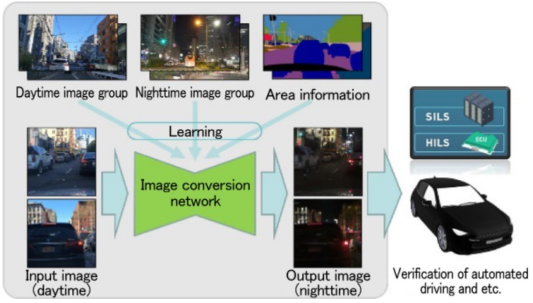
Image Conversion Technology Utilizing Deep Learning & Semantic Segmentation



TOSHIAKI NISHIMORI[*1] **KO HIRADE**[*2]

YOSHINAO TAKAKUWA[*3] **HIROKAZU SHIMIZU**[*4]

KENTA NAKAO[*5]

Driving tests using actual vehicles have been conducted in order to verify automated vehicles. However, there are countless combinations of natural phenomena and traffic flows, and it is required to reduce the time and cost for collecting verification data and improve the efficiency. Therefore, Mitsubishi Heavy Industries, Ltd. (MHI) has developed Artificial Intelligence (AI) algorithm (related patents application) to convert a real image taken in a certain environment into an image of a different environment using deep learning for camera images, which play an important role in object recognition, in collaboration with the University of Tokyo. In this algorithm, semantic segmentation technology is combined with the conventional image conversion technology to improve the conversion accuracy of details. As a result, a large amount of data required for verification of automated vehicles can be generated in a short time, and the time and cost for collecting verification data can be reduced.

1. Introduction

Automated vehicles use various sensors such as Radio Detection and Ranging (Radar) and Light Detection and Ranging (LiDAR) to recognize the environment around the vehicle. Among them, the image sensor (camera) plays an important role in object recognition. In order to verify the camera image recognition technology, driving tests using actual vehicles are used. However, there are countless combinations of natural phenomena and traffic flows, and it is required to reduce the time and cost for collecting verification data and improve the efficiency.

As such, in order to utilize the latest AI technology and contribute to mobility development, MHI has conducted a joint research with the the Kamijo Laboratory of the University of Tokyo Interfaculty Initiative in Information Studies. As a result of the research, we developed a technology to convert a real image taken in a certain environment into an image of a different natural environment such as nighttime and rainy weather using deep learning. As a result, a large amount of data required for verification of automated vehicles can be generated in a short time, and the time and cost for collecting verification data can be reduced.

This report covers the developed image conversion technology and test results.

2. Image conversion

2.1 Generative adversarial network

Generative adversarial network (GAN)[(1)] is an artificial intelligence algorithm that generates non-existent pseudo data by learning the features of input data through competition between a generator network (Generator) and a discriminator network (Discriminator). Due to the flexibility of its architecture, GAN has been studied in various fields, and many methods have been proposed

[*1] Principal Engineer, Electricity & Control Systems Engineering Department, Infrastructure Facilities Business Division, Mitsubishi Heavy Industries Machinery Systems, Ltd.

[*2] Electricity & Control Systems Engineering Department, Infrastructure Facilities Business Division, Mitsubishi Heavy Industries Machinery Systems, Ltd.

[*3] General Manager, Project Promotion Department, Infrastructure Facilities Business Division, Mitsubishi Heavy Industries Machinery Systems, Ltd.

[*4] Principal Engineer, Project Promotion Department, Infrastructure Facilities Business Division, Mitsubishi Heavy Industries Machinery Systems, Ltd.

[*5] Chief Staff Manager, CIS Department, Digital Innovation Headquarters, Mitsubishi Heavy Industries, Ltd.

to convert an input image into another image, which is the purpose of this report.

CycleGAN[(2)] is a method that automatically learns the underlying characteristic differences between two given different image groups A and B and can convert input images belonging to A into images belonging to B. This is "unsupervised learning", which does not require strict correspondence between the image groups A and B, and has the feature that learning data can be easily prepared. Research on applying this method to convert images taken by a vehicle-mounted camera during daytime into ones that appear as if they were taken at night has been conducted[(3)].

2.2 Issue of image conversion

Figure 1 shows an example of image conversion from a daytime image to a nighttime image using CycleGAN described above. (a) is the input daytime image and (b) is the converted nighttime image. Compared to (a), (b) shows an overall decrease in luminance, and looks like a nighttime image. However, it is found by checking the details that there is false information, such as the high-intensity lighting points generated in the air and the non-lit tail lamps of the vehicle running in front, and therefore that the image is not a nighttime image.

