AFSNet: Multi-Scale Adaptive Feature Scaling Convolutional Network for Real-time Object Detection

Md Foysal Haque[1†] **and Dae-Seong Kang**[2††],

Department of Electronic Engineering
Dong-A University, Busan, Korea

Summary
Deep learning approaches showed significant performance in current computer vision tasks particularly on image classification and object detection. Object detection is growing its popularity for video surveillance systems and recognizing objects. Convolutional neural networks (CNNs) opens a wide path for computer vision applications. In this paper, we proposed a method to enable select target classes for detection, produce an initial detection representation by selecting a specific portion of the image and maintain the trainset for detection model. The method achieved noticeable detection results to detect multi-class objects. However, object detection frameworks face difficulties to localize small objects. The main cause of this problem is to adopt exact feature mapping and extracting strategy. Due to the low pixel value feature extractors unable to map and localize the small objects. To eliminate the issue we designed a convolutional network named Adaptive Feature Scaler (AFS) Convolutional network. The network constructed to localize and extract exact feature data to detect multi-class objects.

Key words:
convolutional neural network, object detection, feature scaling, YOLO, AFSNet.

1. Introduction

Mapping and scaling the exact visual part of an image is a crucial task in object detection. In order to achieve the multi-class detection, the framework needs to improve the feature localization and extraction process. Moreover, lack of proper feature extracting strategy the computational complexity is increased which reduces the detection performance of the base network. To extract feature information image classifier uses several extraction methods that divided into two different classes (i) hand-engineered feature extractors [1], and (ii) convolutional neural networks (CNNs) [2]. CNN achieved noticeable success in the different branches of computer vision. Recently, hand-engineered feature extractors are replaced with improved feature computed convolutional networks. Which is more advance and improves the detection performance by extracting high-level feature data. The image classifiers are designed with different types of convolutional layers. These layers are varied by their working principles. Regarding the working strategy, the depth of the layers varies to conduct a specific task properly. Furthermore, the most important part of a visual classifier is the feature extraction block. The feature extractor classifies all visual information of the dataset and collects informative feature information to localize and detect objects.

Regarding the goal of the visual classifier the structure of the network divided into two classes (i) single-stage image classifiers, and (ii) two-stage image classifiers. Single-stage image classifiers [3]-[5] use a simple convolutional structure for the visual classification task. The network performance in relatively better than the two-stage classifier on some specific task. However, the performance of this network is not satisfactory for the complex classification task [6]. On the other hand, two-stage image classifiers [7]-[9] deploys multi-stage localization and classification blocks to solve the visual tasks. These networks assume deep layers for extract and acquire information of each visual part from an image. Deep layers allow hundreds of convolutional layers to extract more precious information from the train and test set. This block builds multi-scale hierarchies feature extractor for collecting feature information.

The outlines of the study highlighted below.

(1) The study conduct to improve the feature classification process without increasing the computational cost.

(2) It adopts the feature scaling method with an adaptive learning strategy. That conducts a unique scaling process to scale the informative part of an object.

2. Related Theories

The deep convolutional neural network showed significant performance on object detection tasks [6]-[8]. Several state-of-the-art object detection methods adopt deep convolutional neural networks to improve the feature mapping and selection process for achieving enhance detection performance. Deep layers construct feature hierarchies to increase the feature maps. These layers were constructed with different scales and sizes to cover all possible parts of an image. The framework includes different extraction and classification blocks to collect sufficient train data to conduct significant detection task.

Manuscript received June 6, 2020
Manuscript revised June 20, 2020

Visual classifiers mainly follow three types of methods to improve the detection accuracy. One of the common methods is adopting a different combination of multi-layer feature classifiers. The second on use different feature layers to predict the different scale of objects. The final one just combines the above two methods to enhance the accuracy. ION [10] uses the above first method of different skip pooling layers to extract information and after that, the information classifies by the combined feature classifier [11]. FPN [12] uses the second method to execute the object location through different feature layers. These layers designed with different scales of convolutional filters to localize the perfect location of each object.

