Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 11, No. 3, 2024

Enhanced Small Target Detection Methodology via Optimized YOLOv8 Framework

Yingjun Zhao [a], Nelson C. Rodelas *

Graduate School, University of the East, Manila, 1008, Metro Manila, Philippines

*** Corresponding author:** Nelson C. Rodelas (Email: nelson.rodelas@ue.edu.ph), [a] liu.xiaolong@ue.edu.ph

Abstract: In the grand palace of computer vision, object detection as a bright pearl, its research and practice value in automation, efficiency improvement and scientific and technological progress occupy a pivotal position. This paper takes YOLOv8 model as the cornerstone of research, and conducts in-depth optimization and exploration on its loss function, network structure and feature extraction mechanism, aiming at significantly improving the model's ability to identify small targets. Specific optimization measures are as follows: (1) The integration of a deformable convolutional module within the YOLOv8 backbone network represents a significant advancement. This strategic modification allows the model to adaptively tailor its receptive field in response to the unique attributes of small targets. Consequently, the model is endowed with the capability to concentrate more precisely on these targets, thereby substantially enhancing detection accuracy. (2) the incorporation of an attention mechanism into the neck structure of the model serves as a sophisticated enhancement. This mechanism functions akin to a discerning filter, adept at extracting salient features from a vast array of information. (3) this research introduces a groundbreaking method for calculating the Intersection over Union (IoU) loss function, termed HIoU. This innovative approach dynamically modulates the weights of the loss function components throughout the training process. The result is a more precise alignment of small targets with their corresponding ground truth bounding boxes, leading to a marked enhancement in the detection performance of small targets.

Keywords: YOLOv8 Algorithm; Small Target Detection; Attention Mechanism; HIoU.

1. Introduction

It is necessary to review the development of traditional target detection methods before discussing modern target detection techniques in depth. These methods were mainly concentrated before 2012, and at their core relied on hand-crafted feature extraction techniques. For example, Haar feature (Viola et al., 2001), local binary model (Ojala et al., 2002) and directional gradient histogram (Dalal et al., 2005) were all widely used feature representation methods at that time. Viola Jones detector (Viola et al., 2004), an algorithm proposed by P. Viola and M. Jones in 2001, is an important breakthrough in the field of real-time face detection. By using integral images, AdaBoost algorithm (Freund et al., 1997) and cascade detection mechanism, the algorithm significantly improves the detection speed and accuracy. Despite its high computational requirements, the Viola Jones detector did set an important milestone in the development of face detection technology, enabling the real-time detection of faces for the first time. DPM (Deformable Part Model) (Felzenszwalb et al., 2008) algorithm as a further evolution of HOG detector was proposed by P. Felzenszwalb in 2008. R. Girshick has deeply optimized DPM (Felzenszwalb et al., 2009, 2010) to further expand its influence in the field of object detection. In the training stage, the DPM algorithm extracts features by dividing the object into multiple subblocks, and in the inference verification stage, it detects the object by recombining these subblocks. This process not only improves the flexibility of detection, but also enhances the ability of the model to adapt to the shape change of the object.

Since 2012, with the rise of convolutional neural networks (CNNs), deep learning technology has experienced unprecedented rapid development and gradually dominated the field of object detection. This period saw the birth of a series of iconic convolutional neural network models, including AlexNet(Krizhevsky et al., 2017), VGG(Simonyan et al., 2014), GoogLeNet(Szegedy et al., 2015; Ioffe et al., 2015; Szegedy et al., 2016, 2017), and ResNet (He et al., 2016), which received widespread attention for their superior performance and innovative design. Convolutional neural networks can effectively extract deep features from images through their unique hierarchical structure, which makes them ideal for object detection tasks. In 2014, Ross Girshick and colleagues (Girshick et al., 2014) proposed R-CNN (Regions with Convolutional Neural Networks) algorithm, marking a major breakthrough in object detection technology. This algorithm is the first to apply convolutional neural network to regional feature extraction, which greatly improves the accuracy and efficiency of target detection. Since R-CNN, the field of target detection has experienced explosive growth, and numerous innovative algorithms have sprung up, driving the rapid development of the entire field. These algorithms not only improve the accuracy of detection, but also optimize the processing speed and the overall performance of the system, which lays a solid foundation for the practical application of target detection technology.

