IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 15, 2022, accepted April 9, 2022, date of publication April 12, 2022, date of current version April 15, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3166928

Pyramid Attention Upsampling Module for Object Detection

HYEOKJIN PARK[1] AND JOONKI PAIK[1,2], (Senior Member, IEEE)
[1]Department of Image, Chung-Ang University, Seoul 06974, South Korea
[2]Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Joonki Paik (paikj@cau.ac.kr)

This work was supported in part by the ICT Research and Development Program of MSIP/IITP (2014-0-00077, development of global multitarget tracking and event prediction techniques based on real-time large-scale video analysis), and in part by the Korea Government (MSIT) through the Artificial Intelligence Graduate School Program of Chung-Ang University under Grant 2021-0-01341.

ABSTRACT The core task of object detection is to extract features of various sizes by hierarchically stacking multi-scale feature maps. However, it is not easy to decide whether we should transmit semantic information to the low layers while reducing the loss of semantic information of the high-level features. In this paper, we present a novel method to reduce the loss of semantic information, and at the same time to improve the object detection performance by using the attention mechanism on the high-level layer of the feature pyramid network. The proposed method focuses on the sparse spatial information using deformable convolution v2 (DCNv2) on the lateral connection in the feature pyramid network. Specifically, the upsampling process is divided into two branches. The first one pays attention to the global context information of high-level features, and the other rescales the feature map by interpolation. Finally, by multiplying the results from the two branches, we can obtain upsampling result that pays attention to semantic information of the high-level layer. The proposed pyramid attention upsampling module has three contributions. First, It can be easily applied to any models using feature pyramid network. Second, it is possible to reduce losses in semantic information of the high-level feature map by performing context attention of the high-level layer. Third, it improves the detection performance by stacking layers up to the low layer. We used MS-COCO 2017 detection dataset to evaluate the performance of the proposed method. Experimental results show that the proposed method provided better detection performance comparing with existing feature pyramid network-base methods.

INDEX TERMS Object detection, feature pyramid network, attention mechanism, deep learning.

I. INTRODUCTION

Object detection is a fundamental but still challenging problem in computer vision. Deep learning-based object detection methods are widely used in various fields such as face detection [1], [2], object tracking [3], [4], pedestrian detection [5], [6], autonomous driving, and medical imaging, to name a few. Convolutional neural networks (CNNs) have been rapidly developed in recent years, starting with Alexnet [7], and have shown significantly improved performance in the field of object detection. Since learning a backbone network from scratch requires large amounts of data and processing time, a pre-trained backbone with a large dataset such as ImageNet is widely used for efficient feature extraction [8]. To realize a lightweight backbone, MobileNetV2 [9] used inverted residuals and linear bottleneck, ShuffleNet [10] used point-wise group convolution and channel shuffle, and Xception [11] applied depth-wise separable convolution to the inception model method that separates channels and spaces. On the other hand, to increase increasing the accuracy by constructing a deeper network, ResNet [12] used a deeper neural network through residual learning, and ResNeXt [13] improved performance by increasing cardinality using grouped convolution based on ResNet. To detect multi-scale objects, the image pyramid [14] method used a multi-scale feature map from high to low layers at the cost of increased computational load and memory space. To solve this problem, research on effectively integrating the feature maps of the high and low layers is being actively conducted.

Figure 1 shows a representative model used for object detection with the structure of the feature pyramid network (FPN) [15]. The deep layers in the backbone extracts

The associate editor coordinating the review of this manuscript and approving it for publication was Sotirios Goudos.

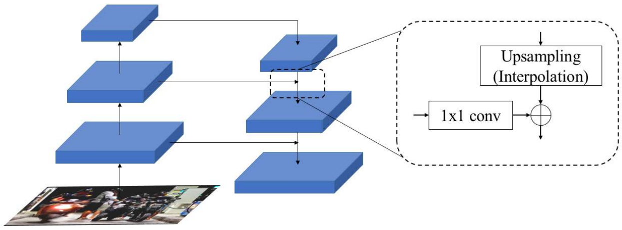

FIGURE 1. Feature pyramid network (FPN).

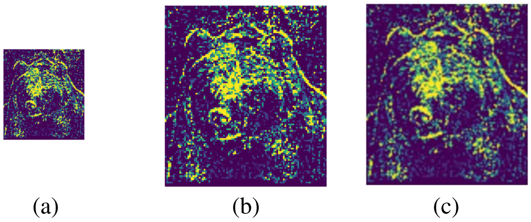

FIGURE 2. (a) Input image, (b) nearset-neghbor interpolation result, (c) bilinear interpolation result.

high-level features such as textures or parts of objects, whereas the shallow layers extracts low-level features such as edges and curves. The FPN uses both high and low-level features effectively as shown in Figure 1 to hierarchically stack feature maps through top-down pathway with lateral connection to detect objects of various sizes. Through this method, the performance of detecting multi-scale objects was improved, and computational cost and memory problems were solved. PANet [16] proposed a stronger FPN structure by adding a bottom-up path, NAS-FPN [17] proposed a new FPN structure through neural architecture search, and BiFPN [18] proposed a fusion method for multi-scale feature maps by building a weighted bi-directional feature pyramid network.

