https://doi.org/10.7840/kics.2024.49.7.976

AED-Net: Attention-Based Detection Model for Disabled Signage Detection

Akhrorjon Akhmadjon Ugli Rakhmonov◆, Barathi Subramanian*,
Bahar Amirian Varnousefaderani*, Jeonghong Kim°

ABSTRACT

The aim of having designated parking spaces for individuals with disabilities is to ensure that only vehicles with proper handicapped signage use them, while preventing unauthorized vehicles from occupying those spaces. To achieve this, real-time monitoring is essential. Existing two-stage object detection models suffer from slow image processing and enhanced backbones with feature pyramid networks are also burdened with expanded parameters. While YOLOv5 model is a compelling choice due to its superior speed and performance compared to existing models. Therefore, this study proposes to make certain modifications to a baseline YOLOv5 model. Instead of the original 9 blocks in the backbone and 4 C3 blocks, we propose to replace them with 6 and 4 EfficientNet blocks, accordingly. These EfficientNet blocks have fewer parameters but still offer higher accuracy in detecting disabled signs, among other types of signs on car windshields. To make up for the reduced number of blocks, we have incorporated an attention mechanism into the proposed architecture before the detection phase. This mechanism enables the model to focus on the crucial regions required for the task. Furthermore, we propose utilizing a more advanced optimizer called AdamW to prevent overfitting. With these enhancements, a novel object detector, attention-based efficient detection model (AED-Net) is proposed. To assess the effectiveness of the proposed approach, we will gather and label a dataset comprising images of cars displaying disabled signage on their windshields. Experiments conducted using this dataset demonstrate that the proposed model achieves a superior F1 score of 0.73 compared to that of baseline model, 0.57. The proposed model utilizes 10 percent fewer parameters compared to the baseline model.

Key Words : Depthwise Separable Convolution, Disabled Signage, Small Object Detection

Ⅰ. Introduction

Detecting small objects in images can be challenging due to limited resolution and contextual information[1]. This is especially true for real-time object detection systems that have constrained computational resources. In parking lots, handicap parking spaces are reserved for vehicles displaying a disabled person sign. However, some drivers unlawfully park in these spots. To address this issue and enforce the benefits for disabled drivers, accurate real-time detection and recognition systems are crucial.

Efforts have been made to improve the detection of smaller objects, but existing methods have limitations[2,18]. Some methods focus on specific image regions or use slower two-stage detectors[3,4]. Single-stage detectors have been developed for

※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1I1A3043970).
◆ First Author : Kyungpook National University, Department of Computer Science and Engineering, r.akhror@knu.ac.kr, 정회원
° Corresponding Author : Kyungpook National University, Department of Computer Science and Engineering, jhk@knu.ac.kr, 종신회원
* Kyungpook National University, Department of Computer Science and Engineering, {barathi.subramanian, bahar.amirian.v}@knu.ac.kr
논문번호 : JKICS202401-004-D-RN, Received December 29, 2023; Revised March 28, 2024; Accepted April 6, 2024

real-time applications[5], and YOLOv5 is a popular choice[6]. However, YOLOv5's accuracy in detecting small objects is reduced.

To enhance YOLOv5's performance in detecting small objects, we propose a modified model. We replace the C3 blocks with more efficient EfficientNet blocks[13], known for their optimized filter sizes and reduced parameters. This maintains real-time processing capabilities while improving accuracy. EfficientNet utilizes depthwise separable convolutions, dividing the operation into depth-wise and point-wise convolutions. This reduces the number of parameters and allows the network to learn spatial and channel-wise relationships independently. The model's size and depth can be adjusted using scaling methods to fit different computational budgets. The proposed model also incorporates attention layers to focus on relevant regions during feature extraction, disregarding less important areas. We aim to identify small disabled signs on car windshields, and attention mechanisms improve the model's focus on these specific regions. To enhance generalization, we use the AdamW optimizer instead of Adam[20]. AdamW addresses weight decay issues and offers improved regularization and generalization performance.

