Sailboat Detection Based on Automated Search Attention Mechanism and Deep Learning Models

Ziyuan Luo, Minh Nguyen, Wei Qi Yan
School of Engineering, Computer and Mathematical Sciences
Auckland University of Technology
Auckland, 1010 New Zealand

***Abstract*—The development of deep learning-based visual object recognition has achived great progress in recent years. Deep learning models based on attention mechanisms are able to further improve the ability to detect the regions of interest (ROI), but creating an appropriate module is a difficult task. Therefore, in this paper, we propose an automated design scheme based on neural architecture search (NAS) to migrate the attention mechanism for visual object detection and obtain better results of sailboat detection by using both public datasets and our own collected datasets. We verify the effectiveness of our proffered method and evaluate the performance compared with other algorithms. The obtained results have demonstrated better robustness in terms of generalization than other deep learning models.**

Keywords—Deep learning, attention, automated machine learning, CNN, visual object detection

I. INTRODUCTION

Auckland has the title of "City of Sails", a great deal of international sailing events were held in this city. For example, one of the most famous sailing competitions in the world, the America's Cup, was held in Auckland, New Zealand. This city has a wealth of beautiful ports, and strong winds blow on the sea. Benefited from this natural and geographical condition, sailing is very deeply developed as an important game. Infected by the sailing culture of this city, we have the idea of combining this ancient sport with deep learning in AI. There are already visual object detection algorithms for sailboats from digital images, but if we implement an algorithm that only detects sailboats, further research work is needed.

Visual object detection means that an algorithm is applied to determine the class label of visual objects, mark the positions on the given image, and annotate them with a bounding box. This is the fundamental work to be accomplished in the field of visual object detection in computer vision, which is also one of important parts in digital video processing. Due to various postures and frequent occlusions of visual objects in digital videos, the irregularity of movements, a depth of field, resolution, lighting, and other conditions, a diversity of forground and background in various scenes, outcomes of visual object detection algorithm will directly affect the effectiveness of object tracking, action recognition, and visual event description. Therefore, with the development of deep learning, the detection is still a very challenging task with a lot of potentials and a large room for improvement. Sailing [1, 2] is a complicated activity, timely and accurate detection using deep learning is extremely needed.

In computer vision, object detection algorithms have gradually shown their importance. The popular algorithms are split into two categories, one is the algorithms of R-CNN (i.e., region-based convolutional neural network) [3] family based on region proposals, the other is one of YOLO (i.e., You Only Look Once) family [4]. In order to accomplish the task of visual object detection, R-CNN takes use of the idea of sliding windows with a region proposal network (RPN). The two-stage algorithm is highly accurate but runs with very slow speed. In response to this shortcoming, YOLO has adopted a one-stage algorithm in which visual object detection is commenced in a single step. However, while the speed of operations is increased, the accuracy rate of these models is dropped.

Since its birth, YOLO has received a series of improvements and demonstrated betterment from YOLOv2 to YOLOv5. This series of models capture visual objects very quickly, acquire the entire information during model training and testing, make use of contextual information, and thus is lessly prone to predicting incorrect object information. Therefore, in this paper, we will follow this series of algorithms and make our improvements based on YOLOv5 to balance the speed and accuracy for sailboat detection.

In recent years, Transformer has become a trendy architecture in deep learning, relying on a simple but powerful mechanism - the attention mechanism - that allows us to focus on parts of the input, therefore, reasoning is much efficiently. DEtection TRansformer (DETR) [5] is the first application of Transformer in the field of visual object detection, enables high-performance detection in the end-to-end way with less prior knowledge. Based on Microsoft COCO [22] dataset, accuracy and efficiency are as same as the highly optimized Faster R-CNN [6]. The results are better than that of Faster R-CNN based on large objects. In addition, unlike most existing detection methods, DETR does not require any custom layers and is therefore easy to be reproduced, including the modules that have been applied to many deep learning frameworks. Numerous results have also shown that combining a transformer with a CNN is able to generate better results.

