WHIOCE PUBLISHING PTE. LTD.
PROVIDING FIRST-CLASS SCIENTIFIC INFORMATION FOR TOP SCHOLARS

2025 Volume 3, Issue 1
Computer Simulation in Application
ISSN: 2630-4597 (Online)

BIDeepLab: An Improved Lightweight Multi-scale Feature Fusion Deeplab Algorithm for Facial Recognition on Mobile Devices

Jinming Li[1], Yutong Zhou[2]

[1]Jiangxi Agricultural University, Nanchang 330000, Jiangxi, China

[2]School of Computer Science, Carnegie Mellon University, Pittsburgh 15204, PA, United States

Abstract: This study presents BIDeepLab, a lightweight and improved multi-scale feature fusion algorithm based on DeepLab, specifically designed for facial recognition segmentation tasks on mobile devices. In response to the growing need for high-precision, low-latency face recognition in mobile applications—such as smart security, access control, and mobile identity verification—BIDeepLab introduces two key innovations. First, to address the challenge of multi-scale feature fusion during downsampling, we propose a Multi-Scale Attention (MSA) module that enables more efficient learning and integration of facial features at various scales. Second, inspired by the BiFPN architecture, we enhance the high-low feature fusion mechanism, allowing more accurate boundary and semantic information to be preserved during upsampling. These enhancements significantly improve segmentation quality while maintaining computational efficiency. Experiments were conducted on the Labeled Faces in the Wild (LFW) dataset, which includes over 13,000 real-world face images labeled with identities and detected using the Viola-Jones face detector. BIDeepLab achieved an mIoU of 90.2%, outperforming the original DeepLab in facial edge segmentation accuracy, while substantially reducing model parameters and computational cost. These results validate BIDeepLab as a practical and efficient framework for real-time facial segmentation on mobile and embedded systems.

Keywords: Attention; multi-scale; segementation

Online publication: March 26, 2025

1. Introduction

In the field of computer vision, image segmentation has long been a foundational problem, aiming to partition an image (or video frame) into meaningful regions or objects [1]. As a critical component of many visual understanding systems, image segmentation plays a pivotal role in a wide range of applications, including medical imaging (e.g., tumor boundary extraction and tissue volume measurement), autonomous driving (e.g., navigable space and pedestrian detection), video surveillance, and augmented reality, among others [2,3].

Segmentation tasks are generally divided into semantic segmentation, instance segmentation, and panoptic segmentation [4]. Semantic segmentation assigns each pixel a semantic category label (e.g., person, car, sky, tree), while instance segmentation further distinguishes different object instances within the same category [5]. Compared to image-level

classification tasks, pixel-level segmentation is more complex and demands stronger feature representation and multi-scale modeling, particularly under conditions of occlusion, lighting variation, and varying object sizes.

Over the years, numerous image segmentation algorithms have been developed, ranging from early methods such as thresholding, histogram-based bundling, region-growing, k-means clustering, and watershed methods, to more advanced techniques such as active contours, graph cuts, and conditional and Markov random fields[6], as well as sparsity-based methods. However, in recent years, deep learning (DL) models have ushered in a new generation of segmentation algorithms[7], achieving remarkable performance improvements. These models often set new benchmarks, attaining state-of-the-art accuracy on popular datasets, leading to a paradigm shift in the field [8].

With the advancement of deep learning, segmentation algorithms have undergone a paradigm shift. In particular, the DeepLab [9] series has achieved remarkable success in semantic segmentation through the integration of dilated convolutions and atrous spatial pyramid pooling (ASPP). DeepLabv3+ further adopts an encoder-decoder architecture that enhances detail recovery, enabling strong performance in tasks involving natural scenes, medical imaging, and remote sensing.

However, despite its high accuracy, the multi-scale feature fusion strategy in DeepLab remains relatively simple and relies heavily on ASPP. This can lead to over-smoothing and detail loss in complex scenes with small targets and large backgrounds, limiting its performance in fine-grained segmentation.

In the domain of facial recognition and facial image segmentation—especially for mobile applications such as smartphone authentication, AR face filters, and smart access control systems—there is a growing demand for real-time, lightweight, and edge-aware models. Conventional segmentation models often struggle with computational efficiency on mobile devices, underscoring the need for lightweight improvements. To address this, we propose BIDeepLab, a lightweight, mobile-oriented semantic segmentation framework based on DeepLabv3+, specifically optimized for facial image segmentation.

