. A.; Wang, O.; and Shechtman, E. 2017b. Toward multimodal image-to-image translation. In *NIPS*.