# CyEDA: CYCLE-OBJECT EDGE CONSISTENCY DOMAIN ADAPTATION

*Jing Chong Beh*[1], *Kam Woh Ng*[2], *Jie Long Kew*[1], *Che-Tsung Lin*[3]
*Chee Seng Chan*[1*], *Shang-Hong Lai*[4,5], *Christopher Zach*[3]

[1] CISiP, Faculty of Comp. Sci. and Info. Tech., Universiti Malaya, Kuala Lumpur, Malaysia
[2] CVSSP, University of Surrey, Guildford, U.K.
[3] Dept. of Electrical Engineering, Chalmers University of Technology, Sweden
[4] Microsoft AI R&D Center, Taiwan
[5] Dept. of Computer Science, National Tsing Hua University, Taiwan

## ABSTRACT

A difficulty of global-level translation is to preserve instance-level details in an image. Although some instance level translation methods can retain the details, most of them require either pre-trained object detection/segmentation network or annotation labels. In this work, we propose a novel method namely *CyEDA* to perform global level domain adaptation that can preserve image contents without any pre-trained networks integration or annotation labels. Specifically, we introduce blending masks and cycle-object edge consistency loss which exploit the preservation of image objects. We show that our approach can outperform other SOTAs in terms of image quality and FID score in both BDD100K and GTA datasets. The code and pre-trained models are publicly available at `https://github.com/bjc1999/CyEDA`.

***Index Terms*—** domain adaptation, image-to-image translation

## 1. INTRODUCTION

Object detection on images under sub-optimal lightning remains a very challenging task despite the rapid advancement of object detection algorithms. This is because images captured in real life are often under the influence of various lighting issues such as weak lightning which complicated the task [1,2]. Image translation-based domain adaptation (DA) is one of the approaches widely studied for object detection in low-light condition. It normally translates normal light images to low-light ones and trains an object detection model with the translated low light images as done by [3–8]. While these methods are self-supervised, they suffer different issues during the translation process.

In this paper, we propose a new approach to achieve instance-level domain adaptation results. Particularly, we introduce blending mask when translating the images from one domain to another domain to ensure that every object is preserved during the translation process without any need for a pre-trained detection network. Next, we also introduce *cycle-object edge consistency loss* to replace the conventional cycle-consistency L1 loss to enforce the model to preserve only instance-level details in terms of content, instead of pixel-level details to prevent information hiding in the generated images. Without any need for annotation labels and pre-trained networks, our model can produce high fidelity results as illustrated in Fig 1. We compare our results both quantitatively via FID [9] score, and qualitatively to show that our approach can outperform current SOTAs. We also successfully perform unsupervised domain adaptation on an object detection model by means of data augmentation to show that the mAP performance of the detection model is boosted with the help of generated images from our model.


**Fig. 1**: Our sample results: day → night on BDD100k [10].

## 2. RELATED WORK

**Image-to-image Translation.** The goal of image-to-image translation (I2I) is to translate an image from its original domain (e.g., daytime) to a target domain (e.g., night-time). Pix2Pix [11] initiated the trend on GAN-based I2I which

*Corresponding author - cs.chan@um.edu.my

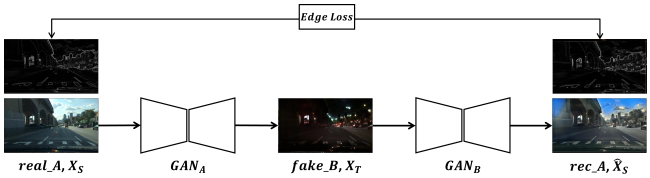
can produce visually impressive results in the target domain given a paired dataset. CycleGAN [3] further improved I2I by introducing cycle-consistency loss to remove the dependency of paired dataset. DRIT [4] proposed to disentangle content and style representation of images to achieve diverse I2I. All these methods have made great progress in improving translation results, but they all suffered from translating content-rich images such as driving images with cars and pedestrians across complex domains (e.g., day-to-night).