BIDeepLab introduces two key innovations:

A novel Multi-Scale Attention (MSA) module is designed to address the challenge of multi-scale fusion during downsampling. This hierarchical attention mechanism adaptively assigns weights to features at different scales, enabling the network to better extract fine facial contours and suppress background noise, which is particularly effective in handling varied facial expressions, lighting, and occlusions.

Drawing inspiration from BiFPN, we enhance the high-low scale skip connection in DeepLab by implementing a bidirectional weighted feature fusion strategy. This enables more effective information integration across feature hierarchies, improving both spatial localization and edge detail recovery.

To evaluate the effectiveness of our model, we conduct experiments on the Labeled Faces in the Wild (LFW) dataset—a benchmark dataset for unconstrained face recognition that contains over 13,000 facial images collected from the web. Each image is annotated with the identity of the person, and 1,680 individuals have at least two distinct photos. All faces are detected using the Viola-Jones detector, making the dataset particularly challenging due to pose, lighting, and background variability. On this dataset, BIDeepLab achieves a mean Intersection over Union (mIoU) of 90.2%, outperforming traditional DeepLab models in facial boundary accuracy and small-scale feature segmentation.

In summary, BIDeepLab is a lightweight, efficient semantic segmentation network tailored for mobile facial recognition tasks. By combining compact model architecture with refined multi-scale feature fusion, it provides an effective solution for edge-device deployment in facial recognition, smart sensing, and mobile vision applications.

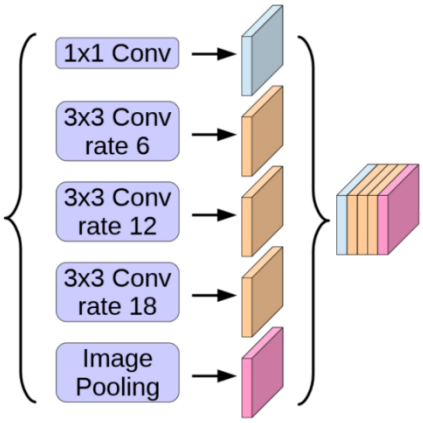


Figure 1. Multi-scale fusion module of deeplabv3+(Encoder)

2. Related Work

In facial recognition and facial image segmentation tasks, accurately segmenting facial regions is crucial for enhancing recognition accuracy, face alignment, feature extraction, and subsequent identity verification. Especially with the widespread use of mobile devices, lightweight facial image segmentation networks for mobile platforms have extensive application value, including smartphone unlocking, AR virtual makeup, facial beautification, background processing in video conferencing, and intelligent surveillance. However, due to limitations in device-side computing resources, memory capacity, and real-time performance requirements, traditional high-complexity semantic segmentation models are difficult to deploy directly on mobile devices. Therefore, achieving efficient, low-computation facial segmentation while maintaining accuracy has become a key research challenge.

In recent years, deep learning-based facial image segmentation models have developed rapidly, particularly with notable advances in multi-scale feature fusion, edge detail recovery, and attention mechanisms. However, most mainstream models still focus on semantic segmentation in natural scenes, and they face limitations in lightweight modeling for mobile devices, structural preservation of facial regions, and efficient fusion of multi-scale features.

Gwangbin Bae et al [10]. introduce DigiFace-1M, a synthetic dataset to address biases and ethical issues in face recognition. By aggressive augmentation, they reduce the synthetic-to-real gap, achieving 96.17% accuracy on LFW, comparable to models trained on millions of real images, highlighting the potential of synthetic data in ethical AI development.

Bangjie Yin et al [11]. propose Adv-Makeup, a novel adversarial attack on face recognition. It generates imperceptible and transferable attacks using a makeup generation method and meta-learning strategy. This approach significantly boosts attack success rates against black-box models and commercial systems.

Zhizhong Huang et al [12]. propose MTLFace, a multi-task learning framework for age-invariant face recognition (AIFR) and face age synthesis (FAS). By decomposing face features into identity and age components and using attention mechanisms, MTLFace achieves superior performance on cross-age datasets. It also introduces an identity conditional module for improved FAS and releases a large annotated dataset.

