Journal of KIIT. Vol. 20, No. 4, pp. 39-47, Apr. 30, 2022. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2022.20.4.39

Severe Weather Traffic Scene Reconstruction using Inherent Augmented Style Encoding Adversarial Network

Md Foysal Haque*, Dae-Seong Kang**

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (NO.2017R1D1A1B04030870).

Abstract

Vision classification and object detection is a critical task for autonomous systems, essentially for driving assistant systems. Lots of research contributing to improving the vision classification system, especially the Artificial Intelligence (AI) visual analysis technologies have been elevated notably the interest in driving assistance systems. The deep learning-based visual classification methods achieved enormous accuracy in classifying visual scenes in the different fields of vision application. However, the visual classifiers still face some difficulties in working under extreme weather conditions. For example, examining the scenes in severe weather conditions, especially during rainy nights and foggy weather, is the main challenge for vision algorithms. Furthermore, the algorithm struggles to identify the common contexts of the scenes. This paper introduced an adversarial scene reconstructing model that restores dark and uncleared scenes to transform explicit scenes like daytime. After that, the reconstructed image is applied to a recognition algorithm to recognize the autonomous vehicle's visual actions.

요 약

비전 분류 및 객체 감지는 자율 시스템, 기본적으로 운전 보조 시스템에 중요한 작업이다. 비전 분류 시스템을 개선하는 데 기여하는 많은 연구가 진행되었다. 특히 인공지능(AI) 시각 분석 기술은 운전 보조 시스템에 대한 관심이 특히 높아졌다. 딥러닝 기반 시각 분류 방법은 시각 응용 분야의 시각 장면을 분류하는 데 있어 엄청난 정확도를 달성했다. 그러나 시각 분류기는 여전히 극한 기후 조건에서 작업하는 데 몇 가지 어려움에 직면해 있다. 예를 들어, 특히 비가 오는 밤과 안개 낀 날씨에서 심한 날씨 조건에서 장면을 연구하는 것은 시각 알고리즘의 주요 과제이다. 게다가, 시각 알고리즘은 장면의 공통 맥락을 식별하기 위해 노력한다. 본 논문은 어둡고 오래된 장면을 복원하여 낮과 같은 명시적인 장면을 변환하는 적대적 장면 재구성 모델을 제안한다. 그런 다음 재구성된 이미지를 인식 알고리즘에 적용하여 자율 주행 차량의 시각적 동작을 인식한다.

Keywords

generative adversarial network, image generation, image reconstruction, scene classification, object detection

* Dept. of Electronic Engineering, Dong-A University
- ORCID: https://orcid.org/0000-0003-0634-9783
** Professor, Electronic Engineering, Dong-A University
- ORCID: https://orcid.org/0000-0003-0186-2430

· Received: Dec. 23, 2021, Revised: Feb. 23, 2022, Recepted: Feb. 25, 2022
· Corresponding Author: Dae-Seong Kang
Dept. of Electronic Engineering, Dong-A University, 37 NaKdong-Daero 550, Beon-gil Saha-gu, Busan, Korea,
Tel.: +82-51-200-7710, Email: dskang@dau.ac.kr

Ⅰ. Introduction

Real-world visual processing applications require collecting information on the state of different objects to localize and identify specific objects for the required application. However, incomplete and inaccurate localization and identification can create a system failure that can hamper the performance of a real-time visual processing application. For example, driving assistant systems and satellite imaginary examination systems rely on real-time visual processing models. Current research interest in autonomous driving systems accelerated the development of modern technologies [1], [2] growing an exciting research domain. Critical challenges in this area are usually related to computer vision, like understanding ongoing progress and managing unpredictable situations even under difficult weather conditions. Fig. 1 shows an example of scene classification and assisting system for autonomous car.

Autonomous driving assistants deal with the real-time video processing system and numerous weather conditions, which is the most challenging task for the vision application. Extreme weather conditions can affect the vision system of the autonomous driving system to localize objects and to making a decision.

Essential components of the driving assistant system heavily depend on constant and intense image-based object recognition abilities building with cameras. Object recognition has been specified by a variety of different methods and is currently ruled by supervised artificial neural network architectures, such as Faster R-CNN[3], EfficientNet[4], and YOLO[5].

However, these models are still insufficient for operating under uncertain conditions like severe weather conditions especially during rainy nights, and foggy weather is the main challenge for vision algorithms[6]. The mentioned issue can be overcome by linking the scene understanding and image reconstruction strategy to evolve the vision-based application's performance. Recent developments in Generative Adversarial Networks (GANs) introduced a promising approach to image generation[7], image reconstruction[8], and style transfer[9], leading to tremendous computer vision applications to solve real-world vision tasks.