The algorithm of combining convolutional neural network (CNN) with attention mechanism allows us to focus on the region of interest (ROI), with a significant potential for feature extraction. Nevertheless, how to design a much suitable model for distinct tasks is a problem that requires a plenty of experimentations and continuous improvement. This process is tedious and does not always lead to the best result. The emergence of neural architecture search (NAS) [7] has resolved this problem. By setting a reasonable search strategy, search space, and evaluation strategy, the NAS is able to achieve the design and evaluation of neural network modules and finally get the optimal network model. Therefore, based on the shortcomings in object detection as well as attention mechanisms and automated design solutions, in this paper,

978-1-6654-0645-1/21/$31.00 ©2021 IEEE

we combine them together to achieve optimal performance. Our contributions are as follows:

- With the realistic acquisition, we construct a realistic sailboat detection dataset, which allows us to evaluate the robustness of our model in real-world applications.
- In order to better focus on the ROI region, we propose a CNN model that combines spatial attention mechanisms together.
- Facing with the diversity of data collected by using various devices and the tedious procedure of attention module design, we create the model by automating the search and design which make the model much robust.

In summary, our proposed model combines and improves YOLO family with the attention mechanisms which are validated at present based on real datasets.

The remainder of this paper is structured as follows: The related work is reviewed in Section II, our methods of this paper are depicted in Section III, Section IV is related to our results, our conclusion is drawn in Section V.

II. Related Work

A. YOLO Family

YOLO[8] (i.e., You Only Look Once) has creatively approached visual object detection task as a regression method, combining the candidate regions and object detection phases into one stage. At a glance, it is possible to know which visual objects are in each image and where they are located. However, this model gains time efficiency at the expense of mAP.

In 2016, YOLOv2 [4] was designed based on Fast R-CNN and SSD (i.e., single shot multibox detector) algorithms, by utilizing a number of training improvements for accuracy betterment, applying the new network model Darknet-19 for speed, using joint training for the classification, combining with the WordTree to expand the model to thousands of classes. YOLOv3 [21] summarizes the improvements made to YOLOv2, uses a residual model to further deepen the network structure, an FPN architecture for multiscale detection has been taken into account.

YOLOv4 [9] has developed an efficient and robust model that allows anyone to train a super-fast and accurate target detector by using a 1080Ti or 2080Ti GPU. The impact of a range of state-of-the-art training methods for visual object detector is verified. The state-of-the-art methods make them much efficient and adaptable if it is trained with a single GPU, including CBN, PAN, SAM, and so on. Compared to YOLOv4, YOLOv5 offers faster speed and higher performance consisting of a family of models, including YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x, YOLOv5x+TTA.

B. Attention

The attention-guided context feature pyramid network (AC-FPN) [23] is an attention-oriented pyramidal network of contextual features, which not only increases the field of view but also makes use of contextual information of visual objects to provide better classification by fusing multiple fields of perception features. SENet [10] was proposed with an attention mechanism between multiple channels, where the importance between channels is calculated through two fully connected layers. Thus, unimportant channel values are filtered out.

DETR [5] is the first algorithm of Transformer model to detect visual objects, this algorithm performs better on large objects and presents poor results in detecting tiny objects. Recently, Swin Transformer model has been proposed and applied to the tasks such as visual object detection, image classification, and image segmentation, all is ranked at the first place, which confirms the importance of the attention mechanism.

In addition, NAS has conducted the work in the design of attention mechanism. Liu, et al. [11] proposed two modules, NAS-SAM and NAS-CAM, to explore the weighting outcomes of space and channel by using a synchronous search strategy that allows the attention module to search for multiple frames at various locations in the same network.

C. Neural Architecture Search

In 2016, Baker et al. [12] creatively proposed the use of reinforcement learning to resolve the design and optimization problems of attention models with dataset CIFAR-10 [13], defeated all manually designed models that brought in the new idea for the development of automatic machine learning.

NAS [24] search is able to automatically design a high-performance neural network based on a set of algorithms. The principle is as follows: A set of candidate neural network structures are constructed the search space, a strategy is applied to search for the optimal network structure in the space. In each iteration of the search process, the network structure is gradually optimized till the optimal sub-network is found. This automation method saves the cost of deep neural network design.