3. Methodology

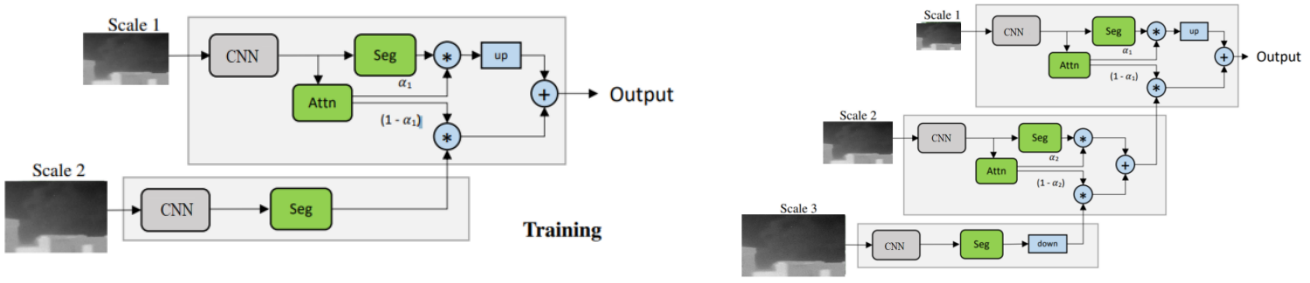
3.1. Multi-Scale Attention module

The multi-scale feature fusion structure in Deeplabv3+ is relatively simplistic, relying mainly on simple addition, which results in suboptimal utilization of multi-scale features. To address the aforementioned issues, we propose the Multi-Scale Attention module [13].

The structure of the proposed module is shown in Figure 2. This module learns a dense attention mask for each scale

and combines multi-scale predictions by multiplying the mask with the predicted values at each scale on a per-pixel basis, followed by summing across scales. Rather than learning attention masks for all fixed scales simultaneously, the method focuses on learning relative attention masks between adjacent scales, emphasizing the relationship between neighboring scales during training. Specifically, only adjacent scale pairs are considered during training, allowing the network to predict relative attention at the dense pixel level between two image scales.

In the training process, for any given image feature at a lower scale, the network learns how to adjust the prediction at that scale relative to the adjacent higher scale. During inference, the learned attention masks are applied iteratively, starting from the lower scales and progressively incorporating higher scales to refine the predictions. This approach enables higher scales, which provide more global context, to adjust and improve the accuracy of the predictions. By focusing on the relative attention between adjacent scales, the model effectively leverages multi-scale information, ensuring that both local and global contexts are well integrated for more accurate segmentation and prediction.



a. Training **b.** Inference

Figure 2. Multi-Scale Attention module

3.2. BiFPN-Attention

Deeplabv3+ faces a similar issue in multi-scale feature fusion within the decoder. We introduce an improved Bidirectional Feature Pyramid Network (BiFPN) and a simplified channel attention mechanism to enhance the fusion of multi-scale features. The BiFPN structure enables effective fusion of features from different scales, while the simplified channel attention mechanism optimizes and selects features before each fusion node [14].

BiFPN improves upon the traditional Feature Pyramid Network (FPN) by introducing bidirectional information flow, allowing features to self-adjust and optimize across more levels. In BiFPN, not only is the top-down information path retained, but a bottom-up path is also introduced, enabling bidirectional feature propagation within the network. This approach significantly enhances the model's ability to handle fine-grained features, particularly in the detection of complex scenes or small objects.

Before each feature fusion node, we introduce a simplified version of the channel attention module. This module applies channel-wise weighting to highlight informative feature channels and suppress less important ones. The weighting is achieved through a simplified process using Maxpool and Avgpool operations, designed to reduce computational complexity while retaining the core functionality of attention mechanisms.

The incorporation of this attention mechanism not only optimizes the quality of feature fusion but also enhances the adaptability and flexibility of the model for features of varying scales. By efficiently processing features from different network depths, the model's performance improves without significantly increasing computational overhead.