The most significant cause of this issue is that CycleGAN is a network that performs learning of the differences between image groups in the image feature space, not learning that takes into account the label information indicating what kind of object something in every area of the image is, and what characteristics it has. In other words, the characteristics of each object shown in the image, such as "this area is the sky and does not glow at night" and "this area is a tail lamp and glows at night," are not taken into account.

Therefore, in this report, we worked on the development of image conversion technology considering the label information of each area.

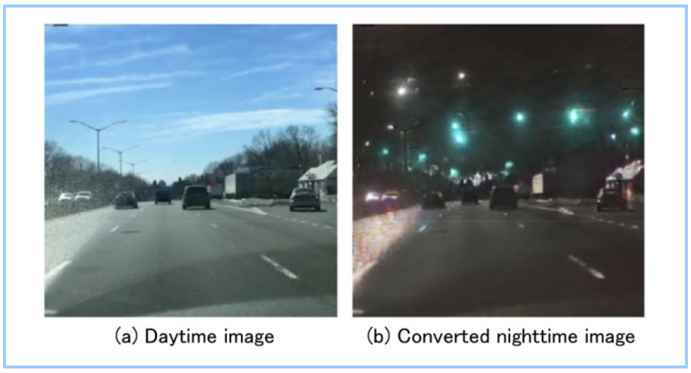

(a) Daytime image
(b) Converted nighttime image

Figure 1 Conversion of daytime image to nighttime image using CycleGAN

2.3 Image conversion that preserves label information of each area

(1) Utilization of semantic segmentation

In order to utilize the label information in each area of an image, we developed a method to allow the CycleGAN generator network to transfer-learn the network trained for semantic segmentation.

Semantic segmentation is a deep learning algorithm that associates labels or categories to all pixels in an image, and is also applied to automated driving and driver assist systems in order to understand traffic scenes. As shown in **Figure 2**, this algorithm can automatically label any input image on a pixel-by-pixel basis by learning the input image and the correct label as a pair[(4)].

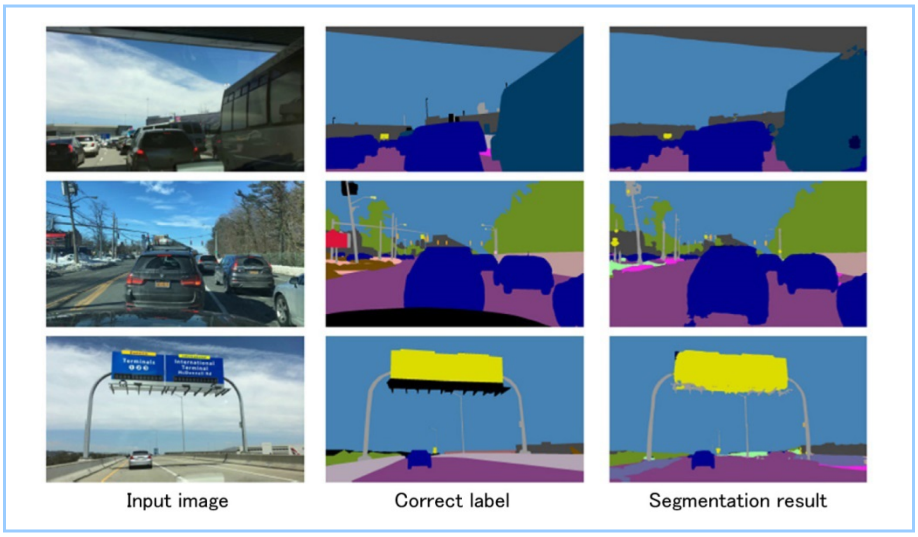


Figure 2　Example of semantic segmentation:

(2) Method to learn networks

Figure 3 shows the architecture of the developed image conversion technology. In Step 1, the network *S* for semantic segmentation is learned using the daytime image X_S and its correct label Y_S. The learning is accomplished by minimizing the loss function *loss1*, which represents the difference between the segmentation result $S(X_S)$ generated by *S* and the correct label Y_S. Thus, as shown in Figure 2, a network capable of pixel-by-pixel labeling can be obtained.

