Leftover Food Measurement using Deep Learning Based Semantic Segmentation

Haruhiro TAKAHASHI[1], Ryuto ISHIBASHI[2], Hayata KANEKO[2] and Lin MENG[1,†]

[1]*College of Science and Engineering, Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga, Japan 525-8577*

[2]*Graduate School of Science and Engineering, Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga, Japan 525-8577*

Abstract

In the super-aging society, the number of caregivers for the elderly is insufficient. Furthermore, the increasing demand for quality care exacerbates the labor shortage problem. Hence, care technology work has become increasingly important. This paper focuses on automatically measuring food intake to monitor the health situation of the elderly. The proposed approach utilizes a semantic segmentation model based on deep learning methods. Specifically, the U-Net architecture is employed as a foundational module for segmenting food intake from the after-meal plate. Attention modules are integrated into the U-Net to enhance the Mean Intersection over Union (mIoU) while minimizing the number of model parameters and FLOPs (Floating Point Operations). Furthermore, an Attention Search method is proposed to search for the optimized Attention module insert position in the U-Net. The experimental results demonstrate that the proposed method achieves a high score, such as mIoU of 87.1%, with only a slight increase in the number of parameters and FLOPs, thus confirming its effectiveness. Optimizing and applying the proposal to practice is important work in the future. To further optimize the number of parameters and FLOPs, the Attention insertion position should be explored with a technique called Neural Architecture Search (NAS).

1. Introduction

The number of people aged 65 and over in relation to the world's population has been rising yearly and is expected to continue to rise [1]. In Japan, about 29% of the total population is 65 and over, making it a super-aged society. Also, only 12.1% of care providers can provide care to people in need of care, and labor shortages in the nursing care industry are becoming serious [2]. In addition to their usual duties, such as caring for residents, caregivers keep food and water intake records. As a problem, recording the amount of food and water is controlled manually by a person, and the standard is not uniform. Recently, AI-based image recognition methods have been widely used and have significant achievements in various fields[3, 4, 5, 6, 7].

The 6th International Symposium on Advanced Technologies and Applications in the Internet of Things (ATAIT 2024), August 19-22, 2024, Kusatsu, Japan

[†]Corresponding author.

ri0126hi@ed.ritsumei.ac.jp (H. TAKAHASHI); ri0097fx@ed.ritsumei.ac.jp (R. ISHIBASHI); ri0099xx@ed.ritsumei.ac.jp (H. KANEKO); menglin@fc.ristumei.ac.jp (L. MENG)

http://www.ihpc.se.ritsumei.ac.jp/ (L. MENG)

0000-0003-4351-6923 (L. MENG)

Therefore, this study mitigates manual tasks by implementing image segmentation with AI and unifying food intake criteria.

This paper proposes the model using semantic segmentation, one of the AI-based image segmentation models to detect the leftover food area. However, due to computer resource limitations, a highly accurate deep-learning image recognition model with AI is computationally expensive and challenging to apply to nursing homes. In addition, to solve this problem, an Attention module that improves accuracy with a small increase in computational complexity is adopted, and the performance is validated. The problem is that searching for the U-Net manually optimized Attention module insertion position takes much time. Therefore, Attention Search is proposed to decide the optimal Attention module insertion position automatically.

The major contributions of this paper are shown as follows:

- Propose AI-based semantic segmentation model to reduce food intake measurement of caregivers' tasks.
- Optimize the Attention module insertion position using Attention Search.
- Proposed U-Net model achieves higher accuracy with less computation than conventional semantic segmentation models.

The remaining parts of this paper are organized as follows. Related works are listed in Section 2. Section 3 proposes our U-Net model. Section 4 shows the dataset, the experimental method, experimental results, and the discussion. Finally, this paper is concluded in Section 5.

2. Related work

2.1. Semantic Segmentaiton Model

Image recognition technology consists of image classification, object detection, and image segmentation. Semantic segmentation is one of the image segmentation methods used in this paper. Semantic segmentation is a method of labeling what appears in each pixel of an image.

U-Net

U-Net [8] is one of the models for semantic segmentation that consists of an encoder and a decoder. U-Net encoder convolves the input image several times to extract features. This part of the network structure is called the backbone, and the model learned by image classification tasks such as ResNet [9] can be used as the backbone. U-Net decoder takes the features extracted by the encoder, performs deconvolution, and outputs a probability map of the same size as the input image. In addition, U-Net concatenates the encoder feature map to the decoder feature map for each hierarchy.

DeepLabv3

DeepLabv3 [10] is one of the models for semantic segmentation. DeepLabv3 uses pre-trained models for image classification tasks as the backbone. The backbone extracts image feature maps, performs parallel Atrous Spatial Pyramid Pooling (ASPP), and then performs several convolutions to output the segmented image. The ASPP is inspired by the Spatial Pyramid Pooling (SPP) [11]. In SPP, pooling is performed on image feature maps in a plurality of scales, and concat and output them. Therefore, it has the advantage that it can cope with input of various sizes and multi-scales. In ASPP, SPP is performed efficiently by applying atrous convolution to pooling while maintaining the advantage of SPP.

