GAN Based Image-to-Image Translation Model for Nighttime Road Scene Dataset

Rebeka Sultana[1], Yuki Hikosaka[1], Gosuke Ohashi[1]

sultana.rebeka.15@shizuoka.ac.jp
[1]Shizuoka Univ., 3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka 432-8561, Japan
Keywords: Image-to-image translation, GAN model, daytime road scene, nighttime road scene, ADAS.

ABSTRACT

A large-scale dataset boosts the performance of deep learning models. However, in-vehicle camera image datasets contain many daytime and few nighttime scenes. Therefore, this study proposes an image-to-image translation model to convert daytime road scenes to nighttime road scenes by preserving visual appearance and contents to increase nighttime dataset.

1 Introduction

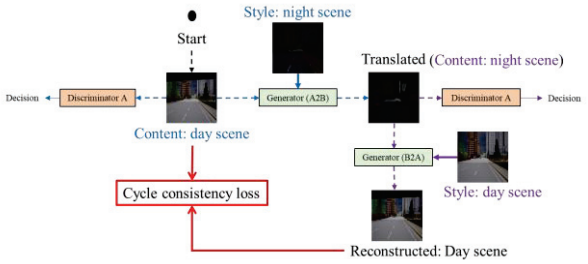
In recent years, driving automation technology has been actively researched. Advanced Driver Assistance System (ADAS) is a low-level driving automation technology consisting of several functions such as adaptive cruise control, night vision, and automatic emergency braking. These functions are executed to reduce driver's stress and increase safe driving. ADAS monitors limited safety rules using sensors such as in-vehicle cameras and LiDAR. Due to the cost-effectiveness of in-vehicle cameras, computer vision is extensively applied to in-vehicle images to complete several tasks, such as traffic-related object detection on the road [1].

In computer vision, deep learning models achieve high accuracy in ImageNet Large Scale Visual Recognition Challenges (ILSVRC) [2]. However, the performance of the deep learning model mainly depends on the scale of the dataset used for training. For example, high accuracy is obtained when models are pre-trained using large-scale datasets in several tasks such as object detection and segmentation [3]. However, most of the in-vehicle camera image datasets currently available contain only daytime scenes and a few nighttime scenes [4]. Therefore, the object detection rate is high when daytime datasets are used. Moreover, the daytime scene contains objects with good visibility due to high illuminance during the day. However, increasing the dataset for nighttime scenes the same as the daytime dataset is a labor-intensive task in object detection [4]. Therefore, we investigate utilizing the daytime dataset for the nighttime task by converting the daytime domain to the nighttime domain.

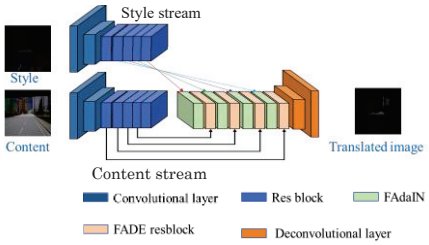
Recently, image domain conversion is completed by General Adversarial Network (GAN) based image-to-image translation models [5]. One of the most popular examples of image domain conversion is to translate a horse image to a zebra image or vice-versa by the CycleGAN model [6]. Therefore, it is possible to convert the daytime road scenes to nighttime by using image-to-image translation model. The merit of domain conversion is to use the same annotations for translated images. However, the content information can be lost in translated images by models such as CycleGAN [6]. Therefore, visual appearance and contents are important in translated images for object detection tasks. Image-to-image translation models apply various concepts such as cycle consistency and content-style separation [6] – [9]. In CycleGAN [6] model, domain A (real image) is first translated to domain B (translated image). Then domain B (translated image) is translated back to domain A (reconstructed image). Therefore, a cycle is completed. CycleGAN [6] learns by imposing that domain A (real image) and domain A (reconstructed image) must be the same. This process is known as cycle consistency. Cycle consistency validates domain conversion by avoiding unnecessary changes in the translated image. On the other hand, the content−style separation concept is applied in TSIT [7] model without using cycle consistency. The content is the shape of object in a scene and the texture or color in the shape are the style of the object. The concept is to use the content of domain A (real image) and the style of domain B (real image) to translate images from domain A to domain B. Therefore, TSIT [7] model takes real images from both domains for translation. The content−style separation improves the accuracy of image conversion by reducing artifacts in translated images. We assume that the content information can be improved by employing both concepts in translated images. Therefore, this study proposes an image-to-image translation model to improve the visual appearance and content information of a translated image by incorporating content and style images in CycleGAN [6] model to convert daytime road scenes to nighttime road scenes.