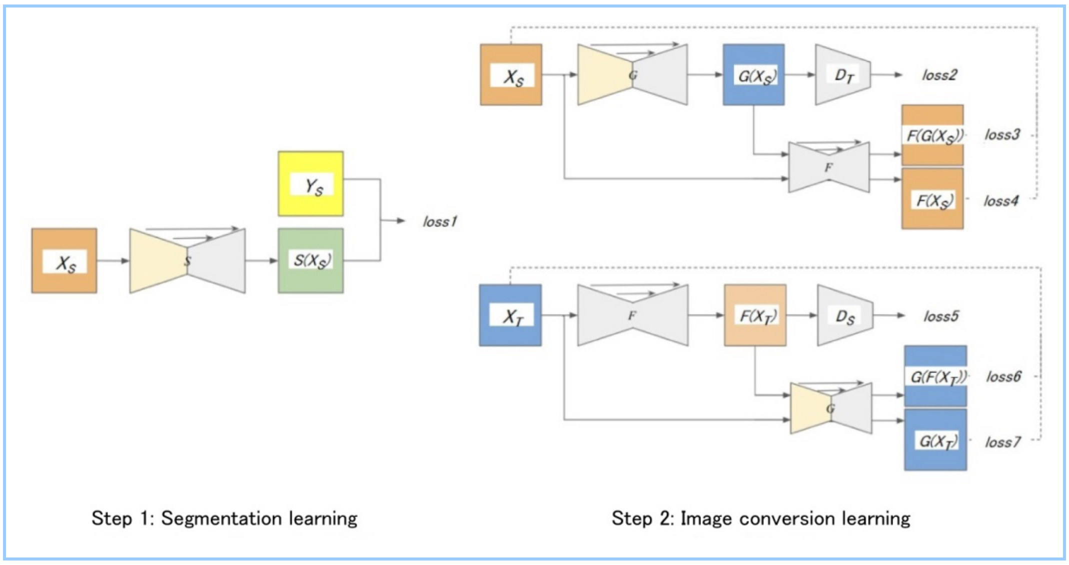


Figure 3　Architecture of learning for developed image conversion technology

Next, in Step 2, based on the network *S* learned in Step 1, networks *G* and *F* of CycleGAN are learned. *G* is a generator that converts a daytime image to a nighttime image and *F* is a generator that converts a nighttime image to a daytime image. X_T is the nighttime image and $G(X_S)$ is the pseudo-nighttime image converted from X_S. D_T is a discriminator for $G(X_S)$, and determines whether $G(X_S)$ is a converted nighttime image or an actual nighttime image. *loss2* is an objective function that expresses the certainty of the discriminator. This value becomes larger when the discriminator judges $G(X_S)$ successfully as a converted nighttime image, and smaller when the discriminator misjudges it as an actual nighttime image. Thus, *G* learns so that *loss2* is minimized and the discriminator learns so that *loss2* is maximized. $F(G(X_S))$ is a daytime image returned from $G(X_S)$ through *F*, which represents the constraint that an image converted by *loss3* needs to be back to the original image when converted further. $F(X_S)$ is an image converted from the daytime image X_S through *F*, which represents the constraint that a daytime image needs to be output as it is when input to *F* where *loss4* converts a nighttime image to a daytime image. *Loss5* to *loss7* are the values when the input image is X_T using the same concept as *loss2* to *loss4*.

Since the network of semantic segmentation is transfer-learned, it is possible to perform the conversion while retaining the label information of each area, which is not considered in general CycleGAN, and thus the conversion accuracy including details is expected to be improved.

3. Experimental results

3.1 Dataset and learning

For the development described in this report, we used Berkeley Deep Drive (BDD) dataset[4], which is widely used in research on automated driving as well as vehicle-mounted camera images of Tokyo metropolitan that we collected independently, as daytime and nighttime images.

The learning of the network for semantic segmentation in Step 1 was performed by using a BDD data set with a correct answer label for each pixel. The learning of the network for CycleGAN in Step 2 was performed in two cases, one using the BDD data set and the other using the Tokyo metropolitan data set.

3.2 Conversion results and evaluation

Figure 4 shows the results of converting daytime images of the BDD and Tokyo metropolitan datasets that were not used for learning to nighttime images. The conventional technology means general CycleGAN-based conversion that is not combined with semantic segmentation. The circles in red show the areas where better conversion results were obtained by the developed technology compared to the conventional technology. In the image converted by the conventional technology, an erroneous conversion occurred by which the streetlights were not lit and the trees were lit. In the image converted by the developed technology, on the other hand, only the streetlights were lit correctly, and the tail lamps of the vehicle running in front are lit more strongly as well. The fact that the conversion to the image unique to nighttime, in which streetlights and tail lamps should be lit was able to be made in this way indicates that the developed technology performed the correct transfer-learning of the label information in the images obtained by segmentation to CycleGAN. In addition, unnatural bright areas that are not seen in the original daytime image are generated on the road surface in the image converted by the conventional technology. In the image converted by the developed technology, on the other hand, no such false signal is generated, which indicates that the label information "road" may suppress erroneous conversion.