The contributions of this study are as follows:

1) A revised YOLOv5 architecture is proposed, replacing C3 blocks with EfficientNet blocks, balancing computational requirements and accuracy.
2) Attention layers are integrated to improve accuracy in detecting small disabled signs, crucial for real-world applications.
3) We conduct experiments that illustrate the superior results of the proposed method on the custom dataset than the baseline method.

The rest of the paper is organized as follows: Section 2 explores existing object detection methods, Section 3 explains our proposed approach, Section 4 presents experimental results, and Section 5 provides conclusions and suggestions for future work.

Ⅱ. Related Work

Object detectors can be categorized into two main types: one-stage detectors and two-stage detectors. Two-stage detectors like Fast R-CNN[14] and Faster R-CNN[8] generate region proposals before classifying them, prioritizing accuracy over inference time. Although there have been attempts to improve the detection of small objects, these methods often sacrifice inference speed. Single shot detectors (SSD)[15], on the other hand, eliminate the need for a separate region proposal stage but face challenges when detecting small objects due to shallow layers lacking deep semantic information.

YOLO has gained popularity as a family of object detectors. YOLOv1[9] introduced faster models by treating detection as a regression task but faced accuracy limitations with small objects. YOLOv2[16] improved recall by removing fully connected layers and introducing anchor boxes for bounding box prediction. YOLOv3[17] incorporated binary cross-entropy loss and a ResNet backbone to enhance class prediction and detect small objects. However, the computational resources required for YOLOv3 limited its real-time deployment.

YOLOv5[10], unrelated to YOLOv4[11], offers similar performance and design but is implemented using PyTorch, making it more accessible and usable in various environments. Furthermore, models in YOLOv5 are significantly smaller, faster to train and more usable in real-world applications. Fig. 1 illustrates the default structure of YOLOv5 model. The YOLOv5 model consists of the backbone, neck, head, and non-max suppression parts. While C3 blocks in YOLOv5 contribute to accuracy, their parameter-heavy nature hinders real-time implementation.

Efforts have been made to prioritize specific regions of the input image for improved resolution[3] and object definition, but this approach is less suitable for real-time systems. Managing feature maps, such as with feature pyramid networks (FPN), can enhance the backbone in different ways[12,13], but it increases the number of parameters and compromises speed.

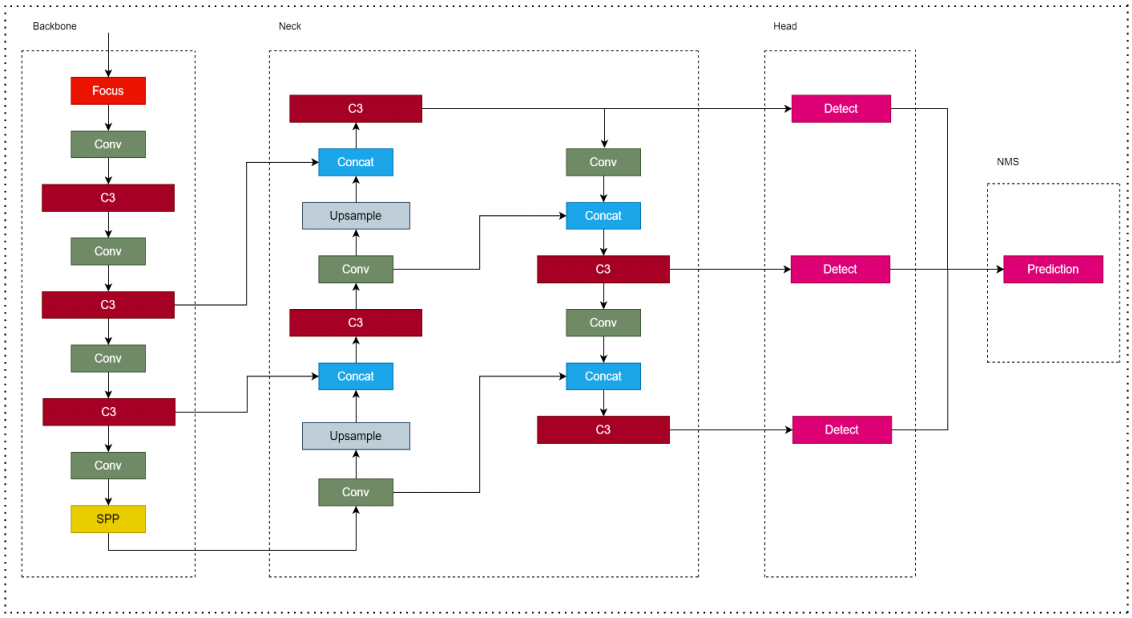

Fig. 1. The default architecture of YOLOv5 model.