The contributions of our work are summarized in three points.

1. We propose an image-to-image translation model using not only cycle consistency but also content-style separation concept to convert the daytime road scene dataset to the nighttime road scene dataset. The converted nighttime dataset shows high fidelity to the annotations of daytime dataset.
2. A driving simulator makes it possible to create daytime and nighttime road scenes of the same situation by setting options except for weather conditions and time. The road scenes created by the driving simulator are used to investigate model performance qualitatively and quantitatively.
3. Experimental results show that proposed model outperforms cycle consistency model (CycleGAN [6]), content-style separation model (TSIT [7]), and cycle consistency and content-style separation models



(a) Cycle consistency



(b) Content-style separation in generator (A2B)

Fig. 1 Proposed model

(DRIT [8], MUNIT [9]).

The rest of the paper is organized as follows. In section 2, the proposed model is described. Section 3 explains all experiments. Finally, the conclusion is presented in section 4.

2 Proposed method

The schematic diagram of proposed model is shown in fig. 1.

2.1 Cycle consistency

The model overview is shown in fig. 1 (a). The GAN based image-to-image translation model consists of two parts: generator and discriminator. The generator generates translated image by taking a real image while getting feedback from the discriminator, which classifies the real images and translated images. The process of the proposed model starts from the start point, as shown in fig. 1 (a). The main training steps of the proposed model are as follows:

1. Generator (A2B) takes daytime road scene as content image and nighttime road scene as style image to generate translated image.
2. Discriminator A classifies the content image as real and the translated image as fake.
3. Generator (B2A) takes translated nighttime road scene as content image and daytime road scene as style image to generate reconstructed image.
4. Losses are very important to learn GAN based model. The cycle consistency loss is used in our model in addition to adversarial loss, identity loss, perceptual loss, and feature matching loss during training [6], [7]. The cycle consistency loss is defined by the following equation.

$$Loss_{cyc}(G,F) = \mathbb{E}_{a\sim p_{data}(a)}\|F(G(a))-a\| + \mathbb{E}_{b\sim p_{data}(b)}\|G(F(b))-b\|$$

$p_{data}(x)$ stands for probability of real x image. Real images are a and b. F and G represent generator B2A and A2B, respectively.

To generate translated images from the daytime scene, only learned Generator (A2B) is used.

2.2 Content-style separation in generator

We adopted the main generator from the TSIT [7] model by excluding the random noise. The generator takes two images as content and style individually. The convolutional layer processes content and style images following res blocks. The following equation defines the res block.

$$y = F(x) + x$$

Here, x is the input feature and F consists of convolution layers. The individual style features from style stream are processed by Feature Adaptive Instance Normalization (FAdaIN) layer [7]. FAdaIN layer is expressed as follows [7].

$$FAdaIN(f_i^s, z_i) = \sigma(f_i^s)\left(\frac{x-\mu(z_i)}{\sigma(z_i)}\right) + \mu(f_i^s)$$

f_i^s, σ, μ, z are style features, mean, standard deviation, and content features, respectively. Content features from the content stream are processed by Feature adaptive denormalization (FADE) resblock layer, which is an element-wise normalization-based operation [7]. Features from different layers in the content stream are used in a skip-connection manner which helps to preserve the content information in the translated image. Finally, the deconvolution layer is used to obtain the target size output image.

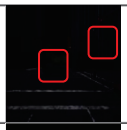
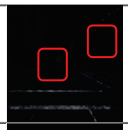
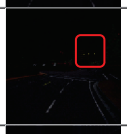
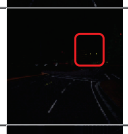
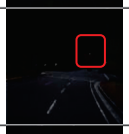
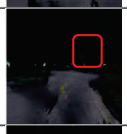
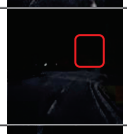
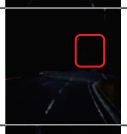
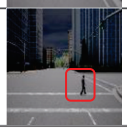
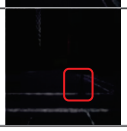
3 Experiment

3.1 Datasets

Two datasets are used in the experiment: the simulator dataset and the real dataset. An open-source simulator named Carla [10], developed by Epic Games, is used to collect the simulator dataset. By setting MAP, objects, and scenes on the Carla, the dataset can be collected for daytime and nighttime scenes of the same area. In real-time road scenes, it is impossible to obtain daytime and nighttime scenes of the same area. Therefore, a simulator dataset is used to compare translated images and ground truths qualitatively and quantitively. Moreover, the real image dataset INIT [5] is used to compare translated images among different models. INIT dataset [5] is proposed for the image conversion task. The images in the dataset were collected in Tokyo, Japan. INIT dataset contains bounding box annotations of traffic-related objects such as cars, persons, and traffic signs.