This method was gradually applied to visual object detection. Inspired by the one-shot NAS [14], DetNAS [15] decouples weight training and structure search for the detector. NAS-FPN [29] was applied to answer the question of how Neck of the detection network is automatically connected in a hierarchical way with visual features to achieve a tradeoff between accuracy and speed. NAS-FPN then is also employed as a reinforcement learning search method by using RNNs as controllers, which was similar to the general NAS method, the agents are applied to speed up the search. The innovation of AutoFPN [25] is in the Neck and Head networks, where Neck is an auto-fusion operation for backbone features first, while auto-head takes advantage of NAS to search a network for the purpose of object classification and region regression.

In summary, we see that NAS has made remarkable progress in both visual object detection and attention mechanisms, but less progress has been achieved to combine the two together. In this paper, we will commence on our efforts and make further betterment for the efficiency and effectiveness of sailboat detection and recognition.

III. Our Methods

An attention model aids us to achieve much accurate visual detection. The usage of NAS method assists us to reasonably design the optimal net for various datasets and resolve the problems of visual object detection and

recognition. Therefore, in this section, we will elaborate on the methods for the detection of sailboats in this paper.

A. Backbone

YOLOv5 is the latest model in the YOLO series, albeit with slightly weaker performance than YOLOv4, however, it is considerably time efficient [30]. Hence, we have chosen this network as our benchmark backbone in this paper. The proposed network consists of three components:

Backbone: A CNN that aggregates and forms image features at multiple image granularities.

Neck: A series of network layers that mix and combine image features and pass them on to the prediction layer.

Output: Prediction of image features, generating bounding boxes and predicting categories.

Pertaining to YOLOv5, Backbone, Neck, and Output are the same no matter it is the YOLO versions v5s, v5m, v5l or v5x. The only difference is the depth and width parameters of the models.

The focus of our proposed module is on adding in the first layer of backbone. Its principal function is to periodically extract pixels from high-resolution images and reconstruct them into low-resolution images, that means, we stack the four adjacent positions of images, focus on width and height information in channel space, improve the receptive field of each point, and reduce the loss of original information. The design purpose of this module is to reduce the amount of calculation.

The third layer of Backbone, BottleneckCSP module, consists of two main parts: Bottleneck [27] and CSP. SPP module (i.e., spatial pyramid pooling) adopts max pooling with kernels 5, 9, and 13, respectively, then concatenation is taken to improve the receptive field.

Neck (PANet) [26] is based on Mask R-CNN and FPN frameworks with information propagation, the ability to accurately retain spatial information, which facilitates the proper positioning of pixels to generate masks. After the setup, we construct the baseline experimental model in this paper.

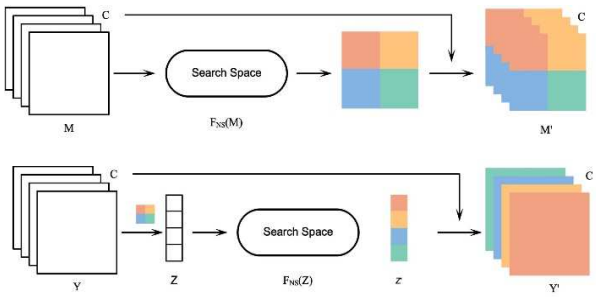


Figure 1. NAS attention modules, the top one is NAS spatial attention module and the bottom one is NAS channel attention module.

B. NAS-Based Spatial and Channel Joint Attention Module

The manually-designed deep neural networks, especially combined across domains, require a wealth of a priori knowledge and extensive experimentation. This is undoubtedly labour-intensive work. Therefore, in this paper, we propose an automated design scheme by using NAS search to accomplish the attentional mechanisms for sailboat detection.

NAS-SCAM consists of two parts, space and channel attention module, which produces a corresponding weighting output.