2.1 Deep Convolutional Neural Network

Deep convolutional networks [13] trained for extract feature map, and execute the activation values of repeated convolutions. The classifier executes all activation results to localize objects. The activation maps are extracted by utilizing convolutional non-linear operations.
Hence the network structure involves a receptive field for executing activation from the input image. These activation results are preserved in fully-connected layers, and latter these data are executed to predict the objects.

2.2 YOLO

YOLO (You Only Look Once) [14]-[16] is a real-time object detection framework. The network constructed with 24 convolutional layers with two fully connected layers. The network examines the input image by $N \times N$ grid. The image classification process of the YOLO network is shown in Figure 1. Each grid cell responsible for predicting a fixed number of bounding boxes. It examines each information and compared it with each bounding box possessing five values$(x, y, w, h, and\ B)$. The principal concept of the YOLO object detector is to develop a single neural network for real-time object detection. The network inspired by the GoogleNet [17]. The layers of the network constructed with 1×1 convolutional layers accompanied by 3×3 Convolutional layers.

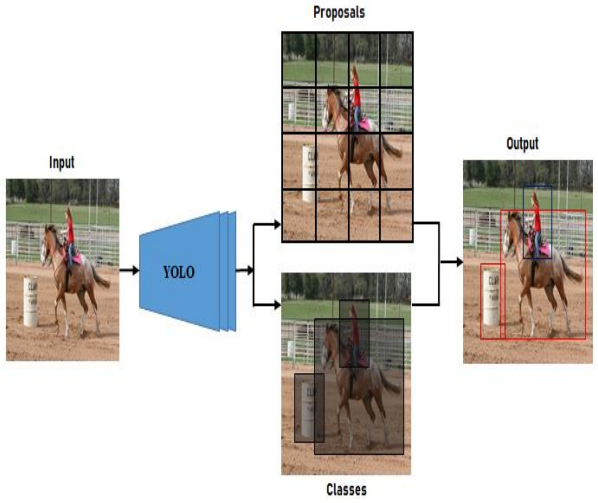


Fig. 1 The image classification process through YOLO object detector.

3. Proposed Algorithm

3.1 Adaptive Feature Scaling Network (AFSNet)

The Adaptive Feature Scaling Network (AFSNet) aims to extract scale-invariant feature data to improve feature quality. It inhabits multi-level feature layers with different scale variations to generate high-level feature data. The proposed method forms as like HyperNet [18], and Feature Pyramid Network (FPN) [13]. It concatenates different features with a variety of different scales to identify the exact object location. The scaler network [19]-[20] employed a variety of convolutional filters with different frame sizes to capture all possible areas of an image. Furthermore, AFSNet involves class-independent methods to generate object hypotheses. The class-independent blocks differentiate those outputs (object hypotheses) by multi-scale filters and collets the feature proposals. Feature classifiers configured to map the image and generate feature proposals. After that, the next block examines all mapped features to locate the exact objects. These proposals are examined by the color and shape of each object. The architecture of AFSNet shown in Fig. 2.

In the initial stage, it explores the process to exploit encoder to classify spatial information. This information transferred to the classification block for localizing the objects. That gives a boost to a bias in the classifier performance to localize the perfect central location of an object is in the input image. The higher the corresponding class score matches to conduct the process. The architecture of the network genetically produces a receptive field for each activation on the source image. The visual area fits

progressively more inferior as it goes deeper into the layers towards the fully connected layers.