2. Principle of YOLOv8 Algorithm

The YOLOv8 algorithm, a groundbreaking advancement in object detection technology, has been meticulously developed by Glenn Jocher. This innovative system builds upon the foundational principles established by its predecessors, the YOLOv3 and YOLOv5 algorithms, while introducing several significant enhancements.

(1) Data preprocessing. The data preprocessing strategy of YOLOv5 was still adopted for YOLOv8. In the training process of the model, we mainly use four kinds of data

enhancement techniques to improve the generalization ability and robustness of the model. These techniques include: Mosaic enhancement, a method of combining multiple images to create new training samples that significantly increase data diversity; Mixup enhancement, which generates new training instances by proportionally mixing two randomly selected images, helps the model learn smoother decision boundaries; Random Perspective enhancement, which simulates images from different perspectives, enhances the model's adaptability to perspective changes; And HSV enhancement, which adjusts the Hue, Saturation, and Value of the image to enhance the model's ability to recognize color changes. The comprehensive application of these enhancement methods can effectively improve the performance of the model in complex environment.

(2) Backbone network structure. The backbone network structure of YOLOv8 can be briefly seen from YOLOv5. The architecture law of the backbone network of YOLOv5 is very clear. In general, each layer of 3×3 convolution with step size of 2 is used to downsample the feature graph, and a C3 module is attached to further strengthen the features Scale according to the size of the model. In YOLOv8, this feature is generally inherited, and the original C3 modules are replaced by the new C2f module, and the C2f module adds more branches to enrich the tributaries of gradient backpass. The following shows the C2f module of YOLOv8 and the C3 module of YOLOv5.

(3) The YOLOv8 model continues to employ the FPN (Feature Pyramid Network) combined with PAN (Path Aggregation Network) architecture to construct its feature pyramid, ensuring the comprehensive integration of multi-scale information. This approach is pivotal for enhancing the detection accuracy across various object sizes. Notably, the modification in YOLOv8 involves the replacement of the C3 module within the FPN-PAN structure with a more advanced C2f module. This change is designed to optimize the feature extraction process, while the overall architecture remains largely consistent with that of YOLOv5, thereby maintaining a familiar framework for performance improvements.

(4) The evolution of the detection head in the YOLO series from YOLOv3 to YOLOv5 was marked by the use of a "Coupled Head" approach, where a single convolutional layer was responsible for both classification and localization tasks. This approach was altered with the introduction of YOLOX, which pioneered the "Decoupled Head" in the YOLO series. Following this trend, YOLOv8 adopts a decoupled head structure, where two separate branches are dedicated to extracting category and position features respectively. These features are then processed through a 1x1 convolutional layer to finalize the classification and localization tasks. This decoupling enhances the model's ability to focus on specific tasks, potentially improving the accuracy and efficiency of the detection process.

(5) Unlike YOLOv5, which employed a candidate box strategy that was dataset-dependent and could lead to inaccuracies if the dataset did not sufficiently represent the data distribution, YOLOv8 adopts a different approach. It focuses on the multi-scale allocation of positive and negative samples for matching, a critical aspect for robust object detection. YOLOv8 utilizes the TOOD strategy, similar to YOLOv6, which is a dynamic label allocation method. This strategy is more adaptive and effective in handling the complexities of object detection in diverse scenarios. In terms of loss functions, YOLOv8 simplifies the process by only considering target bboxes and target scores, omitting the prediction of object presence. The loss function of YOLOv8 is composed of two main components: the classification loss, which is the VFLLoss (Varifocal Loss), and the regression loss, which combines CIoU Loss and DFL Loss, ensuring a balanced focus on both classification accuracy and localization precision.

3. Object Detection based on Deformable Convolution

3.1. Introduction to Deformable Convolution

In the object detection scene, the object to be detected may present a variety of different shapes and sizes, so it is a very challenging task in the field of object detection to recognize different shapes and sizes of the same object. At present, the traditional strategies to solve this problem are generally divided into two categories :(1) by applying a large number of different data extensions to the input image to enhance the model's learning ability of such data. However, there is a problem with this approach, that is, it is very complicated to manually design an expansion strategy, and complex changes in some scenarios are difficult to encode into an expansion strategy. (2) Design some algorithms with invariance, such as the displacement invariance of CNN. However, the complexity of this approach lies mainly in the complexity of the design and learning process of the convolutional kernel.