This paper proposed an adversarial scene reconstruction and recognition algorithm to assist an autonomous driving assistant system.

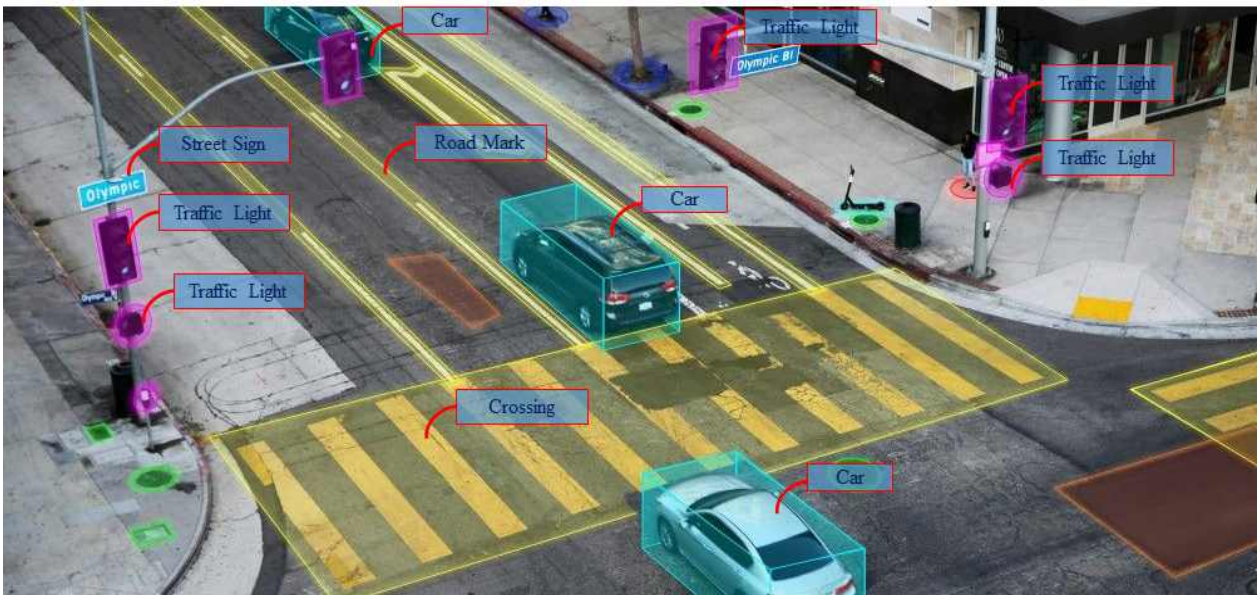


Fig. 1. Scene classification and assisting system for autonomous car

The proposed model consists of two essential components: the GAN-based adversarial network and the object detection module as for scene recognition. These two approaches are employed for scene reconstruction by adversarial network and scene feature extraction, computing optical flow information of the output images for object detection.

II. Related Theories

Scene understanding and activity recognition are core problems of computer vision, which involve several applications, including robot vision, aerospace monitoring, and automotive cars. High-performance computers became the passkey to compute a large amount of data at high speed. The deep learning method derives the feature-based classification approach named the image local features and applies the learning method to perform understanding specific scenarios and action recognition.

Two-stream convolutional models are one of the most typical computer vision solutions for scene classification and recognition. In this paper, two convolutional models are used, one for scene reconstruction. Another one adopts the temporal feature extraction strategy, which examines the actions from images for the feature extraction, which learns from the visible movement vectors of different frames. Then, outputs of the two models are consolidated at the end with the detection task.

2.1 Scene Reconstruction

Context translation aims to learn a function that changes a given image's domain-specific part to the target while maintaining its domain-invariant content [10]. Different vision and graphics applications, e.g., object detection, style transfer, and image super-resolution, can be formulated as image-to-image translation. In recent years, the Generative Adversarial Networks (GANs)[7] have received extensive attention in image-to-image transformation and style transfer tasks.

The proposed adversarial model uses different domain-specific content encoders to learn domain-invariant features for different domains and train the model of each domain using the images from the same domain individually. Although, the domain-invariant features are learned from train images from all the domains. As such, this system effectively translates objects of different scales with multiple backgrounds.