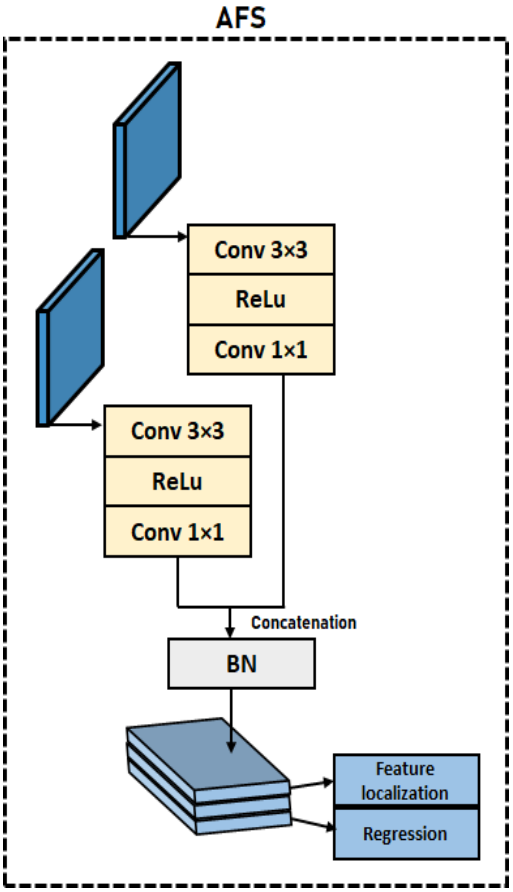


Fig. 2 The architecture of AFSNet.

3.2 Scaling Informative Part

The proposed method aims to extract valuable initial feature maps. The framework examines and map fragments and address the information to classifier. To train the network, a large number of candidates are extracted from the training set. The proposed network considers candidates in the formation of rectangular fragments with multiple sizes and positions to cover every class from an input image. The network examines all possible candidates and selects only the informative part from an image. After that, it examines mapped candidates to scale the informative portion. These scaled feature data latter use to train the network and also help the classifier to classify objects. The process of adaptive feature scaling shown in Fig. 3. The adaptive feature scaler network map and scale feature data through the following equation

$$f_{(i,j)} = c + \sum_{(k,l)\in N(i,j)} \omega_{k,l} \cdot (x - x')_{k,l} \qquad \text{(Eq. 1)}$$

$$S = f_l \sum_{i=1}^{W} \sum_{j=1}^{H} \left(\frac{1}{\omega \times A}\right) \qquad \text{(Eq. 2)}$$

In Eq. 1, and Eq. 2, $x_{i,j}$ denotes the pixel value at the feature location (i,j), $N_{(i,j)}$ carries the neighbors value of (i,j) using the 3×3 convolutional kernel, $\omega_{k,l}$, and c are the convolutional coefficient.

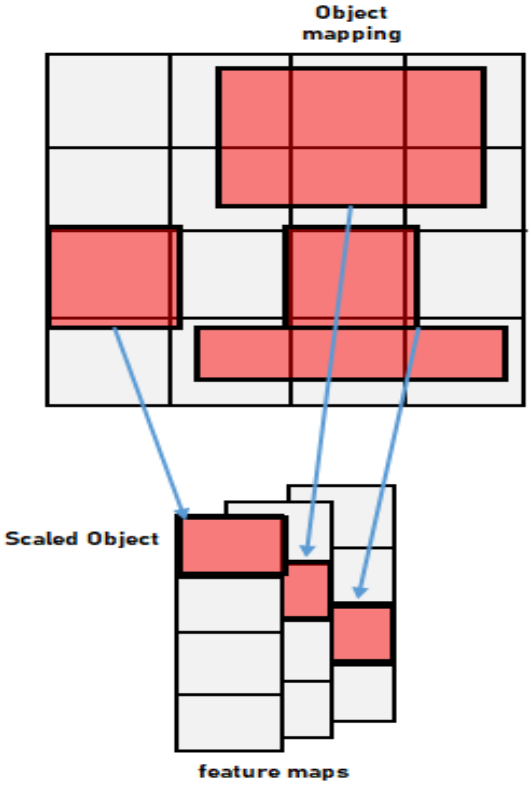


Fig. 3 The process of adaptive feature scaling.

3.2 Network Architecture

The proposed ASFNet introduce to improve the feature extraction process. The network architecture implemented along with YOLO convolutional network. YOLO network struggles to generalize small objects. The network deals with the unusual aspect ratio to localize small objects. That is why the appearance of the framework is not remarkable. For this reason, we combined the proposed network with the YOLO network. The Adaptive Feature Scaling network (AFSNet) considers the fundamental concepts of convolutional filters to forming the perfect convolutional blocks for examining feature information from the test image. The convolutional blocks of the proposed network constructed with admitting the perfect kernel size. The convolutional kernels conduct crucial tasks to examine the visual information.