AugGAN [5] and multimodal AugGAN [6] overcame the structure-consistency by using segmentation annotation to avoid content distortion. Although both methods [5, 6] can preserve image-structure details very well, they required segmentation annotations which are expensive to obtain in most cases. To alleviate this issue, INIT [7] proposed to make use of bounding box annotations to perform instance-level I2I by implementing code bank structure. However, INIT [7] disposed the instance level module after training and thus it lost the instance-level information during inference. DUNIT [8] can preserve instance-level details by integrating an object detection network pre-trained on source domain as a constraint, such that the generated images with annotated instances are still detectable by the detection network. However, the detection network is only pre-trained on the source domain. It is thus arguable whether it can make valid detection on translated images. In this work, our model can preserve instance-level details by integrating blending masks and cycle-object edge consistency loss without relying on any pre-trained networks and annotations.

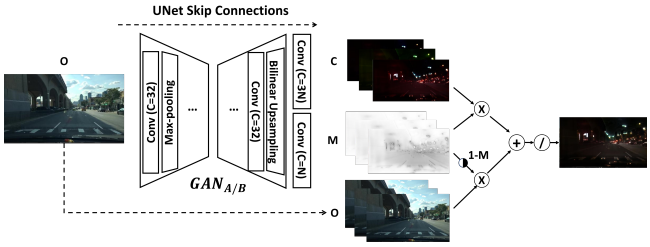
**Domain Adaptation.** Most of the methods discussed aforementioned had been applied to unsupervised domain adaptation for object detection by the mean of data augmentation. This is done by translating a fully labeled dataset to the desired target domain where annotations are not available. Following this, these translated images are used to train a detection model with annotations from the source dataset to improve the detection accuracy without the need of creating a new fully annotated dataset in the target domain. Due to the nature of this problem, the quality of objects after translation is very important as it affects the performance of the detection network. As discussed, methods that are struggling in preserving instance details could not adapt the object detection network to a new domain well. Other methods that can retain instance-level details can perform domain adaptation well but they require either expensive annotation labels or pre-trained networks. Most domain adaptations are experimented on [12–14]. In our work, we used YOLOv5 [14] to assess the quality of the translated images.

## 3. METHODOLOGY

Generally, the goal of unsupervised I2I is to translate an image from a source domain to a target domain with an unpaired




(a) Translation from day-time (night-time) to night-time (day-time).




(b) Mask UNet. The last layer of Mask Unet outputs color changes $C$, and blending masks $M$ which is the degree of the color changes. The image is translated by blending $C$ with the original image $O$.

**Fig. 2**: Our proposed model architecture.

dataset. Given two sets of images $X$ and $Y$ that are unpaired from two different domains $S$ and $T$ , where $X_S \in \mathbb{R}^{H\times W\times 3}$ and $Y_T \in \mathbb{R}^{H\times W\times 3}$, I2I learns a mapping to translate $X_S$ to $X_T$ where $X_T$ should be similar to $X_S$ in terms of contents, and similar to $Y_T$ in terms of style. The same is also applied for $Y_T$ to $Y_S$ in terms of translation. Our work also has a similar structure but with our proposed blending masks and edge loss. Our schematic model is shown Fig. 2.

### 3.1. Mask UNet

Our GAN architecture uses UNet [18] as the backbone because the skip connection design enables precise segmentation as done in their work. To retain instance-level detail during the translation process, we argue that color changes to every pixel of the original image should not be sophisticated. To this end, we modify the last layer of UNet into two separate blocks: (i) a convolution layer that yields $3N$ channels of output followed by *tanh* activation layer, and (ii) a convolution layer that yields $N$ channels of output followed by *sigmoid* activation layer, where $N$ represents the number of masks, inspired by [19]. The block (i) outputs the predicted color changes, $C \in [-1, 1]^{N\times H\times W\times 3}$, to the original image and the block (ii) outputs the predicted degree of color changes, $C \in [0, 1]^{N\times H\times W}$, to the original image. The final image is the normalized blending between element-wise multiplication of $C$ with $M$ and original image $O$ with inverted mask $(1 - M)$. Note that the generation of $M$ neither requires segmentation nor any other form of mask label during the learning. We name this architecture Mask UNet (see Fig. 2b). The process can be represented as:

$$X_T = ((C \otimes M) \oplus (O \otimes (1 - M))) \oslash N. \qquad (1)$$


**Fig. 3**: Quality Evaluation. Left to right: Original daytime image, NICE-GAN [15], Multimodal AugGAN [6], MUNIT [16] and our work (CyEDA). First row: GTA [17] dataset. Second row: BDD100k [10] dataset.

Eq. 1 is essentially computing a translated image by summing the weighted mask of color changes and the inverted weighted mask of the original image. We think that this is easier for the model to translate the original image with all the instances preserved. This is because the translated image is not generated from its hidden embedding. Rather, it is translated from the original image directly to retain the instance level details.

**Comparison with DUNIT [8].** DUNIT integrates a detection subnet to encourage similar detection results. In contrast, our work guides the translation by blending model outputs with the original image to retain instance-level detail. This approach enjoys some benefits over DUNIT: **(i)** DUNIT only encourages the bounding box predicted by the detection subnet from $X_S$ to be similar to the one from $X_T$ using $L1$ loss, but not the content (i.e. it only ensures the detection can still locate the objects but neglects the classification of the objects). **(ii)** The integrated detection subnet is only pre-trained on the source domain which makes the validity of its prediction results on the target domain $X_T$ arguable. **(iii)** Since the architecture requires a pre-trained detection model, the whole model is thus not end-to-end trainable.

### 3.2. CyEDA: Cycle-Object Edge Consistency Loss

In unsupervised I2I translation, a cycle-consistency loss is used to train the model since we do not have any ground truth, e.g., the case in supervised I2I translation. However, [20] shows that CycleGAN learns to hide information in the network to satisfy the cycle consistency requirement. When $L1$ loss is being used as cycle consistency loss, GAN tries to hide/preserve every single details (e.g., leaves) when translating the image. This causes unrealistic translation results.

*We argue that the cycle consistency loss should only enforce the preservation of objects in the images instead of every details.* To this end, we propose cycle-object edge consistency loss to replace $L1$ loss. Instead of computing the $L1$ loss between $X_S$ and $\hat{X}_S$, we first extract the edge information $E_{X_S}$ and $E_{\hat{X}_S}$ from $X_S$ and $\hat{X}_S$ respectively by any differentiable edge detection (Sobel Filter [21] is used in this work, but not limited to). We then compute $L1$ loss between $E_{X_S}$ and $E_{\hat{X}_S}$.

| GAN model | annotation? | GTA | BDD100k |
|---|---|---|---|
| MUNIT [16] | No | 1.066 | 2.461 |
| NICE-GAN [15] | No | 2.466 | 1.913 |
| AugGAN [5] | Yes | 0.825 | 0.332 |
| Multimodal AugGAN [6] | Yes | 1.023 | 0.496 |
| CyEDA (Our work) | No | **0.737** | **0.297** |

**Table 1**: FID score comparison of our approach with SOTAs on GTA (val-night and val-day day-to-night) and BDD100k (det-val-night and det-val-day day-to-night) datasets.

In this way, we only enforce the overall content (e.g., shape) of $X_S$ and $\hat{X}_S$ to be similar, instead of every pixel detail. Our proposed edge loss is formulated as:

$$L_{Edge} = L_1(E_{X_S}, E_{\hat{X}_S}). \tag{2}$$

We also exploit the same domain adversarial loss terms as CycleGAN [3], $L^A_{adv}(G_A, D_A)$ and $L^B_{adv}(G_B, D_B)$ with corresponding domain discriminators $D_A$ and $D_B$. Overall, we minimize the following loss function:

$$L = L^A_{adv} + L^B_{adv} + L_{Edge}. \tag{3}$$

## 4. EXPERIMENTS

### 4.1. Unsupervised Image-to-image Translation

**Dataset**. Datasets used to verify the effectiveness of our I2I model architecture are BDD100k [10] and GTA [17]. Due to GPU limitation, we resize the images in those datasets to $256 \times 455$ for faster training and experiments.