For efficient multi-domain scene translation, the proposed model is constructed with an unsupervised image translation method. The adversarial model utilizes the encoder to encode the domain-invariant features of the target images from multiple domains, and it is named the style encoder. The style Encoder is constructed into the generator, and it conditionally shares among different domains.

2.2 Temporal-Spatial Feature Extraction

The reconstructed images are first fed into a temporal-spatial feature learning module, where the detection framework learns features information. An illustration of the components of the temporal-spatial feature learning module is given in Fig. 2.

According to the typical architecture for feed-forward neural networks, these convolutional neural networks are built by interlacing convolution layers and average pooling layers. First, the convolution layer performs a nonlinear mapping, and the pooling layer learns object information. Then, referring to the learned object information, the detection model examines and performs the detection task.

$$(f \otimes n)_i = \frac{1}{|N(i)|} \sum_{M \epsilon N(i)} f(M).n(u(i,j)) \qquad (1)$$

In equation 1, features denote $f_{(i)}$, $N(i)$ denote the set of neighbors of node i, and $n(u(i,j))$ denote

the weight parameter constructed from the kernel function n.

III. Proposed Algorithm

Image context transformation is one of the crucial tasks in adversarial learning. Several robust algorithms [11], [12] were introduced to solve the image context transformation task. However, those methods face difficulties translating a single image to a set of different images in a target domain. Therefore, the proposed network introduce transforming extreme weather scenes to the normal scenario and performing the detection task through the image.

3.1 Inherent Augmented Style Encoding Adversarial Network

The aim of this model is to learn object context mapping among multiple domains using an adversarial network. The overall network structure of our proposed Inherent Augmented Style Encoding Adversarial Network is shown in Fig. 2. The Inherent Augmented Style Encoding Adversarial Network scene reconstruction model is constructed adopting the architecture of AINN[13]. The adversarial network consists of a Style Encoder, a Generator, and a Discriminator.

Initially, the adversarial model accepts the training image and examines the images using the Style Encoder to map the image context information, and this context information contains the core information of both domains of the image. Then the extracted object context information is delivered to the generator model. The generator model examines the training domains and conducts the scene transformation task by referring to the data of Style Encoder. The adversarial model conducts the scene reconstruction task following equation 2.

$$D_{(a \Leftrightarrow b)} = I_a\left[D(b)-1)^2 + I_b[D(G(a))^2\right] \quad (2)$$

where, I_a and I_b are the image domains, G is generator, and the discriminator is D.

A Discriminator network examines the overall scene reconstruction process. The network examines the reconstructed samples and matches the samples with original images.

However, reconstruction task, the discriminator model cannot reach the satisfactory level to reconstruction task. In that case, the discriminator model sends a loss score as feedback to the generator model, and the generator model continues the reconstruction process until it reaches the reconstruction goal. The reconstruction loss function is shown in equation 3.

$$L_r(G) = E_{x,y}\left[1-\log D_{(r)}(y/x)\right] \quad (3)$$

where, L_r denotes reconstruction loss of generator, x and y is domain samples, D and G denotes discriminator and generator.

3.2 Temporal-Spatial Feature Extraction Network for Detection

After the scene reconstruction, the reconstructed image applies to the detection algorithm for scene classification and object detection. In this term, the detection algorithm formed with the enhanced feature extraction model known as the temporal-spatial feature extractor. The proposed detection network is based on Single-Shot Multibox Object Detector (SSD)[14]. The network architecture of the proposed object detection model is shown in Fig. 2.

The temporal-spatial feature extractor network examines the input samples to examine the contents and collect the content information.