The top-layers of the proposed network contains with initial feature examiners. The prosed network shown in Fig. 2. The primary feature blocks set with a fixed resolution to extract high-level feature information. After that, the initial feature data passes to the AFSNet for detail classification. AFSNet aims two feature levels of the base network. The first input of AFSNet connected with con5 and the second one is connected with conv7. This input block of AFSNet confirms important features detailed over appear repeatedly in the images comprising objects to be localized and seldom in non-class images. The architecture of AFSNet with the base network shown in Fig. 4. The network architecture constructed with light convolutional filters and activation layers. The network first examines the input feature data and generates proposals. Then it classifies all proposal data to examine the exact feature location. After confirming the feature location it divides the feature data and maps only the visual parts form the input image. Similarly, the network examines useful sub-features appear frequently in the areas containing visual patterns or non-objects elsewhere.

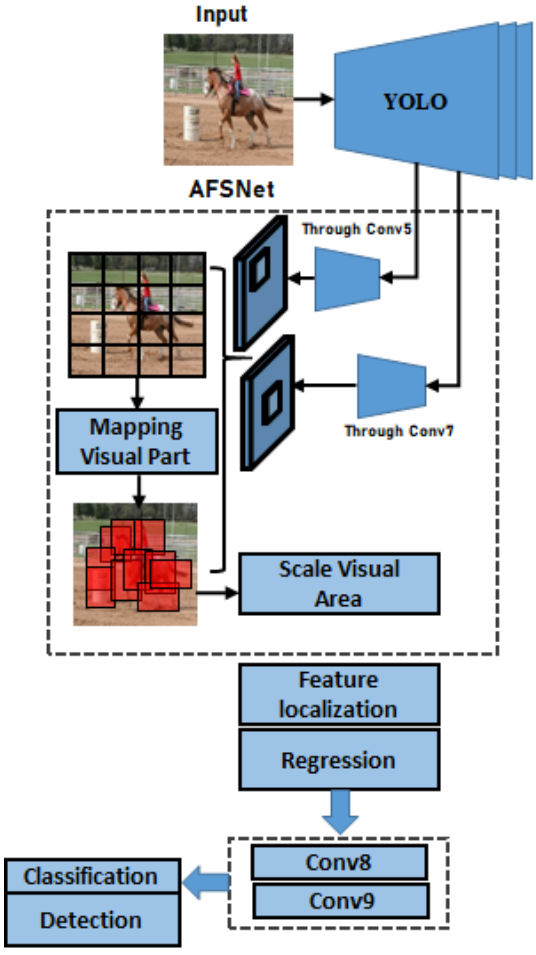


Fig. 4 The process architecture of AFSNet with base network.

To classify these features, it constructs specific fragment f a set of confident examples. Moreover, image regions including the part f, and non-visual fragments are avoided as much as possible. Through the positive features, the network identifies only visual areas and scale the visual parts. The scaled visual information passes to the next convolutional blocks to classify the information for the further convolutional task. The feature information of AFSNet classifies with pre-detection blocks of conv8 and conv9 to object localization and detection tasks.

The localization and detection blocks composed of 7×7 and 1×1 convolutional filters. These layers ensure object detection by comparing the feature proposals and classes. To localize the object the network accompanies the following equation:

$$IOU = p_\tau\left(\frac{class}{object}\right) * p_\tau(object) * C_{pred}^{truth} \quad \text{(Eq. 3)}$$

In Eq. 3, the bounding boxes are predicts by examining each grid cell and also considers the class probability C.

4. Experiments

In this section illustrated the experimental evaluation of the proposed network. The network is evaluated with the multi-scale and single-scale examples. The evolution environments are detailed discussed in this section.