The Deformable Convolutional Network (DCN) series of algorithms have been developed to bolster the model's capacity to acquire invariance to complex objects. In the DCN framework, the convolution kernel transcends the conventional rectangular form, evolving to possess an optimal structure that varies across different stages, feature maps, and even individual pixel points. Consequently, DCN imparts an offset learning mechanism to each point within the convolution kernel, enabling the kernel to adapt its structure in response to diverse data inputs. A visual representation of the DCN's operational mechanism is provided in Figure 1.

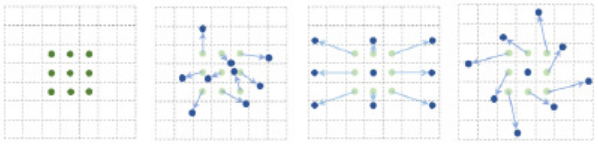

Figure 1. DCN diagram

Deformable convolution emerges as a pivotal advancement in the realm of convolutional neural networks, wherein the traditional convolution process is enhanced by the introduction of learned offsets derived from the input feature map itself. This mechanism entails the convolution operation generating a series of offsets (x, y) for each pixel within the input feature map. Remarkably, these offsets, once predicted, are uniformly applied across the various channels of the identical feature map, thereby introducing a degree of spatial adaptability that is conspicuously absent in conventional convolutional techniques. The intricacies of this innovative module are visually elucidated in Figure 2, offering a clear representation of its operational dynamics.

3.2. Improvement Method

In the object detection task, the bounding box describes the position of the target in each stage of the target detector. Although bounding boxes are easy to calculate, they only

provide a rough positioning of the target and cannot fully fit the shape and attitude of the target. As a result, the features extracted from the regular cells of the bounding box can be heavily influenced by invalid information in the background content or foreground area. This may lead to a reduction in feature quality, which affects the classification performance of target detection.

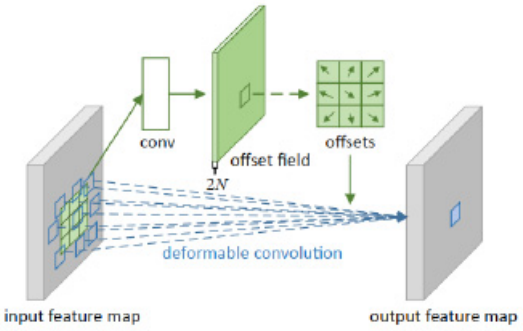

Figure 2. DCN module structure

The purpose of adding deformable convolution is to improve the adaptability of the convolutional neural network to irregularly shaped target objects and broaden its receptive field. Traditional convolution operations can only extract features from regular rectangular receptive fields, while in real scenes, many target objects have irregular shapes, such as faces, cars, animals, etc. In this case, deformable convolution is needed to identify these irregularly shaped target objects. Deformable convolution can adaptively adjust the shape and size of the receptive field according to the irregular shape of the target object, thus improving the robustness of the CNN.

In this study, the integration of the Deformable Convolutional module (DCN block) into the backbone network of YOLOv8s is proposed, primarily for the following reasons: (1) Traditional convolutional neural networks employ fixed-weight kernels, resulting in a uniform receptive field size across different regions of an image. This limitation is particularly problematic when dealing with objects of varying scales or deformations, as these objects may require different receptive field sizes on the feature maps. Therefore, an adaptive adjustment of the receptive field is essential for the network. (2) Deformable convolution offers a more precise alignment with the size and shape of objects during sampling, enhancing robustness, a feature that standard convolution fails to provide. (3) The detection of small targets, which often exhibit small sizes and irregular shapes, is significantly compromised if conventional convolution is employed. This can lead to missed detections and a subsequent decrease in model performance.

In the YOLOv8 architecture, the backbone network serves as the critical component for extracting features from the input image. This feature extraction process is crucial for a comprehensive understanding and detailed description of the image. Typically, the backbone network consists of multiple convolutional layers, each designed to extract features at varying levels of granularity. The integration of these diverse features through fusion and aggregation techniques leads to a more robust and accurate representation of the image, thereby enhancing the overall efficacy of the model.

Given these considerations, this paper advocates for the incorporation of a deformable convolutional module into the backbone network of YOLOv8s. This enhancement allows for the extraction of increasingly detailed image features, setting the stage for more effective feature fusion and prediction processes in the detection head.