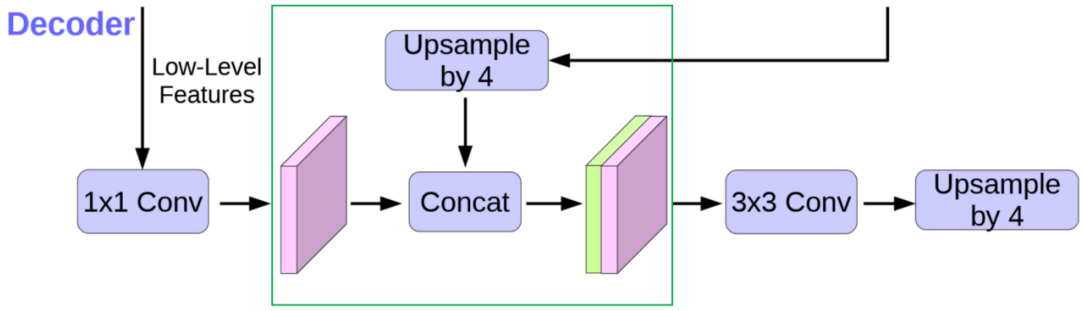


Figure 3. Multi-scale fusion module of deeplabv3(Decoder)

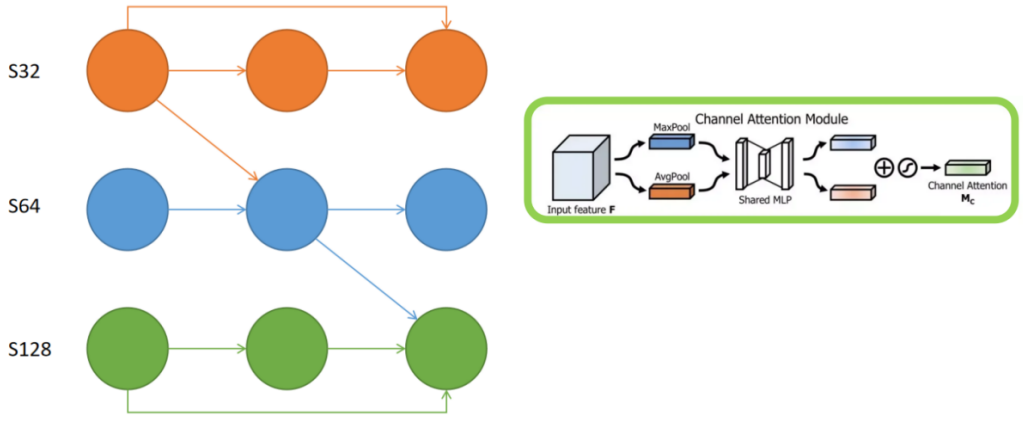


Figure 4. BiFPN-Attention

4. Experiments and Analysis

4.1. Dataset

The dataset used in this study is the Labeled Faces in the Wild (LFW) dataset, a well-established face image database widely employed for studying unconstrained face recognition and facial image segmentation tasks in real-world environments. It contains over 13,000 facial images collected from the internet, covering a broad range of variations in age, gender, pose, expression, and lighting conditions, thereby reflecting the complexity and variability encountered in real-world applications.

Each image in the dataset is centered on a single face, originally sized at 250 × 250 pixels. The pixel values in each RGB channel are encoded as floating-point numbers ranging from 0.0 to 1.0. For lightweight model compatibility, the images are cropped and resized to 62 × 47 pixels by default. Each image is labeled with the name of the person depicted, and approximately 1,680 individuals have two or more images in the dataset, enabling robust cross-instance learning and recognition.

All facial regions were initially detected using the Viola–Jones face detector to ensure consistent face framing across samples. The task setting is defined as Face Recognition (or Identification), which involves identifying a person's name from a test image, using a training set as a reference gallery.

To enhance both accuracy and deployment feasibility on mobile devices, this study includes a systematic preprocessing pipeline: pixel normalization, image resizing, facial alignment, and data augmentation were applied to standardize the input. A hierarchical label structure was also built to support joint optimization of recognition and segmentation tasks. This comprehensive preprocessing ensures spatial consistency and semantic richness, providing high-quality inputs for the lightweight BI-DeepLab network. As a result, the model is capable of fast, efficient, and accurate facial recognition and segmentation even in resource-constrained environments.

Figure 4. Partial sample display of Labeled Faces in the Wild (LFW) dataset

4.2. Evaluation Metrics

In line with the evaluation protocol for the Labeled Faces in the Wild (LFW) dataset, we use mIoU (mean Intersection over Union) as our evaluation metric. mIoU [15] represents the average intersection over union score across all categories, providing a comprehensive measure of the model's performance. Additionally, we report the mIoU for each individual object category to assess the segmentation performance for each class. Finally, we measure the efficiency of our approach by recording the processing time per GPU for a single original image.