In the multi-scale classification, the network managed adaptive feature learning to locate the object area and scale the best window including each object class in the test images. The feature scaler and extractor considered a fixed ratio between the image area including the height and width of all examples of the object class. Moreover, each scaled window containing an object and it simply represented by its width. The criterion of the scaled feature is extended in the declaration of (i', j'), it expresses the area of the window corresponding to the true position with width W, also (i, j) represents the location of the window corresponding to the position of output feature through the detector. Then for the area of (i, j, ω), to be evaluated as a correct detection we require it to perform the feature scaling task.

4.1 Experimental Environment

To improve the feature extraction and detection performance of the proposed network AFSNet coupled with the YOLO network. The adaptive feature scaling convolutional network utilized simple convolutional kernels and activation layers to map and scale the feature information. The experimental parameters are illustrated in Table 1.

Table 1: Experimental Environment

Experimental Parameters	***Set Value***
Input size	448×448
Batch size	32
Weight decay	0.0005
Learning rate	0.001

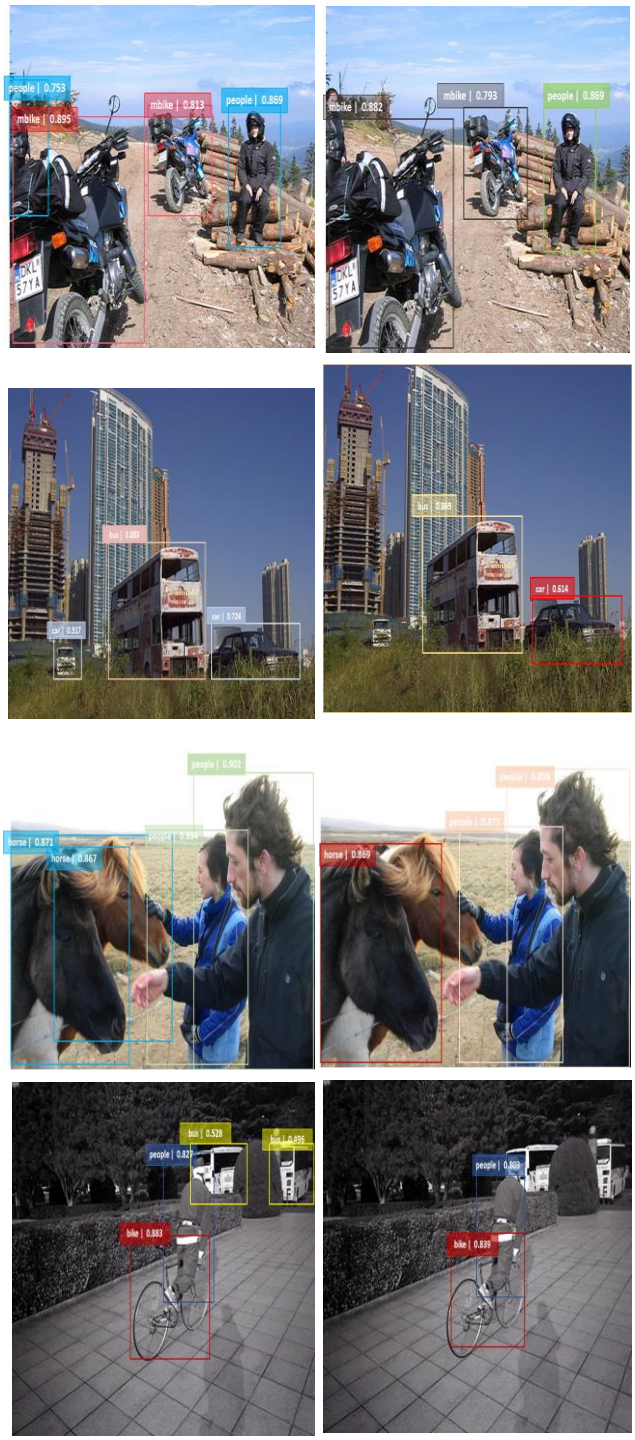
The network train with PASCAL VOC 2012 [21] dataset and the test conducts with the same dataset. The initial learning rate of the network set to 0.001 with the iterations 10000 and 20,000. The momentum sets at 0.9 with the attenuation coefficient of 0.0005, and each batch contains 32 images. Moreover, to enhance the training samples it rotates the image, changes the image saturation, exposure, and hue.