The modified network architecture, post-integration of deformable convolutional modules, is depicted in Figure 3. Specifically, the penultimate and third C2f modules within the backbone network are substituted with deformable convolutional modules. This alteration specifically targets the convolution corresponding to the P3 and P4 detection layers, aiming to bolster the model's sensitivity to small and medium-sized targets. The adaptive change in the receptive field for small targets, facilitated by the deformable convolution module, enables the model to more effectively adjust the regression parameters of the prediction box. This enhancement not only increases the model's focus on small targets but also significantly improves the overall performance of the model.

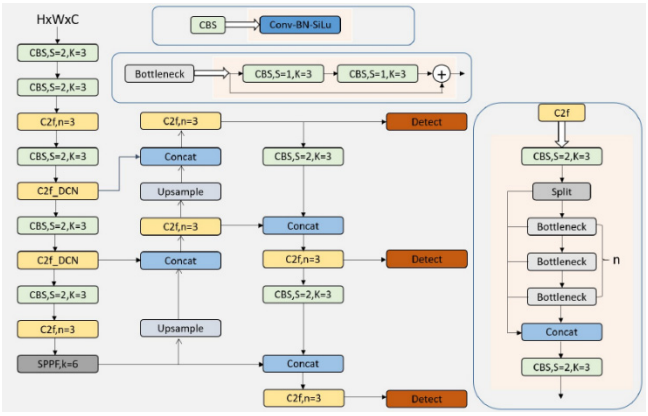

Figure 3. YOLOv8s+DCN network architecture

4. Object Detection based on Attention Mechanism

4.1. Introduction to Attention Mechanism

The attention mechanism serves as a pivotal method in enhancing the learning capabilities of network models by discerning the significance of input data. This technique involves the strategic allocation of varying weights to disparate segments of the input, thereby directing the model's focus towards crucial information. Such a targeted approach not only bolsters the model's performance and accuracy but also mitigates the risk of overfitting, thereby augmenting the model's robustness. Typically, within the realm of neural networks, the implementation of the attention mechanism is facilitated through a dedicated network component, which can be seamlessly integrated into a block structure. It is noteworthy that the attention mechanism lacks a rigid mathematical definition; however, it can be conceptually aligned with traditional techniques such as local image feature extraction and the sliding window method, which can be construed as rudimentary forms of attention.

The attention mechanism can be categorized into three distinct types: (1) Channel attention mechanism, which evaluates the significance of channels by generating and scoring a channel-wise mask, exemplified by the SENet channel attention module. (2) Spatial attention mechanism, which involves the creation and scoring of a spatial mask, as demonstrated by the SAM spatial attention module. (3) Mixed domain attention mechanism, which synergistically considers both channel and spatial attention, with BAM and CBAM attention modules serving as prime examples.

SENet employs an auxiliary neural network to gauge the

importance of each feature channel within the feature map, subsequently assigning a weight to each channel based on its perceived importance. This approach ensures that the model prioritizes significant channels. Nonetheless, SENet's reliance on global averaging pooling inadvertently overlooks substantial spatial information. In contrast, the CBAM module adeptly merges the spatial and channel modules, employing both global average and maximum pooling strategies to curtail information loss effectively. By simultaneously attending to both channel and spatial information, the CBAM module achieves superior results.

The architectural design of the CBAM module is illustrated in Figure 4, delineating its dual-module composition: the initial module focuses on channels, followed by the spatial module, with both modules arranged serially. This configuration underscores the CBAM module's capacity to integrate and leverage both channel and spatial attention, thereby optimizing the model's performance and robustness.

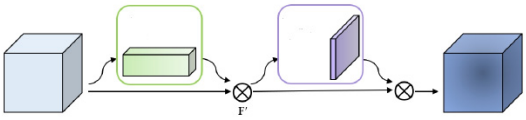

Figure 4. CBAM network structure

4.2. Improvement Method

In this study, the integration of the Convolutional Block Attention Module (CBAM) into the neck portion of the YOLOv8s architecture is justified by several compelling reasons:

(1) Enhanced Feature Representation: The CBAM module is adept at dynamically adjusting both channel and spatial weights within the feature maps, thereby facilitating the effective capture and representation of salient features within the image data. This adaptive capability significantly boosts the model's ability to discern and emphasize critical visual elements.

(2) Superior Computational Efficiency: Relative to other attention mechanisms, CBAM demonstrates a notably reduced computational footprint, resulting in a more efficient processing of features. This efficiency is particularly beneficial in maintaining the real-time processing requirements of the YOLOv8s model without compromising on performance.