Ⅲ. Proposed Method

In this section, we thoroughly present the details of the proposed method.

3.1 Data Pre-Processing

To enhance the dataset's variety, data augmentations such as RandomCrop, RandomGrayscale, RandomHorizontalFlip, and GaussianBlur are employed. The dataset is then annotated manually using the LabelImg tool, involving the cropping of images and the drawing of bounding boxes around the handicapped driver sign.

3.2 Data Learning

Following the dataset's pre-processing, the model is trained to address the objective of detecting disabled signages and accurately outlining bounding boxes around them.

3.3 Inference

In the inference stage, the model is assessed using new data to evaluate its performance in detecting the handicapped signage among other signs located on the rear window of the vehicles.

3.4 Proposed Model Architecture

The proposed modifications aim to enhance the YOLOv5 model's efficiency and accuracy. Key changes include integrating a 6-block-EfficientNet into the backbone for improved efficiency and reduced computational complexity. The spatial pyramid pooling (SPP) layer is maintained in the final layer to ensure consistent feature map size and accurate predictions for objects of different sizes. In the neck, all C3 blocks are replaced with EfficientNet blocks, further boosting efficiency and accuracy. An attention layer is introduced before detection to effectively handle tiny objects and improve accuracy. This attention mechanism, computed using self-attention[19], enables the model to learn the importance of different regions based on their relationships. The attention mechanism can be computed as follows.

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where Q represents a query, K represents a key vector, V represents a value vector, d represents the dimensionality of a key vector. The overall model architecture, including these modifications, is illustrate

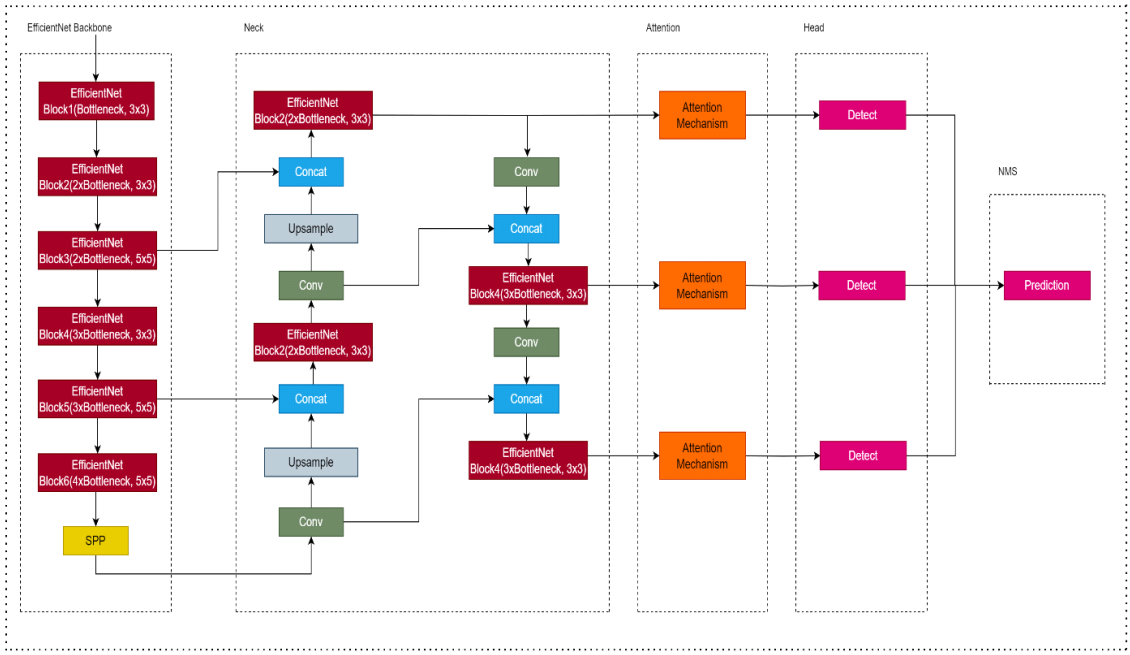

Fig. 2. The architecture of the proposed model.