Existing FPN-based structure uses interpolation [19], [20] or deconvolution [21], [22] to fuse high and low-level feature maps. More specifically, interpolation mainly uses nearest-neighbor or bilinear interpolation. Nearest-neighbor interpolation is calculated quickly by filling the value using pixel values at the nearest location, but as shown in Figure 2b, the edge becomes ambiguous due to the occurrence of aliasing. Bilinear interpolation uses values obtained by multiplying the neighboring four pixel values by weights when the size of the original image is resized by N times, and blur occurs as shown in Figure 2c.

Deconvolution expands the size of the feature map by inversely calculating the convolution, and checker board artifact occurs due to even overlap phenomenon according to stride and keel size.

To solve this problem, Zhao *et al.* proposed a pyramid pooling module (PPM) that applied different size pooling [23], and Chen *et al.* proposed atrous spatial pyramid pooling (ASPP) by superimposing a dilated convolution applied with various rates on an atrous pooling layer [24]. However, PPM has a problem in that information on the pixel location is lost in the process of applying pooling to various sizes to the feature map. In addition, since ASPP uses dilated convolution, a wide receptive field can be utilized at the cost of losing local context information.

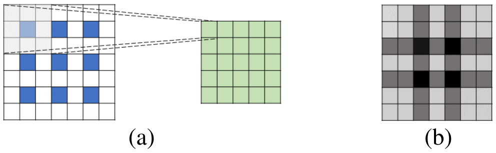

FIGURE 3. (a) Deconvolution, (b) even overlab.

In this paper, we proposed an upsampling method by reducing the loss of semantic information of input images of higher-level feature maps in the FPN structure and attention to local and global context information. The proposed method consists of three steps.

1) We used deformable convolution v2 [25] to the feature map generated in each step to the backbone to generate more prominent context information, and then used lateral connection to reduce semantic information loss to the higher-level feature map.
2) By applying global average pooling to the feature map, we generated through deformable convolution and extracted high-level features using the attention map of global contextual information.
3) In other branches, the high-level feature map is interpolated, and the global context information is multiplied by the attention map to obtain the upsampled feature map.

II. RELATED WORK

A. OBJECT DETECTION

Object detection is usually classified into one-stage and two-stage approaches. The two-stage methods showed high accuracy by sequentially applying regional proposals. Regions with CNN features (R-CNN) is the first two-stage method that uses CNNs to propose candidate regions, and then extracts features from each region [26]. There are many improved variants of R-CNN including, fast R-CNN [27], faster-RCNN [28], Mask R-CNN [29], Cascade R-CNN [30], and Libra R-CNN [31], to name a few. On the other hand, one-stage detection methods are faster than two-stage methods at the cost of lower accuracy. One-stage methods include YOLO [32], SSD [33], RetinaNet [34], and EfficientDet [18]. In addition to one- and two-stage methods, keypoint-based methods, such as CenterNet [35], CornerNet [36] and center-based ATSS [37], FCOS [38], used anchor-free approach. Recently, many object detection methods eliminate non-maximum suppression (NMS) or anchor generation processes using transformer [39] significantly improved detection performance [40], [41].

B. FEATURE PYRAMID NETWORK

Multi-scale object detection is a crucial problem in the object detection field. The image pyramid was a commonly used method in object detection to increase the accuracy of multi-scale object detection [14]. However, independent extraction of features at each level from the image pyramid requires a redundant computational cost. To reduce the computational load and improve the detection accuracy, feature pyramid network (FPN) compensates for the semantic information lost in the forward process using a top-down pathway and lateral connection [15]. Recently, more robust pyramidal structure based on FPN has been proposed. PANet added a bottom-up path augmentation to deliver low-level information to the high-level once more, enabling more precise localization [16]. Natural architecture research-feature pyramid network (NAS-FPN) utilized natural architecture research (NAS) to design a new feature pyramid neural network structure and to improve performance at the cost of increased memory space [17]. Tan *et al.* proposed a weighted bi-directional feature pyramid network (BiFPN) that solves the computational cost problem and fuses multi-scale feature maps more quickly and effectively [18].