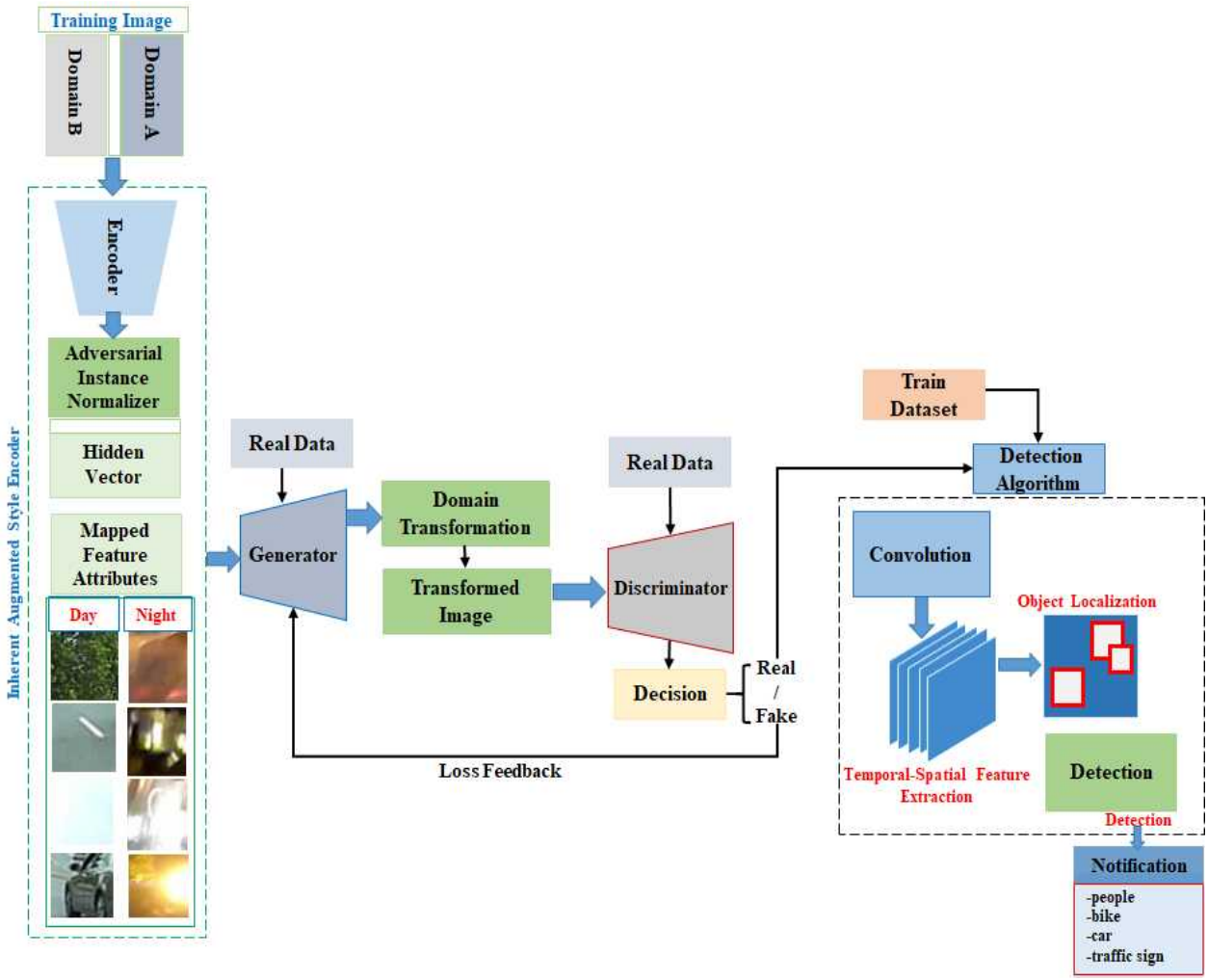


Fig. 2. Network architecture of inherent augmented style encoding adversarial network

Then, this feature information is precisely computed by dense convolutional layers and prepared for the localization process. The localization block examines the feature context information and performs the detection task. Finally, the detection block predicts the detection score and prepares notifications with the detection results. The detection model follows the localization detection process following the equation 4.

$$B_b = \sum_{pos}^{N} C\left(l(x,c) + \varphi G_L(x,y)\right) \quad (4)$$

In equation 4, C is confidence map, l is predicted box, G_L is the ground truth localization, feature coefficient is φ, and object position is x, y.

IV. Experiments

4.1 Experimental Environment

The adversarial model is trained with two different datasets for the scene reconstruction task. First, the Alderley Day/Night dataset[15] is used to reconstruct the rainy night image to the day image. Second, the Cityscape Fog dataset[16] is used to reconstruct the scene from foggy to typical day images. Table 1 shows the datasets statistics in the experiments.

For both adversarial reconstruction experiments, the input image is 300×300 pixels. For optimization the Adam optimizer used to train the model[17] with β1 = 0.5, β2 = 0.999. In each minibatch consists of one image from each domain.

Table 1. Dataset Information for scene reconstruction

	Alderley [15]		Cityscape [16]	
	Day	Night	Foggy	Normal
Train	3000	3000	1800	1800
Test	600	600	600	600

In the detection network, the input size is set to 300×300. Moreover, the detection algorithm training dataset is prepared by combining Cityscape and Alderley Day/Night datasets. The training parameters are set as the initial learning rate 0.001 with SGD, momentum set to 0.9, batch size 32, and weight decay set to 0.0005.