d in Fig. 2.

In addition to architectural modifications, a different optimizer, namely AdamW, is utilized in contrast to the original YOLOv5 model. AdamW is a variant of the Adam optimization algorithm, an extension of stochastic gradient descent. It includes weight decay, which regularizes the model by penalizing large weight values, effectively preventing overfitting.

Ⅳ. Experimental Results

4.1 Dataset Description

Our custom dataset consists of photographs taken with mobile phones, capturing cars with disabled signs on their front windows. The dataset comprises 1025 images with dimensions of 1920x1080, which are resized to 800x800. We divide the dataset into two subsets: the train set and the validation set, which account for approximately 90% and 10% of the total data, respectively.

4.2 Training Details

1) Experimental settings; The proposed model was implemented using Python version 3.9.13 on a personal computer with 32GB of RAM and an Intel i5 2.90GHz CPU, running the 64-bit version of Windows 10, with one 8GB NVIDIA GeForce RTX 2060 SUPER GPU with CUDA 11.0

2) Evaluation metrics: To assess the performance of the model, various loss metrics are employed, encompassing box loss, object loss, and classification loss. These metrics provide insights into different aspects of the model's performance during training and evaluation.

In addition to loss metrics, precision, recall, and F1 score metrics are utilized to gain a more comprehensive understanding of the model's performance. These metrics provide information about the model's ability to correctly identify and classify objects within the dataset. The formulas for them are as follows.

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision \pm Recall} \tag{4}$$

where TP is a true positive, TN is a true negative, FP is a false positive, and FN is a false negative.

4.3 Results

After a mere 20 epochs of training, the proposed model demonstrated superior performance compared to the baseline model in terms of considered evaluation metrics. Table 1 presents a comparison of precision (P), recall (R), and F1 score.

As can be seen the proposed model outperformed the baseline model in P, R, and F1 with 0.71, 0.77, and 0.73, accordingly. Whereas the baseline model achieved 0.6, 0.55, and 0.57 in P, R, and F1, respectively. The results of train and validation losses of the models are illustrated in Fig.3.

As shown the proposed model demonstrated less loss values both during training and validation time owing to the contributions. The proposed model benefits better generalization because of contributions.

Table 1. Performance comparison of the considered models

Model	P	R	F1
Baseline YOLOv5	0.6	0.55	0.57
Ours	0.71	0.77	0.73

It is also important to note that the proposed method owing to the contributions utilizes 10 percent less trainable parameters than the baseline YOLOv5 model. Therefore, it is more suitable for real-time practical applications.

V. Conclusions and Future Work

Efficiently monitoring designated parking spaces for individuals with disabilities is challenging considering the necessity to detect small disabled