C. ATTENTION MECHANISM

Attention mechanism have been actively studied as essential elements for many neural networks in natural language processing. Recently, attention mechanism in computer vision have been used in many fields to improve performance by focusing on regions of interest in images and capturing long-range dependencies. Its application to computer vision includes image classification [42]–[44], image generation [45], [46], and segmentation [47], [48]. Attention mechanism for object detection has been proven through previous studies since it can help to locate and recognize objects in images and improve detection performance [49], [50]. Li *et al.* proposed a MAD unit find neuron activation in high and low streams through aggressive search [49], Zhu *et al.* proposed a new structure of couplenet that integrates global and local information of objects to enhance detection performance [50]. Figure 4 shows the structure of transformer [39]. Recently, transformer-based object detection methods were proposed and provided improved detection accuracy [40], [51].

III. PROPOSED METHOD

A. IMPROVED LATERAL CONNECTIONS

The proposed method can be applied to the original feature pyramid network using attention upsampling module as shown in Figure 5.

Recently, the deep learning-based object detection method uses the FPN to build a robust model against scale change by hierarchically stacking feature map of each stage using lateral connections and top-down paths. The feature map of each stage extracted through a backbone network has detailed characteristics of the object. FPN fixes the number of channels to 256 through point-wise conversion on the feature map of each step extracted from the backbone and proceeds to the decoder stage. The decoder combines a high-level feature map with abundant category information and a low-level feature map to provide a stronger model for multi-scale objects. However, in the process of channel reduction using point-wise convolution to reduce the amount of computation, semantic information about the object of each stage feature map is lost.

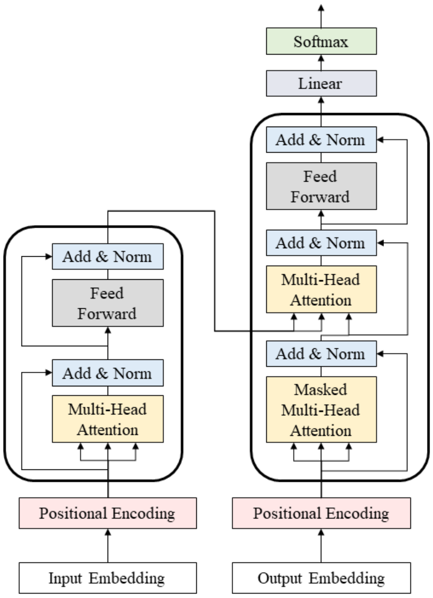

FIGURE 4. Transformer architecture.

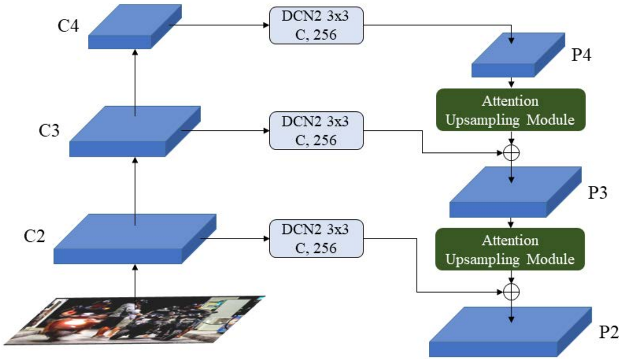

FIGURE 5. Overall architecture of proposed method.

B. ATTENTION UPSAMPLING MODULE

The proposed method changes the number of channels to 256 by using the deformable convolution V2 (DCNv2) when passing the feature map from each stage to the lateral connection to solve the above problem. The DCNv2 is expressed as

$$y(p) = \sum_{k=1}^{k} w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k, \tag{1}$$

where p_k and w_k respectively represent the offset and weight for the k-th position, and $y(p)$ and $x(p)$ respectively represent the characteristics of the position p in the output and input feature maps. Δp_k and Δm_k are the learnable offset and