5. Results

The proposed network adopts the simple unique learning method that uses an adaptive feature scaling process. The scaling method introduced to the network to improve the feature quality that improved the training process of the network along with the detection accuracy. Furthermore, scaling only informative parts of an image proves significant to reduce the complexity of the network. As well as the network takes less time to localization and classification tasks that significant compared with conventional YOLO networks. The evaluation results of the proposed network showed in Table 2, Table 3, and Table 4. The network evaluated with a set of multi-scale test images. The multi-scale evaluation results showed in Fig. 5, and in Fig. 6 illustrates the performance comparison of the proposed network with robust object detection frameworks. The test set generated by considering the low-resolution image and small image that can prove the proposed network actual performance to scale all possible objects with satisfactory detection accuracy.

6. Conclusions

The proposed network adopts the simple unique learning method that uses an adaptive feature scaling process. To summarize, the proposed approach introduced adaptive learning to scale feature data from the image. The AFSNet examines part-based representation to localize visual information. It extracts distinctive object information and construct a set of samples of learning materials to solve the object detection tasks. Moreover, spatial relations observed to locate the visual parts. The feature scaling algorithm is employed to automatically learn a classifier that differentiates between visual parts and non-visual parts. The AFSNet achieved enormous detection accuracy of 84.1 mAP with 36 frames per second (fps).



(i)Proposed Network (ii) YOLO

Fig. 5 Detection results of the proposed network and YOLO object detector.

Table 2: Evaluation results on PASCAL VOC dataset.

Method	*Backbone*	*Evaluation Matrics*		
		Precession	*Recall*	*mAP*
SSD300 [22]	VGG-16	6.6	25.9	41.4
ION [10]	VGG-16	6.4	24.1	38.3
Faster R-CNN [23]	VGG-16	4.1	20.4	35.6
AFS Convolutional Neural Network	YOLO	24.9	28.3	55.7

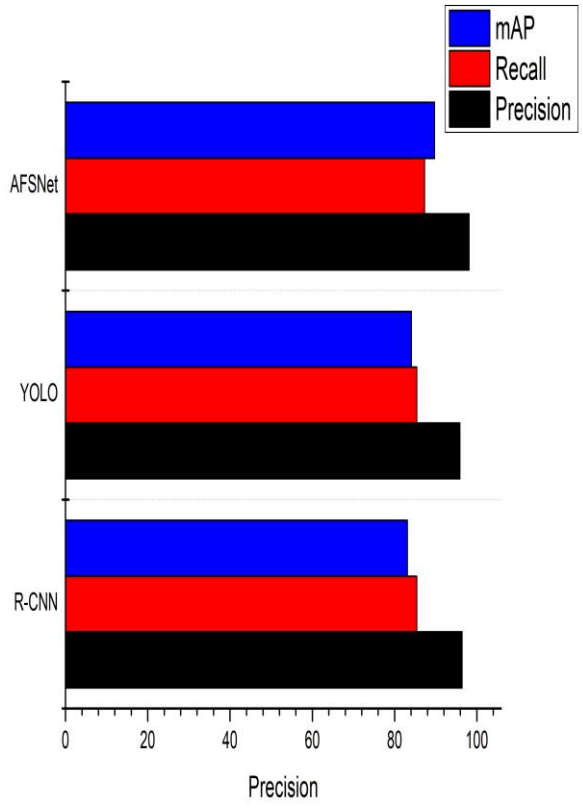


Fig. 6 The average performance comparison with the state-of-the-art methods.

Table 3: The average precision comparison with the robust object detection framework.