V. Results

The proposed model performs well in scene reconstruction with wider diversity. However, the conventional methods are less effective in scene reconstruction tasks that include significant shape variations, e.g, translating the background objects into the image images. The scene reconstruction results are shown in Fig. 3 and 4.

The proposed method is evaluated against the CycleGAN[18], AINN[19], DRIT[10], and SingleGAN [20] models. The CycelGAN and AINN network is capable of the multi-context reconstruction task, but the proposed network achieved a higher evaluation score than the convention models for multi-context reconstruction. The scene reconstruction evaluation results are illustrated in Table 2.

Meanwhile, the detection model is trained with the Cityscape and Alderly dataset to recognize the traffic scenes to identify the object and notify about the detected objects.

Fig. 3. Comparison of reconstructed results between proposed model and SingleGAN

Fig. 4. Reconstructed results of proposed network (foggy to normal day scenario)

Fig. 5. Detection results of the proposed network

Table 2. Comparison of the results of scene reconstruction with robust adversarial models

Method	Dataset	Input size	Reconstruction accuracy(%)
Proposed model	Alderley	300×300	88.1
	Cityscape		88.6
AINN [13]	Alderley		87.3
DRIT [10]			76.9
SingleGAN [20]			79.6
CycelGAN [18]			79.7

The proposed network's detection results are shown in Fig. 5, and the comparison of detection results with the other robust detection model is illustrated in Table 3.

Table 3. Summary of mAP (mean-average precision) for each category, inference speed, and average detection accuracy

Method / Class	Proposed network	FRCN+VGG [21]	ResNet50-Det [22]
	FPS		
	33.8	1.7	22.5
	Average mAP		
	89.5	74.8	76.1
Bus	**91.4**	83.5	81.3
Bike	**89.6**	81.3	80.0
Car	**89.8**	86.1	86.7
People	**91.7**	87.3	89.3
Sign	**84.1**	78.4	80.2
Traffic sign	**85.7**	79.2	78.3

Analyzing all the reconstruction results, the proposed model achieved enhanced performance in translating rainy night scenes into clear daylight scenes and foggy scenarios to normal day scenarios.

Moreover, the reconstructed image detection task was conducted, and the object detection module is trained with six different classes of objects. The scene reconstruction task detection task also achieved higher accuracy in detecting and understanding the scenes.

Ⅵ. Conclusions

This paper introduced a scene reconstruction and scene understanding framework applied to reconstruct the rainy-night image scene to add daylight and foggy images to the clear day scenario image. The network adopts an adversarial network for visual scene reconstruction and detecting objects a temporal-spatial feature extraction detection model used for the driving assisting system. The network reconstructs the daylight image to enhance visibility without harming any context information of the base image, which leads to assisting the module's mainframe in notifying the detection information. The proposed network achieved 88.1 percent accuracy in the rainy-night image to day image reconstruction task and 88.6 percent in the foggy image to day image reconstruction task. In addition, the network achieved 89.2 percent accuracy in the scene understanding object detection task. In the future, we will aim to include the pedestrian and road patterns information to increase the working area of the proposed mode by improving the network architecture.

References

[1] F. Rundo, et al., "Advanced Car Driving Assistant System: A Deep Non-local Pipeline Combined with 1D Dilated CNN for Safety Driving", In IMPROVE, pp. 81-90, Jan. 2021. http://dx.doi.org/10.5220/0010381000810090.

[2] V.Q. Nguyen, et al., "A Study on Real-time Detection Method of Lane and Vehicle for Lane Change Assistant System using Vision System on Highway", Engineering science and technology an international journal, Vol. 21, No. 5, pp. 822-833, Oct. 2018. https://doi.org/10.1016/j.jestch.2018.06.006.

[3] S. Ren, et al., "Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks", Advances in neural information processing systems, Vol. 28, pp. 91-99, Jun. 2015. https://doi.org/10.48550/arXiv.1506.01497.

[4] M. Tan, et al., "Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks", In International Conference on Machine Learning, pp. 6105-6114, May 2019. https://doi.org/10.48550/arXiv.1905.11946.

[5] A. Bochkovskiy, et al., "Yolov4: Optimal Speed and Accuracy of Object Detection", arXiv preprint arXiv:2004.10934, Apr. 2020. https://doi.org/10.48550/arXiv.2004.10934.