The architecture of NAS-SAM is shown in Fig. 1. The input feature map is $M = [m_1, m_2, \ldots, m_c]$, which has width W, height H, and channel C. M is transformed into spatial weight map $n \in \mathbb{R}^{H \times W}$ by one or multiple convolutional operations and nonlinear operations $F_{NS(\cdot)}$ in NAS search space. In the search space, n is related to the spatial weighting information. Finally, we are use of the multiplication operation to fuse spatial weight map n into the input feature map M and generate output feature map M' as shown in eq. (1).

$$M' = n \otimes M = [nm_1, nm_2, \ldots, nm_c] \quad (1)$$

Correct choice $F_{NAS}(\cdot)$ is the key operation and has a great impact on the weight effect. However, there are many choices which is difficult to find the best one. Therefore, we recommend choosing $F_{NAS}(\cdot)$ by using the NAS to find suitable framework.

The structure of NAS-CAM is shown in Fig. 1. In order to produce channel weighting information without changing spatial information. Suppose the input eigenmap is $Y = [y_1, y_2, \ldots, y_c]$, $y_i \in \mathbb{R}^{H \times W}$, we take use of a global average pool along spatial dimensions as shown in eq.(2), and generate a vector $z \in \mathbb{R}^{1 \times 1 \times C}$.

$$Z_i = Avgpool(y_i) = \frac{1}{H \times W} \sum_{p=1}^{H} \sum_{q=1}^{W} y_i(p, q) \quad (2)$$

NAS-SAM and NAS-CAM are employed for generating the feature maps, the two output feature maps retain important weights by using max pooling in the fusion. In order to better fit the attention search mechanism for the channels and spaces, a new search space has been created as shown in Table I.

TABLE I. THE OPERATIONS OF NAS-SAM AND NAS-CAM

	NAS-SAM	NAS-CAM
Zero (No connection)	√	√
Conv2D 1	√	√
Conv2D 3	√	√
Conv2D 5	√	√
Conv2D 9		√
Conv2D 15		√
Atrous Conv2D 3	√	√
Atrous Conv2D 5	√	√

Because of the architectures of NAS-SAM and NAS-CAM, we choose different operations between the two nodes in the NAS search space. Therefore, the operations of NAS-SAM and NAS-CAM are shown in Table 1. For NAS-SAM, because we need to extract information from spatial dimensions, we take consideration of 2D convolution with different filter having various sizes to extract information from perception fields. We also take use of dilated convolution to improve distance information. For NAS-SAM model, one-dimensional convolution is applied to extract channel information due to global average pooling. In addition, zero calculation is also performed for

NAS-SAM and NAS-CAM, which indicates that there is no connection between the two nodes. The algorithm based on gradient is stated as follows:

Input: Training dataset

Step 1: Firstly, determine the number of the nodes of the model.

Step 2: Load all operations into the connection path of the node through mixed operations to form a multipath neural network.

Step 3: Apply different weights to each path and solve the discrete optimization problem, update the selection probability and weight of the combined mixed operation at the same time.

Step 4: Select the final network structure according to the probability from the mixed operation. The latter will introduce the probability calculation and function selection of this paper.

Output: The best net structure

In order to better accommodate cross-domain combinations, we therefore relax the discrete operation space [17] in this paper to form a continuous gradient-accessible search space, complete with hyperparametric optimization of structures and operations.

In order to adopt the optimal architecture of attention modules in the same network, we propose a synchronous search strategy to search for each attention module independently. In general, after the architecture of attention module is designed, the modules of the same architecture will be inserted at the end of each upper and lower sampling block [18, 19]. However, due to convolution and pooling operations, feature maps at different locations in the network have distinct semantic meanings. Therefore, searching for different attention module architectures makes it much suitable for different locations in the network, such as upsampling and downsampling blocks.

The synchronous search strategy supports to design a unique attention module for each upsampling and downsampling block, and optimize it separately. Because the attention module architecture is adjusted by optimizing the continuous variable α, which will have different gradients in the optimization process, thus optimizations in different directions will generate a much suitable architecture.

C. Loss Function and Evaluation Function

In this paper, we are use of a combination of losses to better fulfill the task. Firstly, BECLogits loss function was employed to calculate the loss of objectness score, cross-entropy loss function (BCEclsloss) [28] was used for the class probability score, and GIOU Loss [16] was taken for predicting the bounding box. In the combined loss, in order to keep the equilibrium accelerated convergence and have better performance, the weights are set as c_{iou}=0.05, g_{iou}=1.00, and b_{ce}=0.50, respectively.