These results confirm the effectiveness of the developed technology incorporating semantic segmentation that estimates pixel-by-pixel label information.

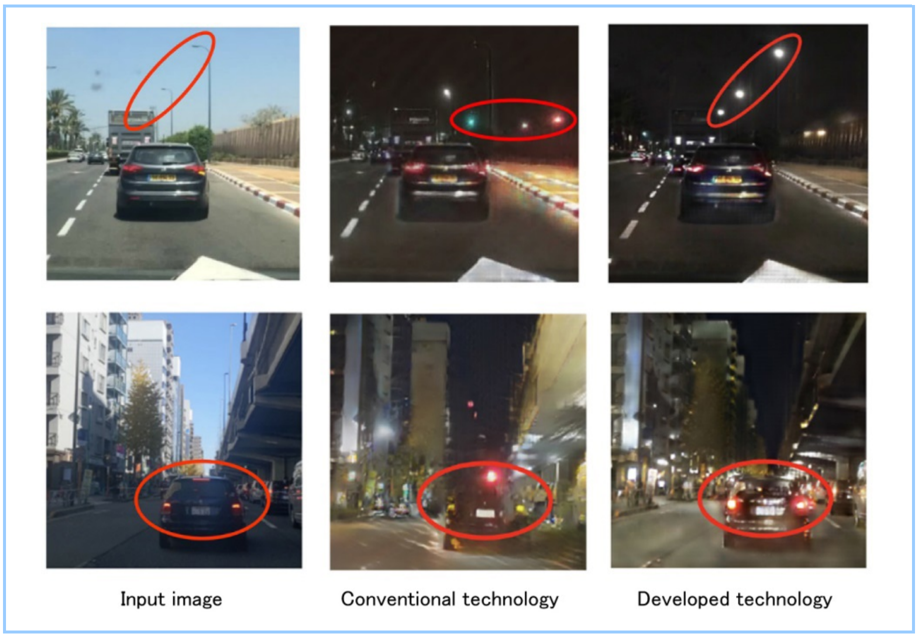


Figure 4 Conversion results for BDD data (upper) and Tokyo metropolitan data (lower)

Table 1 shows the results of quantitative evaluation of the conversion results using the Fréchet Inception Distance (FID)[5], which is used to evaluate the image generation by GAN. FID is the calculated distance between a set of images (A) and another set of images (B) in the feature space. The shorter this distance (smaller the value), the more similar both image sets are. For the dataset used in this development, the FID between the daytime image before conversion and the nighttime image used for learning was 3000, but the FID between the converted nighttime image and the nighttime image used for learning was decreased to 200, which confirms that the conversion was able to make daytime images more similar to nighttime images. In the same manner, the FID between the daytime image before conversion and the daytime image used for learning was 80, and the FID between the converted nighttime image and the daytime image used for learning was 2400, which confirms that the conversion was able to make daytime images more similar to nighttime images.

Table 1 Similarity comparison using FID between converted images

	Image for learning (nighttime)	Image for learning (daytime)
Image before conversion (daytime)	3000	80
Converted image (nighttime)	200	2400

(The smaller the value, the more similar the images are.)

4. Conclusion

We have developed an image conversion method that improves the conversion accuracy of details by utilizing the label information in each area of the image. The results of converting daytime images to nighttime images using this method were quantitatively evaluated by FID, and it was confirmed that this conversion method was able to make daytime images more similar to nighttime images.

By using learning images such as heavy rain, fog, snow, etc., in this method, it will be possible to generate images of different weather conditions in a short time that can be used for verification of automated vehicles. In the future, we will work on the generation of complex disturbance images by learning the time of day and weather conditions separately and combining multiple conversions.

In addition, we are also developing a method to generate bad-weather sensor data for Radar and LiDAR by using deep learning. We would like to contribute to further reduction of time and cost for collecting verification data.

Finally, we would like to express my sincere gratitude to Associate Professor Shunsuke Kamijo Mr. Jinho Lee, and Mr. Daiki Shiotuka of the Kamijo Laboratory of the University of Tokyo Interfaculty Initiative in Information Studies for their cooperation in this research.

References

(1) I. Goodfellow, et al., Generative Adversarial Nets, In NIPS, 2014

(2) J. Zhu, et al., Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, In ICCV, 2017

(3) Shinta Masuda et al., Image Modality Translation for Enriching Virtual Space, The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

(4) F. Yu, et al., BDD100K: A diverse driving video database with scalable annotation tooling, arXiv preprint arXiv, 1805.04687, 2018

(5) M. Heusei, et al., GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, NIPS, 2017