**Benchmark**. We compare our results with several SOTA unpaired I2I methods as follows: (i) MUNIT [16], which disentangles the image feature into content (domain invariant) and style (domain-specific) and then switches content and style features between two domains to generate the translated images. (ii) NICE-GAN [15], which reuses the discriminator model for image encoding. (iii) AugGAN [5] and multimodal AugGAN [6], which integrate segmentation subnets to ensure


**Fig. 4**: Qualitative: Ablation study on different settings.

structure-consistency in translated image. However, INIT [7] and DUNIT [8] do not provide complete source code; thus, they are not compared here.

**Results**. We compare our results with other SOTAs in terms of image quality in Fig. 3. It shows that our work outperforms other SOTAs in terms of fidelity. SOTAs, such as NICE-GAN [15] and MUNIT [16], that did not use any extra detection or segmentation subnet to exploit the instance consistency suffer from preserving the image-objects in image-translation. While multimodal AugGAN [6] successfully preserves the instances after translation, it fails to produce high fidelity translation result even though they leveraged segmentation labels in their work. Our result is better not only in terms of salient objects such as vehicles, but also the backgrounds such as buildings. To objectively evaluate the quality of generated images, we also compare the FID score [9] of whole images among our proposed model and other SOTAs. As shown in Table 1, our model outperforms other SOTAs, yet does not require any pre-trained detection subnet and annotation labels. Thanks to the blending mask, our model only has to learn how to fine-tune the color of the original image instead of redrawing the whole image in the target domain which makes it easier to produce a better quality image.

**Ablation Study**. To understand the effects of the proposed blending mask and edge loss, we train models with different combinations of both proposed modules. With CycleGAN [3] as baseline (Fig. 4(a)), it is the model without both proposed blending mask and edge loss. The generated image is dark and unrealistic and the instance details are blurry. In Fig. 4(b), when we use edge loss without the masking module (like a typical UNet), the loss alone can only produce a gray-scale version of the original daytime image. It can be attributed to the edge loss which cannot guide the model to translate across domains because the model only exploits on the edge (the shape) information instead of the pixel details. In Fig. 4(c), we can see that color contrast is well preserved compared to 4(a) and 4(b) with the help of masking. However, some unnecessary details like leaves were kept inside the image which look unrealistic for a night-time image. We argue that this is because the model has to translate all the details

| Experiment Settings | FID |
|---|---|
| No mask + L1 (CycleGAN + UNet) | 0.551 |
| No mask + Edge | 0.579 |
| Mask + L1 | 0.389 |
| **Mask + Edge (CyEDA)** | **0.297** |

**Table 2**: Quantitative: Ablation study on different settings.

| Training Dataset (BDD100k) | mAP (whole) | AP (car) |
|---|---|---|
| det-train-night | 0.444 | 0.627 |
| + det-val-day day-to-night | **0.465** | **0.644** |

**Table 3**: Object detection domain adaptation at night-time. Testing on BDD100k-det-val-night.

back to the original daytime image (i.e., $X_S$ to $X_T$ to $\hat{X}_S$) due to L1 cycle consistency loss, which worsens the learning. In Fig. 4(d), our CyEDA model can produce high-quality images with instance details retained. The translated images is a weighted summation of the original image and the color changes. We believe this has simplified the translation process where instance details are directly kept from the original images.

### 4.2. Domain Adaptation

We make use of the generated night-time images from our proposed model to train a YOLOv5 [14] object detection model to examine the effect of including the generated images in training object detector in the target domain. We experimented in two settings: training the small YOLOv5 model (YOLOv5s) using BDD100k real night-time training images (around 2000 images) with and without the translated images (around 1000 BDD100k-det-val-day-to-night images) generated by our proposed model. The YOLOv5s model is pre-trained on COCO128 [22] dataset. The result of the experiment is presented in Table 3, and it shows that include the generated images as training images does help to boost the performance of YOLOv5s in detecting objects in the target domain (night-time) as the model has more data to learn from.