(3) Versatility Across Diverse Tasks: The CBAM module is inherently versatile, capable of being seamlessly integrated into a wide array of visual recognition tasks. This adaptability is crucial for the YOLOv8s network, where the neck region serves as a pivotal link between the backbone network and the output prediction head.

The neck region of the YOLOv8s network is structurally unique, facilitating the comprehensive integration of features across various scales. This integration is fundamental to the subsequent prediction stages, and thus, the architectural design of the neck region substantially influences the overall algorithm performance.

The revised network architecture, incorporating the CBAM module, is depicted in Figure 5. Specifically, the CBAM module is strategically positioned after the upsample structure in the up-sampling phase of the PAN-FPN, following each C2f module in the down-sampling phase, and prior to the convolution operations of the CBS module. This placement ensures that feature attention is enhanced just before feature fusion, allowing the optimized model to concentrate more effectively on small target information. By doing so, the model not only increases its focus on the target but also significantly improves the accuracy of both recognition and localization of small targets.

This strategic integration of the CBAM module within the YOLOv8s architecture not only bolsters the model's feature processing capabilities but also enhances its overall performance in object detection tasks, particularly in scenarios requiring the detection of small or subtle targets.

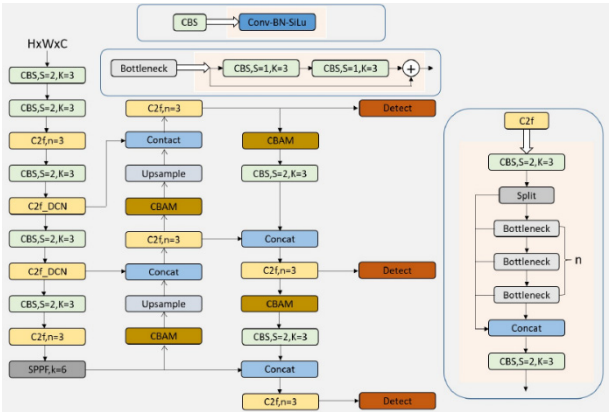

Figure 5. YOLOv8s+DCN+CBAM network structure

5. Object Detection based on Improved IoU

5.1. Improved IoU

In the realm of object detection, the Intersection over Union (IoU) metric serves as a critical tool for assessing the congruence between the predictions generated by a model and the actual annotations of objects. This metric is derived through a mathematical process that involves determining the intersection and union of the predicted regions against the ground truth labels. The IoU is then quantified by dividing the area of intersection by the area of union, resulting in a value that ranges from 0 to 1. A higher IoU value indicates a greater degree of similarity between the model's predictions and the true annotations. When the IoU surpasses a predetermined threshold, it is deemed that the predicted bounding box aligns satisfactorily with the actual object. Moreover, during the training phase, IoU is frequently integrated into the loss function, playing a pivotal role in guiding the network to refine its performance in object detection and segmentation tasks. This integration ensures that the model not only identifies objects but also delineates their boundaries with precision, thereby enhancing the overall accuracy and reliability of the detection system.

5.2. Improvement Method

Given the presence of substandard instances within the training dataset, the application of geometric metrics, including distance and aspect ratio, tends to disproportionately penalize these low-quality examples. This disproportionate penalty can detrimentally affect the model's ability to generalize. To mitigate this issue, it is imperative that the loss function be designed to lessen the impact of these geometric measures when the predicted bounding box aligns closely with the target box. By doing so, the loss function can minimize undue interference during the training process, thereby enhancing the model's overall generalization capabilities.

In view of WIoU's idea of dynamically processing IoU loss

function, this paper proposes a new calculation criterion for IoU loss function, named HIoU and defined as:

$$L_{\mathrm{HIoU}}=R_{\mathrm{W\ IoU}}\times L_{\mathrm{IoU}}+\frac{1}{2}\left(\frac{\left(x-x_{gt}\right)^2+\left(y-y_{gt}\right)^2}{W_g^2+H_g^2}+\alpha v)\right)$$

In this paper, we introduce the novel HIoU metric, which stands out by dynamically adjusting the bounding box regression loss, akin to the WISE-IOU approach. During the initial phases of model training, when the Intersection over Union (IoU) between the predicted bounding box and the ground truth annotation is notably low, the model's primary focus should be on refining the candidate frame to enhance this IoU. The HIoU mechanism, through its initial RWIoU×LIoU component, effectively magnifies the model's penalty for low IoU values, thereby encouraging more accurate initial predictions.