In order to evaluate the performance of the proposed deep neural network, the object detection probability, false detection probability, which refers to the probability that one object in detection results in a false detection, F1[20] score, precision and reacall are defined as eq.(3~5):

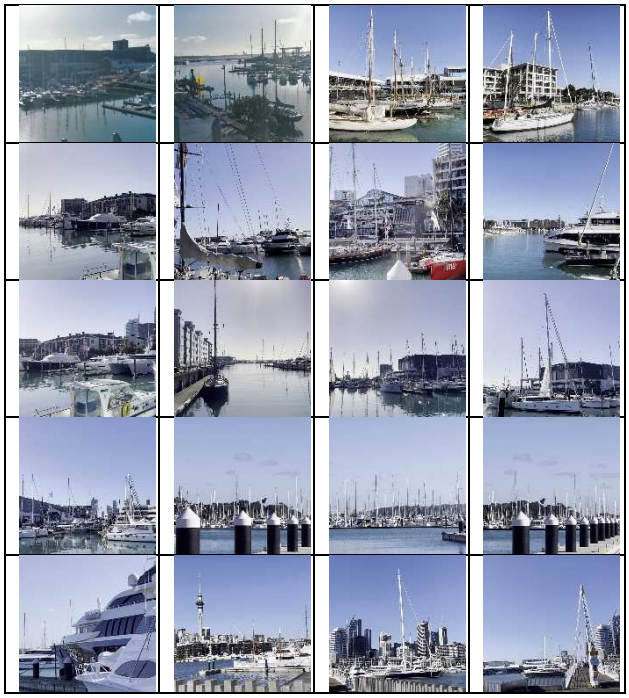
$$P_d = \frac{N_{td}}{N_{ground_truth}} \tag{3}$$

$$P_f = \frac{N_{\mathrm{fd}}}{N_{\mathrm{total\_target}}} \tag{4}$$

$$F1 = 2 \times \frac{P_d \times (1-P_f)}{P_d + (1-P_f)} \tag{5}$$

where N_{td} is the number of true detections, $N_{\mathrm{ground_truth}}$ is the total number of ground truth, N_{fd} is the number of false detections, and $N_{\mathrm{total\_target}}$ is the number of detections in total.

TABLE II. THE IMAGE SAMPLES SHOT AT LOCAL HHARBOUR



IV. OUR RESULTS

A. Our Dataset

In this paper, we firstly construct a real sailing dataset by shooting videos using a mobile camera along seashore. It is used to evaluate the performance of our proposed deep learning models. Secondly, we retrieve and select the America's Cup sailing competition videos in the past three years and split them into frames. In our dataset for training, the images contain 1,484 labeled ships, 348 ships in images were taken into account for testing. For the training dataset, the labelled images are resized to 512 × 512 without overlapping. Patches with ships are sent to the network for model training. The test images are processed in the same way. In order to ensure that the model works fastly and correctly, the images were resized to a uniform size 640×640. Table II shows the image samples of the dataset from our local harbour. Our dataset also includes the video frames from the America's Cup sailing competition as shown in Table III.

B. Our Setup

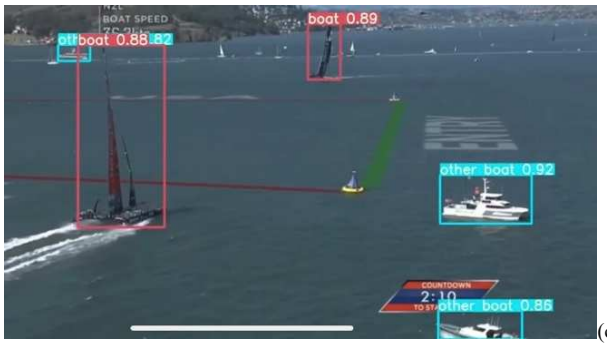
We carried out our experiment by using PyTorch. The baseline of the proposed net is YOLOv5. Regarding NAS search, the process is appended at the end of each downward or upward sampling block.