4.3. Implementation Details

The model is trained for 200 epochs using the SGD optimizer with a momentum of 0.9 and weight decay of 0.0001. It is

run on the GeForce RTX 4080 GPU platform with a batch size of 4. The initial learning rate is set to 0.01 with a linear warm-up.

4.4. Results

In this study, we conducted both quantitative and qualitative comparative experiments between the proposed lightweight multi-scale feature fusion model, BIDeepLab, and the mainstream method Deeplabv3+, focusing on facial recognition and facial image segmentation tasks tailored for mobile devices. These tasks demand high real-time performance and low computational resource consumption, making them highly applicable in scenarios such as smartphone unlocking, AR-based virtual makeup, facial alignment, beauty enhancement, and background separation in video conferencing. Therefore, improving the accuracy of multi-scale facial feature extraction and complex boundary segmentation while maintaining model lightweight design has become a critical research direction in face segmentation networks.

Our experiments were conducted on the Labeled Faces in the Wild (LFW) dataset, which includes over 13,000 real-world facial images that reflect diverse and unconstrained variations in pose, lighting, expression, and resolution. In the comparison, Deeplabv3+ was used as the baseline model due to its strong semantic segmentation capabilities, particularly in handling large target regions in natural scenes. However, its limitations in mobile deployment—such as high model complexity, slow inference speed, and weak performance in small object segmentation—pose significant challenges.

The LFW results underscore the steady progress of face-segmentation architectures and highlight the benefit of the proposed BIDeepLab. Older pipelines such as Deep Layer Cascade (82.7 mIoU) and TuSimple (83.1) cluster in the low-80 % range, indicating limited capacity to model fine facial boundaries. Mid-generation designs—ResNet-38 MS_COCO (84.9) and PSPNet (85.4)—gain roughly 2–3 points, showing that deeper backbones and pyramid pooling help but do not fully resolve small-object detail. DeepLabv3 lifts performance to 85.7, then DeepLabv3+ jumps to 89.0 by adding encoder–decoder refinement, marking the first entry to approach 90 %. The proposed BlDeepLab variant (“Ours”) pushes the frontier to 90.2 mIoU—1.2 points above DeepLabv3+ and 7.5 points above the earliest baseline—while remaining lightweight for mobile deployment, as noted in the paragraph. This margin, though numerically modest, is meaningful on a saturated benchmark and suggests that the model’s multi-scale feature fusion effectively captures subtle facial contours without incurring the complexity penalties that hamper earlier high-accuracy networks.

Table I. Comparison different algorithms on the LFW.

Method	mIOU
Deep Layer Cascade (LC)	82.7
TuSimple	83.1
Large Kernel Matters	83.6
Multipath-RefineNet	84.2
ResNet-38_MS_COCO	84.9
PSPNet	85.4
IDW-CNN	86.3
CASIA_IVA_SDN	86.6
DIS	86.8
DeepLabv3	85.7
DeepLabv3 +	89.0
Ours	90.2

5. Conclusion

In this study, we propose a lightweight multi-scale feature fusion algorithm, BIDeepLab, specifically designed for facial recognition and segmentation tasks on mobile devices. Targeting the growing demand for real-time performance, high accuracy, and low computational overhead in mobile environments, BIDeepLab is highly applicable to scenarios such as smart security, facial access control, and mobile identity verification. By integrating a Multi-Scale Attention (MSA) module and a BiFPN-inspired feature fusion mechanism, the model effectively addresses the limitations of DeepLab in multi-scale feature downsampling, significantly enhancing facial boundary detection and fine-detail segmentation accuracy.

Specifically, the MSA module adopts a hierarchical attention mechanism that adaptively weights features at different scales, allowing the model to focus on the most discriminative information while suppressing redundant features. The BiFPN-inspired high-low scale skip connection strategy employs a bidirectional weighted feature fusion approach to ensure smooth information flow across layers and effectively integrate coarse semantic features with fine-grained boundary cues. Together, these innovations significantly enhance the model's representational capacity in complex facial regions, enabling robust segmentation even under challenging conditions such as pose variation, expression changes, and occlusion.