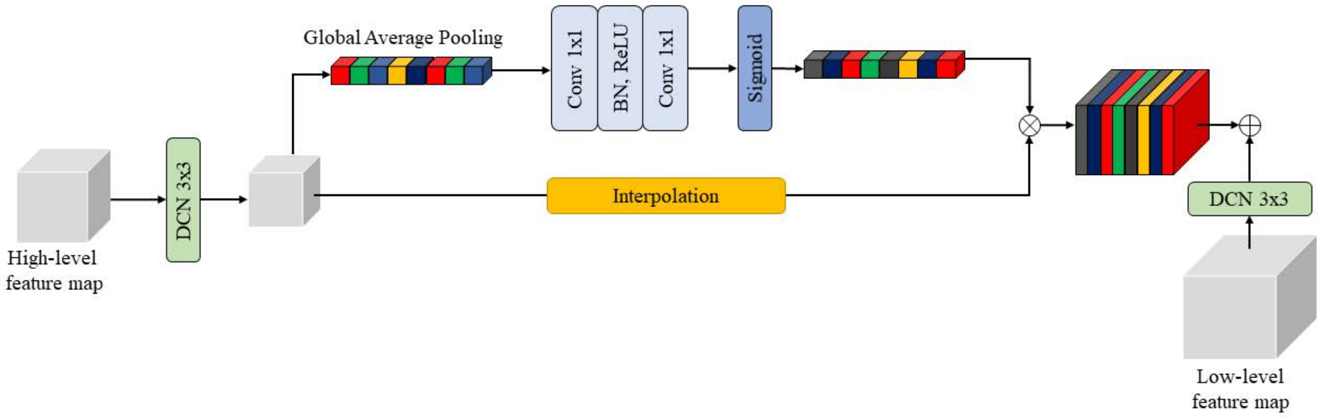

FIGURE 6. Attention upsampling module.

modulation scalar for the k-th position. DCNv2 improves the deformable convolution by finely adjusting the spatial support region with a modulation scalar in the range of [0, 1] for an offset with a real number of unrestricted ranges.

DCNv2 uses a flexible kernel that can be easily changed. In addition, the receptive field can be finely adjusted through the modulation scalar. Therefore, DCNv2 can focus more powerfully and efficiently on sparse spatial location than fixed kernels. When applied to the lateral connection, the representation ability of features generated in each stage can be improved, and object detection performance can be improved by extracting high-level features as a feature map that is more robust to geometric deformation.

In this section, we propose the attention upsampling method by dividing the upsampling process into two branches to focus on semantic information by applying global average pooling to the high layer feature map and to reduce information loss that occurs during the upsampling process as shown in Figure 6.

In general, FPN is used in object detection to extract multi-scale features. FPN upsamples the high-level feature map and fuses it with the low-level feature maps. Usually, interpolation is used to resize the feature map, but the interpolation method has a problem in that semantic information is lost due to aliasing and blur. To solve that the problem, an alternative upsampling method uses the transposed convolution. However, it cannot avoid checkerboard artifacts due to overlapping, which results in performance degradation. To compensate for the information loss that occurs during the upsampling process, the size of the feature map is changed using interpolation with convolution operation. However, above mentioned methods are computationally expensive and use complex convolution structures. Furthermore, because of unidirectional upsampling, the semantic information of the high-level layer is lost due to interpolation and deconvolution.

The proposed method focuses on global context information by applying global average pooling to higher-level feature maps and fusion with low layers to reduce semantic information loss of higher-level feature maps. Average pooling can extract globally important features by summarizing and reflecting spatial information in consideration of the entire image. Global average pooling can be applied to high-level feature maps extracted through the lateral connection to use high-level functions including abundant category information. The proposed attention upsampling module sends the feature map of the high layer to two branches. The first branch extracts global context information at $1 \times 1 \times C$ through global average pooling and aggregates information through 1×1 convolution followed by batch normalization and ReLU. Another branch resizes the feature map by applying nearest-neighbor interpolation to the higher-level feature map. By multiplying the attention map of global context information from each brunch and the upsampling feature map, the semantic information of the high-level feature map can be focused to obtain the upsampled result. Finally, a modified pyramid network is constructed by the elemental-wise sum of the attention feature map of the high layer and the feature map of the low layer.

IV. EXPERIMENTS

A. EXPERIMENT SETUP

MS-COCO 2017 dataset was used for evaluation. The dataset consists of 118K training images and 5K verification images. MMDetection implemented in pytorch was used for the experiment [52]. The input training images are maintained at a ratio and the size of the width and height is adjusted to 1333 and 800. The initial learning rate is 0.005, and for the stability of learning during the initial 500 iterate, the warm-up ratio is 0.01, and 8 epochs and 11 epochs are multiplied by 0.1. The optimizer used SGD, weight-decay is set to 0.0001, momentum is 0.9, and batch size is set to 4. We used a single GPU, 2080ti.

B. EVALUATION METHOD

For performance evaluation of the proposed method, precision, recall, and average precision were used. Precision and

TABLE 1. Comparison with state-of-the-art object detectors (minival).