3.2 Experimental conditions

In the simulator train set, the number of Carla daytime and nighttime scenes are 17,687 images and 17,687 images, respectively. In the simulator test set, the number of Carla daytime and nighttime scenes are 5,895 images and 5,895 images, respectively. In the real image train set, the number of INIT daytime and nighttime scenes are 10,000 images and 10,000 images, respectively. In the real image test set, the number of INIT daytime and nighttime scenes are 33,370 images and 2000 images, respectively. In the experiment, the batch size is set to 1, the learning rate is 0.0002, the

Example	Daytime	Nighttime	CycleGAN [6]	TSIT [7]	DRIT [8]	MNUIT [9]	Proposed
(a)							
(b)							
(c)							

(a) Qualitative results

Model	Epoch	Original (day)	Original (night)	Translated (night)	FID (↓)
-	-	✓	✓	×	102.8
CycleGAN [6]	20	×	✓	✓	65.8
TSIT [7]	10	×	✓	✓	84.6
DRIT [8]	20	×	✓	✓	176.7
MUNIT [9]	20	×	✓	✓	147.7
Proposed	20	×	✓	✓	**62.6**

(b) Quantitative result

Fig. 2 Experimental results (Carla dataset [11])

optimization function is ADAM, and the loss function is L1 during training. The images are resized to 256×256 pixels. The model is trained for 20 epochs. Network is implemented on PyTorch framework using python language. The CPU configuration of the computer environment is Intel(R) Core(TM) i9-9900. The GPU configuration is NVIDIA GeForce GTX 2080Ti 11GB. The memory is 32GB.

3.3 Evaluation metric

Fréchet Inception Distance (FID) is used to evaluate the performance of model [11] with unpair images. FID accesses the quality of images generated by generative models such as GAN. FID can be expressed as follows [10]:

$$\mathrm{FID} = \left\|\mu_x - \mu_y\right\|^2 + \mathrm{Tr}\left(\Sigma_x + \Sigma_y - 2\Sigma_x\Sigma_y\right)$$

Here, x and y are feature vectors from input images. μ and Σ are the mean and covariance of the feature vector, respectively. The number of feature vectors is 2048 for each input image. A pre-trained inception model is used to obtain feature vectors from real and translated images of the same domain.

3.4 Comparison models

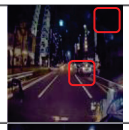
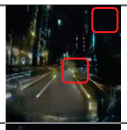
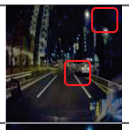
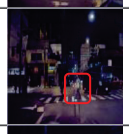
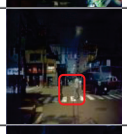
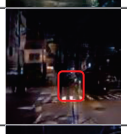
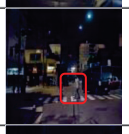
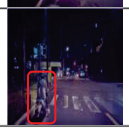
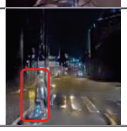
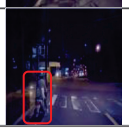
The proposed model is compared with other image-to-image translation models: CycleGAN [6], TSIT [7], DRIT [8], and MUNIT [9]. CycleGAN [6] employs cycle consistency concept. On the other hand, TSIT [7] uses the content-style separation concept. DRIT [8] and MUNIT [9] use cycle consistency and content-style separation concepts.

3.5 Experimental results

The experimental results are shown in Fig. 2 and Fig. 3. The red box indicates about the objects in the translated image to be discussed. The best score is shown in bold. (↓) indicates the lower the score is, the better the model performance is.

Qualitative and quantitative results of the simulator dataset are shown in Fig. 2 (a) and Fig. 2 (b), respectively. In example (a), translated car by the proposed method shows a loss of content information than only translated image by TSIT [7] compared with the ground truth image. However, the content information of the building in the translated image by the proposed model is better than translated image by TSIT [7]. In example (b), the traffic light and building information in the translated image are better than other models. Similar behavior of the proposed model is seen in example (c) according to the content information of the pedestrian. In examples (a)-(c), translated images by the proposed model are closer to ground truth visually. Quantitative results in Fig. 2 (b) show that the proposed model achieves the best FID score.