[6] M. Schutera, et al., "Night-to-Day: Online Image-to-Image Translation for Object Detection Within Autonomous Driving by Night", IEEE Transactions on Intelligent Vehicles, Vol. 06, No. 3, pp. 480-489, Nov. 2020. https://doi.org/10.1109/TIV.2020.3039456.

[7] I. J. Goodfellow, et al., "Generative Adversarial Networks", Communications of the ACM, Vol. 63, No. 11, pp. 139-144, 2020.

[8] A. Pajot, et al., "Unsupervised Adversarial Image Reconstruction", In International Conference on Learning Representations, 2018.

[9] Y. Viazovetskyi et al., "Stylegan2 Distillation for Feed-forward Image Manipulation", in European Conference on Computer Vision, Glasgow, UK, pp. 170-186, Aug. 2020.

[10] H. Y. Lee, et al., "Diverse Image-to-image Translation via Disentangled Representations", ECCV 2018, arXiv preprint arXiv:1808.00948, Munich, Germany, Sep. 2018. https://doi.org/10.48550/arXiv.1808.00948.

[11] Y. Choi, el al., "Stargan v2: Diverse Image Synthesis for Multiple Domains", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, pp. 8188-8197, Jun. 2020. https://doi.org/10.1109/CVPR42600.2020.00821.

[12] N. Takano, et al., "Srgan: Training Dataset Matters", arXiv preprint arXiv:1903.09922, Mar. 2019. https://doi.org/10.48550/arXiv.1903.09922.

[13] M. F. Haque, et al., "AINN: Adversarial Instance Normalization Network for Image-to-Image Translation", In Proceedings of KIIT Conference, Cheongju, Korea, pp. 117-120. Sep. 2020.

[14] M. F. Haque, et al., "Multi Scale Object Detection based on Single Shot Multibox Detector with Feature Fusion and Inception Network", Journal of KIIT, Vol. 16, No. 10, pp. 93-100, Oct. 2018. https://doi.org/10.14801/jkiit.2018.16.10.93.

[15] M. J. Milford, et al., "SeqSLAM: Visual Route-Based Navigation for Sunny Summer Days and Stormy Winter Nights", In IEEE international conference on robotics and automation, Saint Paul, MN, USA, pp. 1643-1649, May 2012. https://doi.org/10.1109/ICRA.2012.6224623.

[16] M. Cordts, et al., "The Cityscapes Dataset for Semantic Urban Scene Understanding", In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, pp. 3213-3223, Jun. 2016. https://doi.org/10.1109/CVPR.2016.350.

[17] D. P. Kingma, et al., "Adam: A Method for Stochastic Optimization", arXiv preprint arXiv:1412.6980, Dec. 2014. https://doi.org/10.48550/arXiv.1412.6980.

[18] J.Y. Zhu, et al., "Unpaired Image-to-image Translation using Cycle-consistent Adversarial Networks", in Proceedings of the IEEE International Conference on Computer Vision, pp. 2223-2232, Mar. 2017. https://doi.org/10.48550/arXiv.1703.10593.

[19] M. F. Haque, et al., "Adversarial Scene Reconstruction and Object Detection System for Assisting Autonomous Vehicle", arXiv:2110.07716, Oct. 2021. https://doi.org/10.48550/arXiv.2110.07716.

[20] X. Yu, et al., "SingleGAN: Image-to-image Translation by a Single-generator Network using Multiple Generative Adversarial Learning", arXiv preprint arXiv:1810.04991, Oct. 2018. https://doi.org/10.48550/arXiv.1810.04991.

[21] K. Ashraf, et al., "Shallow Networks for High-accuracy Road Object Detection", arXiv:1606.01561, Jun. 2016.

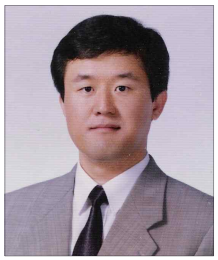
[22] B. Wu, et al., "SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, pp. 129-137, Jul. 2017. https://doi.org/10.1109/CVPRW.2017.60.

Authors

Md Foysal Haque

2018 ~ 2020 : M.Sc. in Electronic Engineering, Dong-A University.
2020 ~ present : Ph.D. candidate in Electronic Engineering, Dong-A university
Research interests : Digital image processing, computer vision and pattern recognition

Dae-Seong Kang



1994 : Ph.D. in Electrical Engineering, Texas A&M University
1995 ~ present : Professor at Dong-A University
Research interests: Image processing and compression