Method	Backbone	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
FCOS [38]	Resnet50-FPN	36.6%	56.0%	38.8%	21.0%	40.6%	47.0%
PANet [16]	Resnet50-FPN	37.5%	58.6%	40.8%	21.5%	41.0%	48.6%
Faster R-CNN [15]	Resnet50-FPN	37.4%	58.1%	40.4%	21.2%	41.0%	48.1%
Cascade R-CNN [30]	Resnet50-FPN	40.3%	58.6%	44.0%	22.5%	43.8%	52.9%
Libra R-CNN [31]	Resnet50-FPN	38.3%	59.5%	41.9%	22.1%	42.0%	48.5%
FCOS	**Resnet50-ours**	**37.2%**	**56.3%**	**39.5%**	**20.7%**	**40.6%**	**48.7%**
PANet	**Resnet50-ours**	**39.1%**	**60.2%**	**42.5%**	**22.7%**	**42.5%**	**51.0%**
Faster R-CNN	**Resnet50-ours**	**39.2%**	**60.4%**	**42.7%**	**22.4%**	**42.2%**	**51.4%**
Cascade R-CNN	**Resnet50-ours**	**41.4%**	**60.2%**	**45.3%**	**23.7%**	**44.9%**	**54.5%**
Libra R-CNN	**Resnet50-ours**	**39.2%**	**59.9%**	**43.2%**	**22.4%**	**42.5%**	**51.5%**

TABLE 2. Comparison with state-of-the-art object detectors (test-dev).

Method	Backbone	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
FCOS [38]	Resnet50-FPN	36.9%	56.6%	39.3%	20.6%	39.5%	46.0%
PANet [16]	Resnet50-FPN	38.0%	59.0%	41.3%	22.1%	41.2%	46.9%
Faster R-CNN [15]	Resnet50-FPN	37.7%	58.7%	40.8%	21.7%	40.6%	46.7%
Cascade R-CNN [30]	Resnet50-FPN	40.6%	59.2%	44.0%	23.0%	43.4%	51.1%
Libra R-CNN [31]	Resnet50-FPN	38.6%	60.0%	42.0%	22.4%	41.3%	47.7%
FCOS	**Resnet50-ours**	**37.3%**	**56.6%**	**39.9%**	**20.5%**	**39.6%**	**47.4%**
PANet	**Resnet50-ours**	**39.4%**	**61.0%**	**42.7%**	**22.5%**	**41.8%**	**49.5%**
Faster R-CNN	**Resnet50-ours**	**39.5%**	**61.2%**	**42.9%**	**22.9%**	**41.9%**	**49.6%**
Cascade R-CNN	**Resnet50-ours**	**41.6%**	**60.9%**	**45.2%**	**23.9%**	**44.0%**	**52.7%**
Libra R-CNN	**Resnet50-ours**	**39.8%**	**60.9%**	**43.6%**	**22.3%**	**42.6%**	**50.5%**

recall can be respectively expressed as

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

and

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

where TP means true positive, FN false negative, and FP false positive. In order to calculate the precision and recall, the intersection over union (IoU) is used to set the thresholds of the ground truth box and the prediction box to determine the truth. The equation for IoU can be expressed as.

$$IOU = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})}, \tag{4}$$

where B_p is the predicted box, and B_{gt} is the ground truth box. The mAP of MS-COCO increases by 0.05 per step from IoU $= 0.5$ for 80 classes and is calculated as an average AP up to 0.95. The equation for mAP is expressed as.

$$mAP = \frac{mAP_{0.50} + mAP_{0.55} + \cdots + mAP_{0.95}}{10}, \tag{5}$$

where mAP_{50} and mAP_{75} are calculated as fixed values of $IoU = 0.5$ and $IoU = 0.75$, respectively.

C. RESULTS

The performance the proposed method was compared with FCOS [38], PANet [16], faster R-CNN using FPN [15], Cascade R-CNN [30] and Libra R-CNN [31]. The evaluation uses AP, AP_{50}, AP_{75}, AP_S, AP_M, and AP_L as evaluation criteria for MS-COCO. Table 1, 2 shows the quantitative evaluation performance of mAP, mAP_{50}, and mAP_{75} when the existing method and the proposed method are applied. Experimental results of the proposed method showed the best performance in most cases. In some small, medium, and large evaluations. according to object size, accuracy was not improved, but all of the results were improved in mAP.

Qualitative evaluation was evaluated by comparing images as a result of detecting objects with a score of 0.5 or higher. Figure 7. shows the qualitative evaluation comparing the proposed method and FPN. The figures in the first and second rows compare images containing objects of various sizes. When the proposed method was applied through the test images, even small objects were successfully detected. The figures in the third and fourth rows compared images containing large objects. FPN erroneously detected multiple objects in a single large object. On the other hand, the proposed method successfully detected the large object because feature