Method	*Backbone*	AP_S^{bbox}	AP_M^{bbox}	AP_L^{bbox}
AFSNet	YOLO	24.9	28.3	55.7
YOLO[15]	VGG-16	6.4	24.1	38.3
Faster R-CNN [23]	VGG-16	4.1	20.4	35.6
ION [22]	VGG-16	24.9	28.3	55.7

Acknowledgments

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

References

[1] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", international Journal of Computer Vision, vol. 60, no. 2, pp. 91-100, 2004.
[2] C. Nebauer, "Evaluation of Convolutional Neural Networks for Visual Recognition", IEEE Transactions on Neural Networks, vol. 9, no. 4, pp. 685-696, 1998.
[3] W. Liu et al., "SSD: Single Shot Multibox detector", ECCV, pp. 21-37, 2016.
[4] J. Redmon, A. Farhadi, " YOLO9000: Better, Faster, Stronger", arXiv preprint arXiv:1612.08242, 2016.
[5] K. Simonayan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", arXiv preprint arXiv:1409.1556, 2014.
[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", CVPR, pp. 770-778, 2016.
[7] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via regionbased fully convolutional networks," in Proc. Adv. Neural Inf. Process.Syst., Barcelona, Spain, 2016, pp. 379–387.
[8] R. B. Girshick, "Fast R-CNN," ICCV, pp. 1440–1448, Dec. 2015.
[9] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scaledeep convolutional neural network for fast object detection", European Conference on Computer Vision, Amsterdam, pp. 354–370, 2016.
[10] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," CVPR, pp. 2874–2883, 2016.
[11] Z. Jie, et al., " Scale-AwarePixel-Wise Object Proposal Network," IEEE Transaction on Image Processing, vol. 25, no. 10, pp. 4525-4539, 2016.
[12] T.Y. Lim et al., " Feature Pyramid Networks for Object Detection", IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117-2125, 2017.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks", NIPS, pp. 1097-1105, 2012.
[14] J. redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-time Object Detection", IEEE Conference on Computer Vision and Pattern Recognition, pp. 779-788, 2016.
[15] M. F. Haque, H. Y. Lim, and D. S. Kang, "Real Time Object Detection Based on YOLO with Feature Filter Bank", The Journal of Korean Institute of Information Technology, vol. 17, no. 5, pp. 91-97, 2019.
[16] J. Redmon, and A. Farhadi, "YOLOv3: An Incremental Improvement", arXiv preprint arXiv:1804.02767, 2018.
[17] C. Szegedy, et al., "Going Deeper with Convolutions," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1-9, 2015.
[18] T. Kong, et al., "HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection", IEEE Conference on Computer Vision and Pattern Recognition, pp. 845-853, 2016.
[19] M. F. Haque, and D. S. Kang, "Adaptive Feature Scaler Convolutional Network for Object Detection", In Proceedings of KIIT Conference, pp. 241-244, 2019.
[20] Z. Bhutto, et at., "Scaling of Color Fusion in Stitching Image", International Journal of Computer Science and Network Security, Vol. 19, No. 4, pp. 61-64, April 2019.
[21] M. Everingham, et al., "The pascal visual object classes challenge: a retrospective", International Journal of Computer Vision, Vol. 111, No. 1, pp. 98-136, Jan. 2015.
[22] W. Liu, et al., " SSD: Single Shot Miltibox Detector," In Europian Conference on Computer Vision, pp. 21-37, 2016.
[23] S. Ren, K. He, R. Girshick, and J. sun, "Faster R-CNN: Towards Real-Time Object Detection with region proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, June, 2017.

Md Foysal Haque received M.S. degrees in Electronic Engineering from Dong-A University in 2020. Currently studying Ph.D. in Electronic Engineering at Dong-A University, Busan, Korea from 2020. Research interests in Digital Image Processing and Pattern recognition.

Dae-Seong Kang, He received a B.S. degree from Kyungpook National University, Daegu, South-Korea, in 1984. M.S. degree and Ph.D. in Electrical Engineering from Texas A&M University, USA, in 1991 and 1994, respectively. He is currently Professor at the Department of Electronic Engineering, Dong-A University, Busan, Korea. His research interests in Image processing and compression.