Qualitative and quantitative results of the real image dataset are shown in Fig. 3 (a) and Fig. 3 (b), respectively. The ground truths are not available. Therefore, the translated images will be discussed only. The content information in translated images by TSIT [7] is better than the other models. However, the domain conversion is not good by TSIT [7] model, as shown in translated images. In example (a), the sky area generated by TSIT [7] model indicates that TSIT [7] model fails to convert the domain as expected. However, the proposed model is able to convert domains successfully while preserving content information as much as possible. In examples (a)-(c), the objects such as car, pedestrian, and rider are translated properly by the proposed model in terms of domain conversion and preserving content information than CycleGAN [6], DRIT [8], and MUNIT [9]. However, domain

Example	Daytime	Nighttime	CycleGAN [6]	TSIT [7]	DRIT [8]	MNUIT [9]	Proposed
(a)		-					
(b)		-					
(c)		-					

(a) Qualitative results

Model	Epoch	Original (day)	Original (night)	Translated (night)	FID (↓)
-	-	✓	✓	×	114.7
CycleGAN [6]	20	×	✓	✓	56.7
TSIT [7]	20	×	✓	✓	72.9
DRIT [8]	20	×	✓	✓	96.4
MUNIT [9]	20	×	✓	✓	58.3
Proposed	20	×	✓	✓	**56.5**

(b) Quantitative result

Fig. 3 Experimental results (INIT dataset [5])

conversion by the proposed model is better than TSIT [7] model. Quantitative results in Fig. 3 (b) show that the proposed model outperforms other models according to the FID score.

4 Conclusions

This study proposes an image-to-image translation model to convert daytime road scenes to nighttime road scenes to improve the visual appearance and content information. The effectiveness of the proposed model is confirmed by the simulator dataset (Carla) [10] and the real image dataset (INIT) [5]. Generated images by the proposed model are compared with generated images by CycleGAN [6], TSIT [7], DRIT [8], and MUNIT [9]. The proposed model shows visually better results than other models according to cycle consistency and content-style separation. Moreover, quantitative results show that the proposed model achieves the best score in terms of the FID score.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP20K11828.

References

[1] N. Kosaka and G. Ohashi, “Vision based nighttime vehicle detection using CenSurE and SVM,” IEEE Transactions on Intelligent Transportation Systems, Vol. 16, No. 5, pp. 2599-2608, 2015.

[2] J. Deng, W. Dong, R. Socher, L. Li, K. Li and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” IEEE conference on computer vision and pattern recognition, pp. 248-255, 2009.

[3] C. Sun, A. Shrivastava, S. Singh and A. Gupta, “Revisiting unreasonable effectiveness of data in deep learning era,” IEEE international conference on computer vision, pp. 843-852), 2017.

[4] H. Lee, M. Ra, and W. Kim, "Nighttime data augmentation using GAN for improving blind-spot detection," IEEE Access 8, pp 48049-48059, 2020.

[5] Z. Shen, M. Huang, J. Shi, and X. Xue, and T. S Huang, “Towards instance-level image-to-image translation,” IEEE/CVF conference on computer vision and pattern recognition, pp. 3683-3692, 2019.

[6] J. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks,” IEEE international conference on computer vision, pp. 2223-2232, 2017.

[7] L. Jiang, C. Zhang, M. Huang, C. Liu, J. Shi and C. C. Loy, “TSIT: A simple and versatile framework for image-to-image translation,” European Conference on Computer Vision, pp. 206-222, 2020.

[8] H. Lee, H. Tseng and J. Huang, M. Singh, M. Yang, “Diverse image-to-image translation via disentangled representations,” European conference on computer vision, pp. 35-51, 2018.

[9] X. Huang, M. Liu, S. Belongie and J. Kautz, “Multimodal unsupervised image-to-image translation,” European conference on computer vision, pp. 172-189, 2018.

[10] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez and Vladlen Koltun. “CARLA: An open urban driving simulator,” In Conference on robot learning, PMLR, pp. 1-16, 2017.

[11] M. Heusel, H. Ramsauer, T. Unterthiner and B. Nessler, “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium,” Advances in neural information processing systems 30, 2017.