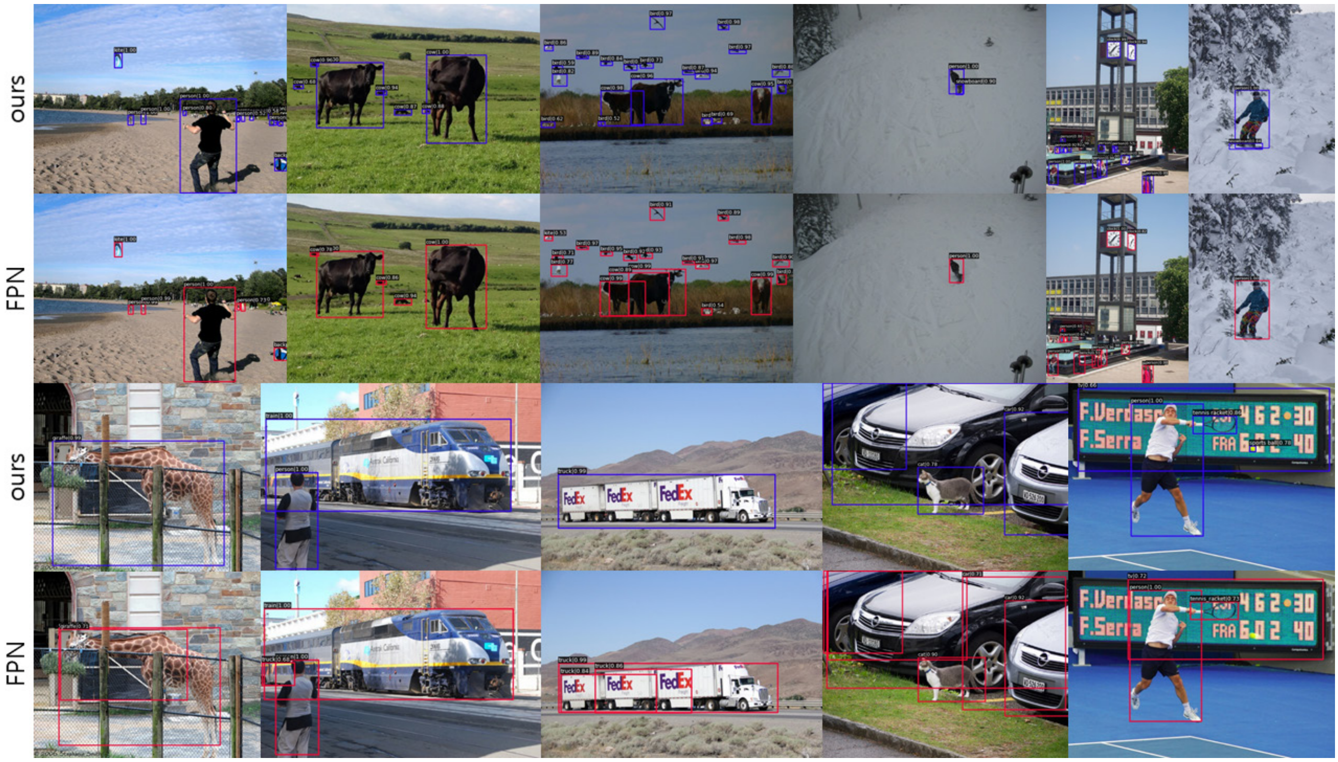

FIGURE 7. Qualitative comparisons of the proposed attention upsampling module with FPN.