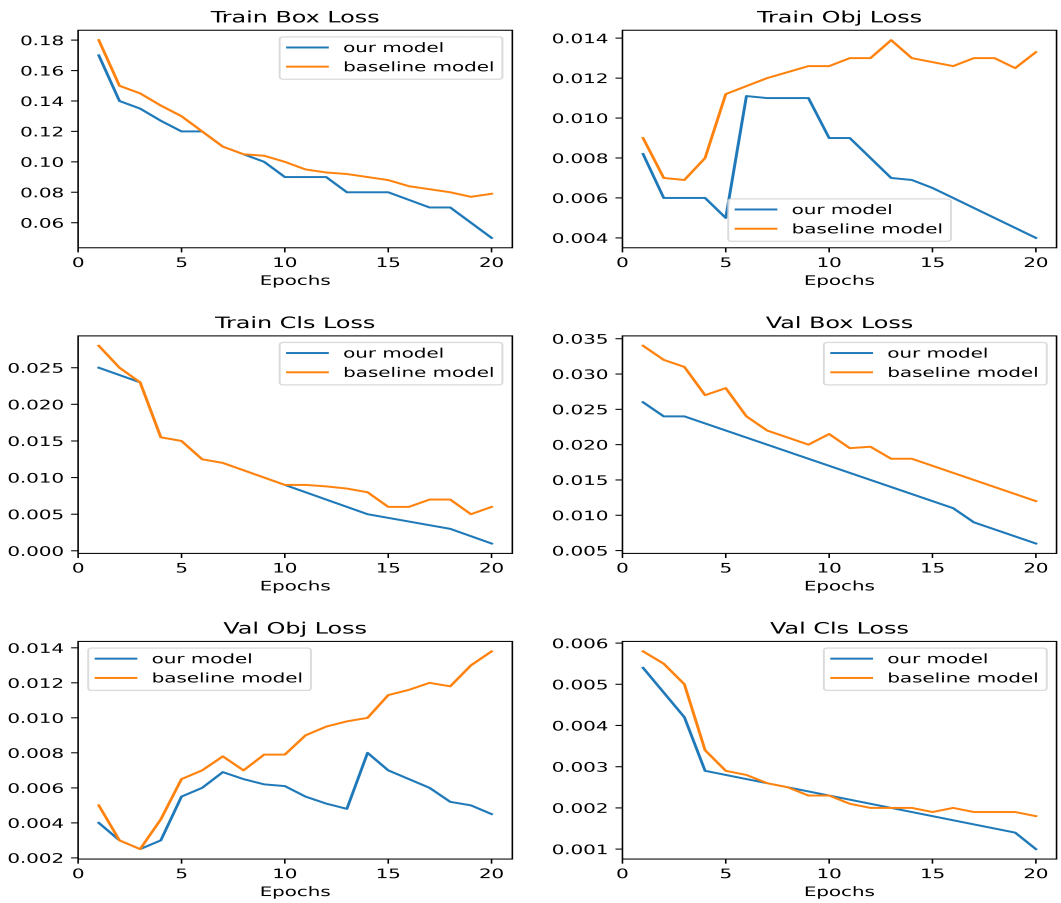

Fig. 3. The performance of the models in terms of the losses.

signage. YOLOv5 possesses numerous advantages compared to existing methods such as faster image processing and better superior latency. In this work, we proposed the AED-Net for real-time object detection. We replaced C3 layers in the backbone and neck of YOLOv5 with more efficient EfficientNet blocks, used a different optimizer and added an attention mechanism for better focus. As a result, our model outperformed the baseline with improved precision, recall, F1 score, and less loss values. The future can be to use more comprehensive techniques to handle tiny objects and evaluate the method with other benchmark datasets additionally.

References

[1] G. Cao, X. Xie, W. Yang, Q. Liao, G. Shi, and J. Wu, "Feature-fused SSD: Fast detection for small objects," in *Ninth ICGIP 2017*, vol. 10615, pp. 381-388, SPIE, 2018. (https://doi.org/10.48550/arXiv.1709.05054)

[2] N. D. Nguyen, T. Do, T. D. Ngo, and D. D. Le, "An evaluation of deep learning methods for small object detection," *J. Electr. and Comput. Eng.*, vol. 2020, pp. 1-18, 2020. (https://doi.org/10.1155/2020/3189691)

[3] B. Singh, M. Najibi, A. Sharma, and L. S. Davis, "Scale normalized image pyramids with autofocus for object detection," *IEEE Trans. Pattern Anal. and Mach. Intell.,* vol. 44, no. 7, pp. 3749-3766, 2021. (https://doi.org/10.48550/arXiv.2102.05646)

[4] B. Singh, M. Najibi, and L. S. Davis, "Sniper: Efficient multi-scale training," *Advances in NIPS*, vol. 31, 2018. (https://doi.org/10.48550/arXiv.1805.09300)