2.2. Attention Module

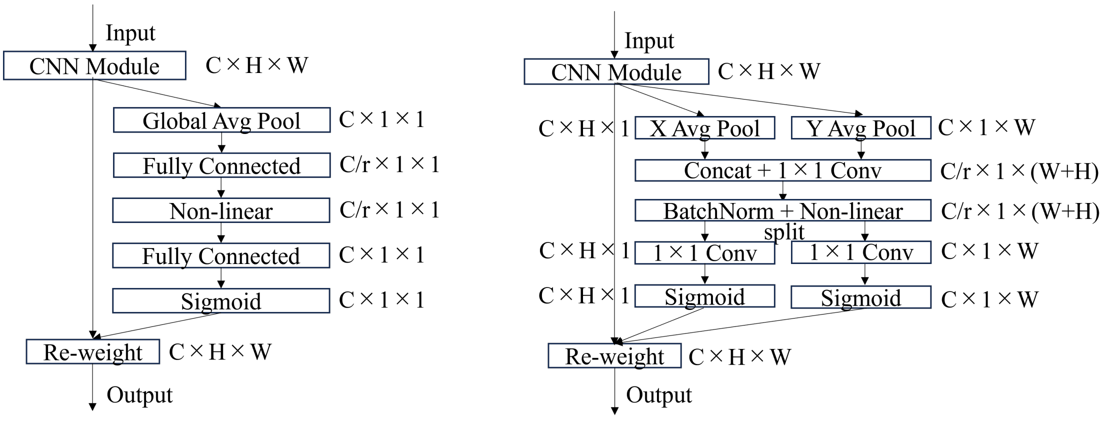
Attention module is a module for improving accuracy with a small number of parameter increases. It is inserted into several points of the model. In Figure 1, **r** is the compression rate of the number of channels, and the computational complexity becomes smaller as **r** is increased.

Squeeze and Excitation Module

Squeeze and Excitation Module (SE-Module) [12] is one of the Attention modules and its structure is shown in Figure 1 (a). The Squeeze part uses global average pooling to take advantage of the feature size of the entire image. In the Excitation part, the obtained feature size of each channel is passed through several layers to obtain the weight of each channel, which is then multiplied by the input feature maps and output.

Coordinate Attention Module

Coordinate Attention Module(CA-Module) [13] is one of Attention modules and its structure is shown in Figure 1 (b). Unlike the SE-Module, the CA-Module is divided into W and H directions for pooling and weighting. By dividing, not only the computational complexity is reduced, but also spatial information can be included in the weighting.



(a) (b)

Figure 1: (a) Structure of Squeeze-and-Excitation Module (b) Structure of Coordinate Attention

2.3. MobileNetV3

MobileNet is one of the models that can be installed in a mobile phone with a small number of parameters and can be used for various tasks. MobileNetV1 [14] uses Depthwise Separable Convolution, a convolution that performs Depthwise convolution followed by Pointwise convolution, to reduce the number of parameters compared to normal convolution. MobileNetV2 introduced an Inverted Residual to reduce the large calculation amount of Pointwise Convolution in MobileNetV1. Inverted Residual is an improvement on ResNet's Residual Block, approximating Depthwise Separable Convolution with a small number of calculations by sandwiching Depthwise Convolution with Pointwise Convolution. In MobileNetV2, the number of channels is reduced compared with MobileNetV1, but only in Inverted Residual, the number of channels is increased to sufficiently extract the feature quantity, thereby reducing the parameter quantity. MobileNetV3 [15] introduces SE-Module for Inverted Residual and changes some non-linear activation from ReLU to Hardswish. In addition, hardware-enabled Neural Architecture Search (NAS)[16] is used to tailor the model to the phone's CPU.

3. Proposed Method

In this paper, U-Net model for semantic segmentation and Attention Search that optimizes the position of the Attention module are proposed to realize higher accuracy with less computation.

3.1. Our U-Net model

This paper proposes a U-Net model with Attention module that increases the accuracy by a small number of parameters, without increasing the number of channels. Its structure is shown in Figure 2. In Two Convolution Block(TCB) of Figure 2, 3×3 convolution, batch normalization, and ReLU (CBR) are performed twice, and Attention Point to insert Attention module is after each CBR.

3.2. Attention Search

Attention Search is performed as a method to insert an Attention module at the optimal position to improve accuracy. Attention module used in this paper is the SE-Module and the CA-Module. For each SE-Module and CA-Module, this method searches which TCB is the best position to insert the attention module at Attention Point 1 and 2, respectively.