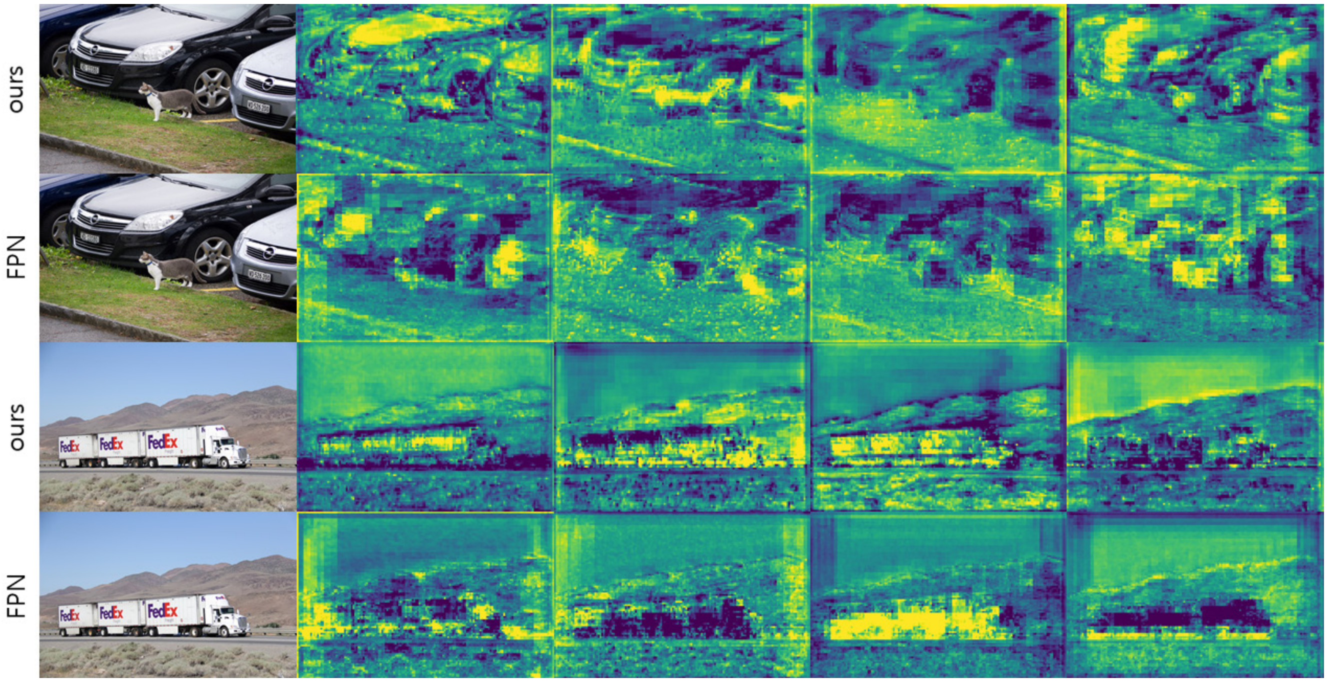

FIGURE 8. Comparison of the proposed attention upsampling module and the pyramid last layer (P2) feature map of FPN.

fusion is performed by paying attention to global context information.

Figure 8 shows the comparison of the proposed attention upsampling module and the pyramid last layer (P2) feature map of FPN. If the FPN uses only nearest-neighbor interpolation, aliasing is unavoidable, which makes context information insufficient to distinguish objects. However, the proposed method shows better results because it improves

TABLE 3. Ablation study of training avg/max pooling with/without.

Max pooling	Average pooling	AP	AP_{50}	AP_{75}
✓	✓	38.5%	60.1%	41.6%
✓		39.0%	60.3%	42.2%
	✓	**39.2%**	**60.4%**	**42.7%**

global context information from the high-level feature map to attention upsampling. The proposed method improves lateral connection and attention upsampling global context information to propose a low-level feature map and fusion method. As a result of the experiment, it compensated for information loss caused by the conventional method and showed better results in object detection of various sizes. Therefore, when using the attention upsampling module proposed in this paper, quantitative evaluation prove that it is a more robust model for multi-scale objects than conventional methods.

D. ABLATION STUDY

The effect of the pooling method on the proposed attention upsampling module was tested. Table 3 shows the performance of the proposed method according to the use of max pooling and average pooling. ResNet-50 was used as backbone, and the performance was the lowest when both max pooling and average pooling were used. Although the performance difference between max pooling and average pooling is small, average pooling shows the better performance.

V. CONCLUSION

In this paper, we proposed a pyramid attention upsampling module for object detection that can be used in feature pyramid-based neural networks. Our method focuses on sparse spatial information by applying DCNv2 to improve the physical connection. In addition, the loss of semantic information that occurs in the process of upsampling high-level feature maps was improved. In the upsampling process, we presented a technique for performing the upsampling process in two branches to reduce the loss of semantic information in the high-level feature map and to attention to the global context information. The proposed method can be easily applied to the object detection model using the FPN structure, and it shows better performance on MS-COCO benchmark by applying it to various object detection models. Since the one-stage method showed a slight better performance compared to that of the two-stage method, further research is needed to improve the attention upsampling module.

REFERENCES

[1] Y. Zhu, H. Cai, S. Zhang, C. Wang, and Y. Xiong, "TinaFace: Strong but simple baseline for face detection," 2020, *arXiv:2011.13183*.

[2] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5203–5212.

[3] H. Hu, B. Ma, J. Shen, H. Sun, L. Shao, and F. Porikli, "Robust object tracking using manifold regularized convolutional neural networks," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 510–521, Feb. 2019.

[4] K. Chen and W. Tao, "Learning linear regression via single-convolutional layer for visual object tracking," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 86–97, Jan. 2019.

[5] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao, "Generalizable pedestrian detection: The elephant in the room," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11328–11337.

[6] Y. Qian, M. Yang, X. Zhao, C. Wang, and B. Wang, "Oriented spatial transformer network for pedestrian detection using fish-eye camera," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 421–431, Feb. 2020.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[10] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[11] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[13] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.

[14] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid methods in image processing," *RCA Eng.*, vol. 29, no. 6, pp. 33–41, 1984.

[15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[16] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[17] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7036–7045.

[18] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10781–10790.

[19] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4203–4212.

[20] S.-W. Kim, H.-K. Kook, J.-Y. Sun, M.-C. Kang, and S.-J. Ko, "Parallel feature pyramid network for object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 234–250.

[21] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*.

[22] S. Woo, S. Hwang, and I. S. Kweon, "StairNet: Top-down semantic aggregation for accurate one shot detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1093–1102.