## 5. CONCLUSION

This paper introduces an approach to retain instance-level detail when translating images to a target domain by generating blending masks from UNet and performing color fine-tuning on original images according to the masks. We also proposed *cycle-object edge consistency loss* to remove information hiding in generated images, which gives extra capacity to perform more realistic image-translation. Our model is able to produce SOTA results in terms of FID score and image quality. In the future, we plan to investigate the use of masking in skip connections of UNet so that the effects of masking can be applied to different levels of features.

## 6. REFERENCES


[1] Chongyi Li, Chunle Guo, Ling-Hao Han, Jun Jiang, Ming-Ming Cheng, Jinwei Gu, and Chen Change Loy, "Low-light image and video enhancement using deep learning: A survey," *TPAMI*, pp. 1–1, 2021.

[2] Yuen Peng Loh and Chee Seng Chan, "Getting to know low-light images with the exclusively dark dataset," *Computer Vision and Image Understanding*, vol. 178, pp. 30–42, 2019.

[3] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[4] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang, "Diverse image-to-image translation via disentangled representations," in *ECCV*, September 2018.

[5] Sheng-Wei Huang, Che-Tsung Lin, Shu-Ping Chen, Yen-Yi Wu, Po-Hao Hsu, and Shang-Hong Lai, "Auggan: Cross domain adaptation with gan-based data augmentation," in *ECCV*, September 2018.

[6] Che-Tsung Lin, Yen-Yi Wu, Po-Hao Hsu, and Shang-Hong Lai, "Multimodal structure-consistent image-to-image translation," *AAAI*, vol. 34, no. 07, pp. 11490–11498, Apr. 2020.

[7] Zhiqiang Shen, Mingyang Huang, Jianping Shi, Xiangyang Xue, and Thomas Huang, "Towards instance-level image-to-image translation," in *CVPR*, 2019.

[8] Deblina Bhattacharjee, Seungryong Kim, Guillaume Vizier, and Mathieu Salzmann, "Dunit: Detection-based unsupervised image-to-image translation," in *CVPR*, 2020, pp. 4786–4795.

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *NIPS*, vol. 30, 2017.

[10] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *CVPR*, 2020, pp. 2633–2642.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, vol. 28.

[13] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[14] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Jiacong Fang, imyhxy, Kalen Michael, Lorna, Abhiram V, Diego Montes, Jebastin Nadar, Laughing, tkianai, yxNONG, Piotr Skalski, Zhiqiang Wang, Adam Hogan, Cristi Fati, Lorenzo Mammana, AlexWang1900, Deep Patel, Ding Yiwei, Felix You, Jan Hajek, Laurentiu Diaconu, and Mai Thanh Minh, "ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference," Feb. 2022.

[15] Runfa Chen, Wenbing Huang, Binghui Huang, Fuchun Sun, and Bin Fang, "Reusing discriminators for encoding: Towards unsupervised image-to-image translation," in *CVPR*, June 2020.

[16] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz, "Multimodal unsupervised image-to-image translation," in *ECCV*, September 2018.

[17] Stephan R. Richter, Zeeshan Hayder, and Vladlen Koltun, "Playing for benchmarks," in *ICCV*, 2017, pp. 2232–2241.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[19] Oran Gafni, Lior Wolf, and Yaniv Taigman, "Live face de-identification in video," in *ICCV*, 2019, pp. 9377–9386.

[20] Casey Chu, Andrey Zhmoginov, and Mark Sandler, "Cyclegan, a master of steganography," *arXiv preprint arXiv:1712.02950*, 2017.

[21] Irwin Sobel, "An isotropic 3x3 image gradient operator," *Presentation at Stanford A.I. Project 1968*, 02 2014.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.