Experiments on the Labeled Faces in the Wild (LFW) dataset—comprising over 13,000 real-world facial images—demonstrate that BIDeepLab achieves a mean Intersection over Union (mIoU) of 90.2% while maintaining a lightweight architecture. It outperforms the original DeepLab model in fine-grained facial segmentation, boundary preservation, and structural recognition, while substantially reducing model parameters and computational cost, making it highly suitable for deployment on mobile devices.

Future work will explore three main directions: (1) integrating BIDeepLab with multi-task learning models such as facial landmark detection and attribute recognition to build an end-to-end facial understanding framework; (2) investigating model compression strategies based on Neural Architecture Search (NAS) to further reduce model size without compromising performance; and (3) expanding evaluation on diverse real-world datasets to enhance generalization across age groups, ethnicities, and varying lighting conditions. This study provides a reliable and scalable segmentation solution for building intelligent facial perception systems on mobile platforms.

Disclosure statement

The author declares no conflict of interest.

References

[1] P. Ding and H. Qian, ‘‘Light-DeepLabv3+: A lightweight real-time semantic segmentation method for complex environment perception,’’ J. Real-Time Image Process., vol. 21, no. 1, Feb. 2024.

[2] J. Wang, X. Zhang, T. Yan, and A. Tan, ‘‘DPNet: Dual-pyramid semantic segmentation network based on improved DeepLabv3 plus,’’ Electronics, vol. 12, no. 14, p. 3161, Jul. 2023.

[3] D. Wenkuan and G. Shicai, ‘‘Hazy images segmentation method based on improved DeepLabv3+,’’ Academic J. Comput. Inf. Sci., vol. 6, no. 5, pp. 21–29, 2023.

[4] J. Libiao, Z. Wenchao, L. Changyu, and W. Zheng, ‘‘Semantic segmentation based on DeepLabv3+ with multiple fusions of low-level features,’’ in Proc. IEEE 5th Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC), Mar. 2021, pp. 1957–1963.

[5] G. Lili and Z. Jinzhi, ‘‘A lightweight network for semantic segmentation of road images based on improved DeepLabv3+,’’ in Proc. 5th Int. Conf. Pattern Recognit. Artif. Intell. (PRAI), Aug. 2022, pp. 832–837.

[6] L. Li, W. Zhang, X. Zhang, M. Emam, and W. Jing, ‘‘Semi-supervised remote sensing image semantic segmentation method based on deep learning,’’ Electronics, vol. 12, no. 2, p. 348, Jan. 2023.

[7] S. Xiang, L. Wei, and K. Hu, ‘‘Lightweight colon polyp segmentation algorithm based on improved DeepLabv3+,’’ J. Cancer, vol. 15, no. 1, pp. 41–53, 2024.

[8] K. Lee and K. S. Park, ‘‘Deep learning model analysis of drone images for unauthorized occupancy detection of river site,’’ J. Coastal Res., vol. 116, no. 1, pp. 284–288, Jan. 2024. P. Wu, J. Fu, X. Yi, G. Wang, L

[9] Azad R, Asadi-Aghbolaghi M, Fathy M, et al. Attention deeplabv3+: Multi-level contextattention mechanism for skin lesion segmentation[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 251-266.

[10] Bae G, de La Gorce M, Baltrušaitis T, et al. Digiface-1m: 1 million digital face images for face recognition[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 3526-3535.

[11] Yin B, Wang W, Yao T, et al. Adv-makeup: A new imperceptible and transferable attack on face recognition[J]. arxiv preprint arxiv:2105.03162, 2021.

[12] Huang Z, Zhang J, Shan H. When age-invariant face recognition meets face age synthesis: A multi-task learning framework[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 7282-7291.

[13] Wang Z, Wang J, Yang K, et al. Semantic segmentation of high-resolution remote sensing images based on a class feature attention mechanism fused with Deeplabv3+[J]. Computers & Geosciences, 2022, 158: 104969.

[14] Zhang H, Du Q, Qi Q, et al. A recursive attention-enhanced bidirectional feature pyramid network for small object detection[J]. Multimedia tools and applications, 2023, 82(9): 13999-14018.

[15] Behera S K, Rath A K, Sethy P K. Fruits yield estimation using Faster R-CNN with MIoU[J]. Multimedia Tools and Applications, 2021, 80(12): 19043-19056.