During the search, the total epoch is 120. The updates of w alternate with updates of the continuous variable α. The learning rate is 1.00×10^{-3} while w is being updated, the rate is 1.00×10^{-2} whilst α is being updated. After trained an optimized architecture, we reconstructed the network based on learning rate α. In the reconstruction process, the total number of elements is 300, the learning rate is 1.00×10^{-3}. We saved the best-performance model after the validation and set it as the best model for test. All experiments were carried out for three times.

TABLE III. THE IMAGE SAMPLES FROM AMERICA'S CUP VIDEOS

(a)

(b)



(c)

Figure 2. The screenshots of demonstration videos (a) Video-1 (b) Video-2 (3) Video-3

In Fig.2, we show various results for sailboat detection in our test. Fig. 2(a) is a video from our harbour, Fig.2 (b) is a video from American's Cup 2021, Fig. 3(c) is a video to detect various sailboats.

C. Experimental Results

In this paper, in oder to better compare the performance of our proposed model on the proposed attention mechanism, we have conducted a horizontal and vertical comparison. Horizontal comparison means that we compared the performance of the same model on the public dataset and the dataset collected by ourselves. On the other hand, we also compared the performance of different models on the same dataset, which is a vertical comparison. Thus, we searched the model in the public dataset. In order to verify the effectiveness of our proposed NAS-SCAM and synchronous search strategy, we compare the performance of the proposed models.

TABLE IV. COMPARISONS OF DETECTION PERFORMANCE ON A PUBLIC DATASET.

Model	$\mathbf{P_d}$(%)	P_f(%)	$F1$
Baseline	71	18.5	0.75
NAS-SAM + Baseline	72.49	19.30	0.77
NAS-SAM + Baseline	73.50	20.50	0.78
NAS-SCAM + Baseline	72.50	18.60	0.76

TABLE V. COMPARIONS OF OBJECT DETCTION PERFORMANCE BASED ON OUR DATASET

Model	$\mathbf{P_d}$(%)	P_f(%)	$F1$
Baseline	67.00	17.00	0.69
NAS-SAM + Baseline	68.49	18.30	0.72
NAS-SAM + Baseline	66.00	16.50	0.70
NAS-SCAM + Baseline	71.00	19.60	0.77

In Table IV, we see that NAS-SCAM is able to achieve better performance than the baseline model without any attention. Furthermore, we notice that the channel-based attention mechanism has a better performance than the space-based attention mechanism. In order to further confirm the performance of our model in terms of generalization, we tested the attention-based model in our collection of real-world images and compared it as follows.

It is easy to see from Table V that the model obtained from the neural architecture-based search has better robustness in terms of generalization than the typical CNN model. In the comparative experiment, because other parameters are kept unchanged, we conclude from the comparison that the performance of the proposed model that removes the neural architecture search mechanism has decreased. This confirms the effectiveness of the method proposed in this paper.

V. DISCUSSION AND CONCLUSION

This paper presents a YOLO algorithm for saiboat detection that introduces an attention mechanism for automated search design: NAS-SCAM-YOLO, which has a lower cost for improving detection accuracy and efficient network design. By using the distinctive features of sailboat images that distinguish them from other ships, it is able to effectively identify sailboats regardless of whether there are other ships in the scene. Our main idea is to select the

extracted feature vectors and optimize the connection in the attention model whilst maintaining the fast prediction of the regression detection algorithm so that the whole network is able to better filter out the feature vectors for subsequent detection. Meanwhile, the attention mechanism involved in NAS-SCAM-YOLO and the improved feature fusion method are migrated to other feature extraction networks.

The next step of this work will be to investigate in-depth visualization of the attention mechanism and present a more intuitive representation of the vsual features within the model [31, 32].

REFERENCES

[1] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, "A SAR dataset of ship detection for deep learning under complex backgrounds," Remote Sensing, vol. 11, no. 7, pp. 765, 2019.