[5] B. Wu, F. Iandola, P. H. Jin, and K. Keutzer, "Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving," in *Proc. IEEE Conf. CVPR*, pp. 129-137, 2017. (https://doi.org/10.48550/arXiv.1612.01051)

[6] A. Benjumea, I. Teeti, F. Cuzzolin, and A. Bradley, "Yolo-z: Improving small object detection in yolov5 for autonomous vehicles," *arXiv preprint arXiv:2112.11798*, 2021. (https://doi.org/10.48550/arXiv.2112.11798)

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in NIPS*, vol. 28, 2015. (https://doi.org/10.48550/arXiv.1506.01497)

[8] G. Plastiras, C. Kyrkou, and T. Theocharides, "Efficient convnet-based object detection for unmanned aerial vehicles by selective tile processing," in *Proc. 12th Int. Conf. Distrib. Smart Cameras*, pp. 1-6, 2018. (https://doi.org/10.48550/arXiv.1911.06073)

[9] J. B. G. Jocher and A. Stoken, ultralytics/ yolov5: v5.0 - YOLOv5-P6 1280 models, 2021. (https://doi.org/10.5281/zendo.4154370)

[10] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv: 2004.10934*, 2020. (https://doi.org/10.48550/arXiv.2004.10934)

[11] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. CVPR*, pp. 2117-2125, 2017. (https://doi.org/10.48550/arXiv.1612.03144)

[12] A. A. U. Rakhmonov, B. Subramanian, and J. H. Kim, "Airy YOLOv5 for disabled sign detection," in *2023 Fourteenth ICUFN*, pp. 869-874, 2023. (https://doi.org/10.1109/ICUFN57995.2023.10200853)

[13] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Int. Conf. Mach. Learn. PMLR, 2019*, pp. 6105-6114, 2019. (https://doi.org/10.48550/arXiv.1905.11946)

[14] R. Girshick, "Fast r-cnn," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 1440-1448, 2015. (https://doi.org/10.1109/ICCV.2015.169)

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Comput. Vision-ECCV 2016: 14th Eur. Conf.,*

Amsterdam, The Netherlands, Oct. 11-14, 2016, Proc., Part I 14. Springer, pp. 21- 37, 2017. (https://doi.org/10.48550/arXiv.1512.02325)

[16] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. IEEE Conf. CVPR*, pp. 7263-7271, 2017. (https://doi.org/10.48550/arXiv.1612.08242)

[17] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018. (https://doi.org/10.48550/arXiv.1804.02767)

[18] Y. H. Kim, J. H. Kim, and H. H. Park, "Light-CAM: A lightweight model for weakly supervised object localization of embedded devices," *J. KICS*, vol. 47, no. 8, 2022. (https://doi.org/10.7840/kics.2022.47.8.1144)

[19] A. Vaswani, et al., "Attention is all you need," in *Proc. 31st Annu. Conf. Neural Inf. Process. Syst.*(NeurIPS), pp. 5998-6008, 2017. (https://doi.org/10.48550/arXiv.1706.03762)

[20] I. Loschilov, F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019. (https://doi.org/10.48550/arXiv.1711.05101)

Akhrorjon Akhmadjon Ugli Rakhmonov

Sep. 2022~Current : Ph.D. student, Kyungpook National University

<Research Interests> Anomaly Detection, Computer Vision, Deep Learning, Machine Learning

[ORCID:0009-0002-0392-0307]

Barathi Subramanian

Feb. 2024 : Ph.D. degree, Kyungpook National University

<Research Interests> Gesture Recognition, Computer Vision, Deep Learning, Machine Learning

Bahar Amirian Varnousefaderani

Feb. 2017~Current : M.S. degree, Kyungpook National University

<Research Interests> Anomaly Detection, Computer Vision, Deep Learning, Machine Learning

Jeonghong Kim

Feb. 2001 : Ph.D. degree, Chungnam National University

<Research Interests> Anomaly Detection, Computer Vision, Deep Learning, Machine Learning

[ORCID:0000-0002-7466-1376]