As the training progresses and the IoU between the predicted and actual object boxes stabilizes at a high level, the initial RW in the RWIoU×LIoU component diminishes, allowing the model to shift its attention to refining the center point and the aspect ratio of the candidate box. This strategic shift enables the model to more precisely align the predicted box with the ground truth, thereby enhancing overall model performance.

Furthermore, HIoU innovatively refines the intersection and merge operations inherent in traditional IoU calculations. It strategically reduces penalties associated with geometric metrics such as distance and aspect ratio, ensuring that multiple factors—including IoU, position, size, and shape—between the predicted and actual boxes are more comprehensively considered. This holistic approach significantly boosts the accuracy of target detection.

By reducing the influence of the second and third loss functions by half, HIoU subtly modulates the model's response to geometric measures without overly influencing the broader training process. This delicate balance ensures that the model retains strong generalization capabilities.

In summary, HIoU demonstrates superior adaptability and robustness over CIoU and WISE-IoU in specific scenarios, offering enhanced evaluation metrics for object detection and classification tasks. This advancement underscores HIoU's potential to significantly improve the precision and reliability of object detection systems.

6. Conclusion

In the realm of object detection, the YOLOv8 algorithm stands out as a pinnacle of performance. This paper delves into the architectural framework of YOLOv8, with a particular emphasis on enhancing its capability to detect small targets—a persistent challenge in the field. To address this issue, the paper proposes several key optimizations to the YOLOv8 model. The paper presents an innovative approach to the loss function of the network. Recognizing the need for adaptive balancing of different components of the loss function across various stages of detection, the paper introduces IoU Rules. These rules are designed to dynamically adjust the contribution of each part of the loss function, ensuring that the network's focus remains aligned with the detection of small targets throughout the detection process.

This paper not only outlines the structural intricacies of the YOLOv8 algorithm but also presents a series of targeted optimizations aimed at elevating its performance in detecting small targets. Through the strategic integration of advanced convolution modules, attention mechanisms, and adaptive loss function adjustments, the YOLOv8 algorithm is poised to set a new standard in the field of object detection.

References

[1] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C].Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001:volume 1.:IEEE,2001:I-I.

[2] Ojala T, Pietikainen M, Maenpaa T. Multi resolution gray-scale and rotation invariant texture classification with local binary patterns[J].IEEE Transactions on pattern analysis and machine intelligence,2002,24(7):971-987.

[3] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C].2005 IEEE computer society conference on computer vision and pattern recognition(CVPR'05):volume 1.:IEEE,2005:886-893.

[4] Viola P, Jones M J. Robust real-time face detection[J]. International journal of computer vision,2004,57:137-154.

[5] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of computer and system sciences,1997,55(1):119-139.

[6] Felzenszwalb P, Mcallester D, Ramanan D. A discriminatively trained, multiscale, deformable part model[C].2008 IEEE conference on computer vision and pattern recognition.: IEEE, 2008:1-8.

[7] Felzenszwalb P F, Girshick R B, Mcallester D. Cascade object detection with deformable part models[C].2010 IEEE Computer society conference on computer vision and pattern recognition.:EEE,2010:2241-2248.

[8] Felzenszwalb P F, GirshicK R B. Object detection with discriminatively trained part-based models[J].IEEE transactions on pattern analysis and machine intelligence, 2009, 32 (9):1627-1645.

[9] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J].Communications of the ACM,2017,60(6):84-90.

[10] Simonyan K, Zisserman A.Very deep convolutional networks for large-scale image recognition[J].arXiv preprint arXiv: 1409. 1556,2014.

[11] Szegedy C, Liu W, Jia. Going deeper with convolutions [C]. Proceedings of the IEEE conference on computer vision and pattern recognition.2015:1-9.

[12] Ioffe S, Szegedy C. Batch normalization:Accelerating deep network training by reducing internal covariate shift[C]. International conference on machine learning.:pmlr, 2015:448-456.

[13] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C].Proceedings of the IEEE conference on computer vision and pattern recognition.2016:2818-2826.

[14] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception resnet and the impact of residual connections on learning[C].Proceedings of the AAAI conference on artificial intelligence: volume 31.2017.

[15] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C].Proceedings of the IEEE conference on computer vision and pattern recognition.2016:770-778.

[16] Girshick R, Donahue J,Darrell.Rich feature hierarchies for accurate object detection and semantic segmentation [C]. Proceedings of the IEEE conference on computer vision and pattern recognition.2014:580-587.