[2] P. W. Vachon, J. Campbell, C. Bjerkelund, F. Dobson, and M. Rey, "Ship detection by the RADARSAT SAR: Validation of detection model predictions," Canadian Journal of Remote Sensing, vol. 23, no. 1, pp. 48-59, 1997.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580-587.

[4] J. Redmon, A. Farhadi, "YOLO9000: Better, faster, stronger," in IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7263-7271.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in European Conference on Computer Vision: Springer, 2020, pp. 213-229.

[6] Y. Xu, G. Yu, Y. Wang, X. Wu, and Y. Ma, "Car detection from low-altitude UAV imagery with Faster R-CNN," Journal of Advanced Transportation, 2017, pp. 1-10.

[7] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," The Journal of Machine Learning Research, vol. 20, no. 1, pp.1997, 2017.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in IEEE Conference on Computer Vision & Pattern Recognition, 2016.

[9] A. Bochkovskiy, C. Y. Wang, and H. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv:2004.10934.

[10] H. Jie, S. Li, S. Gang, and S. Albanie, "Squeeze-and-excitation networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, no. 99, 2017.

[11] Z. Liu, H. Wang, S. Zhang, G. Wang, and J. Qi, "NAS-SCAM: Neural architecture search-based spatial and channel joint attention module for nuclei semantic segmentation and classification," in MICCAI 2020, pp. 263-272.

[12] B. Baker, O. Gupta, N. Naik, and R. Raskar, "Designing neural network architectures using reinforcement learning," arXiv:1611.02167.

[13] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Handbook of Systemic Autoimmune Diseases, vol. 1, no. 4, 2009.

[14] H. Pham, M. Y. Guan, B. Zoph, Q. V. Le, and J. Dean, "Efficient neural architecture search via parameter sharing," in International Conference on Machine Learning, pp. 4095-4104, 2018.

[15] Y. Chen, T. Yang, X. Zhang, G. Meng, X. Xiao, and J. Sun, "DetNAS: Backbone search for object detection," arXiv:1903.10979.

[16] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[17] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-Net: Going deeper with nested U-structure for salient object detection," Pattern Recognition, vol. 106, p. 107404, 2020.

[18] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in European Conference on Computer Vision, 2018.

[19] A. G. Roy, N. Nav Ab , and C. Wachinger, "Concurrent spatial and channel squeeze & excitation in fully convolutional networks," in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2018.

[20] N. Wawrzyniak, T. Hyla, and A. Popik, "Vessel detection and tracking method based on video surveillance," Sensors, vol. 19, no. 23, pp. 5230, 2019.

[21] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv: 1804.02767, 2018.

[22] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in European Conference on Computer, pp. 740-755, 2014.

[23] J. Cao, Q. Chen, J. Guo, and R. Shi, "Attention-guided context feature pyramid network for object detection," arXiv: 2005.11475.

[24] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L. J. Li, F. F. Li, A. Yuille, J. Huang, and K. Murphy, " Progressive neural architecture search," in European Conference on Computer Vision (ECCV), pp. 19-34, 2018.

[25] H. Xu, L. Yao, W. Zhang, X. Liang, and Z. Li, "Auto-fpn: Automatic network architecture adaptation for object detection beyond classification," in IEEE/CVF International Conference on Computer Vision, pp. 6649-6658, 2019.

[26] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in IEEE/CVF International Conference on Computer Vision, pp. 9197-9206, 2019.

[27] J. Park, S. Woo, J. Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," arXiv: 1807.06514, 2018.

[28] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," IEEE Access, vol. 8, pp. 4806-4813, 2020.

[29] G. Ghiasi, T. Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7036-7045, 2019.

[30] K. Liu, H. Tang, S. He, Q. Yu, Y. Xiong, N. Wang, "Performance validation of YOLO variants for object detection," in International Conference on Bioinformatics and Intelligent Computing, pp. 239-243, 2021.

[31] W. Yan. Computational Methods for Deep Learning Theoretic, Practice and Applications. Springer, 2021.

[32] W. Yan. Introduction to Intelligent Surveillance Surveillance Data Capture, Transmission, and Analytics. Springer, 2021.