[23] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[24] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[25] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9308–9316.

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[27] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.

[29] K. He, G. Gkioxari, P. Dollár, and R. Girshick, ''Mask R-CNN,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[30] Z. Cai and N. Vasconcelos, ''Cascade R-CNN: Delving into high quality object detection,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[31] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, ''Libra R-CNN: Towards balanced learning for object detection,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 821–830.

[32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, ''You only look once: Unified, real-time object detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[33] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, ''SSD: Single shot MultiBox detector,'' in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, ''Focal loss for dense object detection,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[35] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, ''CenterNet: Keypoint triplets for object detection,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6569–6578.

[36] H. Law and J. Deng, ''CornerNet: Detecting objects as paired keypoints,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[37] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, ''Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9759–9768.

[38] Z. Tian, C. Shen, H. Chen, and T. He, ''FCOS: Fully convolutional one-stage object detection,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[40] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, ''End-to-end object detection with transformers,'' in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, ''Swin transformer: Hierarchical vision transformer using shifted windows,'' 2021, *arXiv:2103.14030*.

[42] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, ''Training data-efficient image transformers & distillation through attention,'' in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[43] H. Zhao, J. Jia, and V. Koltun, ''Exploring self-attention for image recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10076–10085.

[44] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ''An image is worth 16×16 words: Transformers for image recognition at scale,'' 2020, *arXiv:2010.11929*.

[45] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, ''Generative pretraining from pixels,'' in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1691–1703.

[46] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, ''Image transformer,'' in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4055–4064.

[47] M. Ren and R. S. Zemel, ''End-to-end instance segmentation with recurrent attention,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6656–6664.

[48] H. Li, P. Xiong, J. An, and L. Wang, ''Pyramid attention network for semantic segmentation,'' 2018, *arXiv:1805.10180*.

[49] H. Li, Y. Liu, W. Ouyang, and X. Wang, ''Zoom out-and-in network with map attention decision for region proposal and object detection,'' *Int. J. Comput. Vis.*, vol. 127, no. 3, pp. 225–238, 2019.

[50] Y. Zhu, C. Zhao, H. Guo, J. Wang, X. Zhao, and H. Lu, ''Attention couplenet: Fully convolutional attention coupling network for object detection,'' *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 113–126, Jan. 2019.

[51] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, ''Deformable DETR: Deformable transformers for end-to-end object detection,'' 2020, *arXiv:2010.04159*.

[52] K. Chen *et al.*, ''MMDetection: Open MMLab detection toolbox and benchmark,'' 2019, *arXiv:1906.07155*.

HYEOKJIN PARK was born in Incheon, South Korea, in 1994. He received the bachelor's degree in electric engineering from the Korea National University of Transportation, South Korea, in 2020. Currently, he is pursuing the Master of Science degree in digital imaging engineering with Chung-Ang University. His research interests include object detection, object recognition, artificial intelligence, and unsupervised learning.

JOONKI PAIK (Senior Member, IEEE) was born in Seoul, South Korea, in 1960. He received the B.S. degree in control and instrumentation engineering from Seoul National University, in 1984, and the M.Sc. and Ph.D. degrees in electrical engineering and computer science from Northwestern University, in 1987 and 1990, respectively. From 1990 to 1993, he was with Samsung Electronics, where he designed image stabilization chipsets for consumer camcorders. Since 1993, he has been a member of the Faculty of Chung-Ang University, Seoul, where he is currently a Professor with the Graduate School of Advanced Imaging Science, Multimedia, and Film. From 1999 to 2002, he was a Visiting Professor with the Department of Electrical and Computer Engineering, The University of Tennessee, Knoxville. Since 2005, he has been the Director of the National Research Laboratory in the field of image processing and intelligent systems. From 2005 to 2007, he served as the Dean for the Graduate School of Advanced Imaging Science, Multimedia, and Film. From 2005 to 2007, he was the Director of the Seoul Future Contents Convergence Cluster established by the Seoul Research and Business Development Program. In 2008, he was a full-time Technical Consultant of the System LSI Division of Samsung Electronics, where he developed various computational photographic techniques, including an extended depth of field systems. He has served as a member for the Presidential Advisory Board for Scientific/Technical Policy with the Korean Government and is currently serving as a Technical Consultant for the Korean Supreme Prosecutor's Office for computational forensics. He was a recipient of the Chester-Sall Award from the IEEE Consumer Electronics Society twice, the Academic Award from the Institute of Electronic Engineers of Korea, and the Best Research Professor Award from Chung-Ang University. He has served the Consumer Electronics Society of the IEEE as a member of the Editorial Board, the Vice President for International Affairs, and the Director for Sister and Related Societies Committee.