Attention Search is performed as shown in Algorithm 1. A includes the SE-Module and CA-Module, P includes the Attention Point 1 and 2, N indicates the number of TCBs in our U-Net model. a indicates Attention module to be inserted, p indicates Attention Point to be inserted, m indicates set of models to save, i indicates the number of the pattern to be executed, μ indicates the model to be saved, M indicates the maximum score model in the set m, $\mathscr{M}$ indicates set of maximum score models to save. Then, an explanation of Algorithm.1 is described. In line 2, Attention Module is selected as SE-Module or CA-Module. In line 3, Attention Point is selected as 1 or 2. In line 5, selecting the pattern to insert Attention module to be executed In line 6 through 8, inserting Attention module into the TCB corresponding to the selected

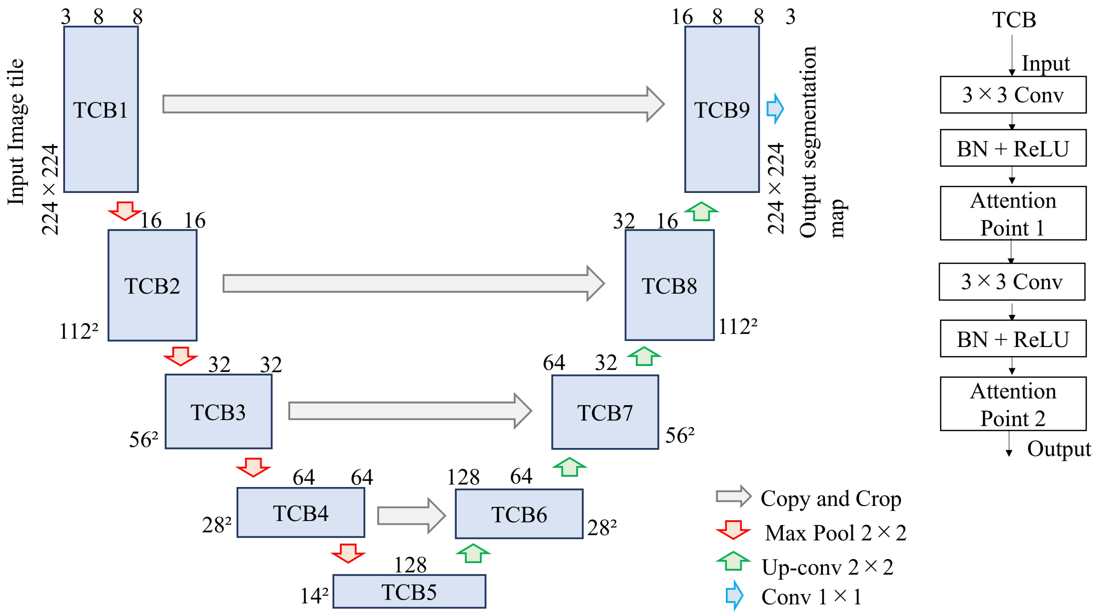


Figure 2: Structure of Our U-Net Model

pattern. In line 10 through 16, the selected model is trained, and is saved when the epoch is $T \times 0.8$ or greater. In line 18 through 19, the model with the maximum score from the saved models is selected. The score used is mIoU. In all processing of lines 1 through 21, Attention Search explores the model with the maximum score.

4. Experiments

4.1. Dataset

The dataset used in this study was created by taking pictures before, during, and after meals and labeling the areas of rice and miso soup. The image sizes are 456 (or 455) ×608 and 608×456 (or 455), and the input image of the model is resized to 224×224. There are a total of 241 samples and about 85% (206 samples) of the total are used in the training and about 15% (35 samples) of the total are used in the test. The dataset with the correct labels is shown in Figure 3.

4.2. Experiments Method

Implementational Conditions

Experiments in this paper are conducted on Intel(R) Xeon(R) Gold 6342 CPU, a single NVIDIA RTX A6000 GPU, and NVIDIA Jetson-Nano. The operating system is Ubuntu 22.04.6 LTS, the CUDA version is 12.2, and the jetpack version is 4.6.1 (the tensorRT version is 8.2). PyTorch library is used as the framework for implementing deep learning.

Algorithm 1 Attention Search

```
Input: $A$: sets of Attention modules,
    $P$: sets of Attention Points,
    $N$: number of TCBs in U-Net,
    $T$: total epochs,
Output: $\mathscr{M}$: Best Models
$\mathscr{M} = \{\}$
for $a$ in $A$ do
    for $p$ in $P$ do
        $m = \{\}$
        for $i = 0$ to $2^N - 1$ do
            if $i \equiv 1 \pmod 2$ then
                $a$ is inserted in $p$ of TCB[$j$]
            end if
            $i = i//2$
            for $epoch = 1$ to $T$ do
                Training
                if $epoch \geqq T \times 0.8$ then
                    $\mu \leftarrow$ Model
                    $m \leftarrow m \cup \{\mu\}$
                end if
            end for
        end for
        $M \leftarrow$ Maximum Score of $m$
        $\mathscr{M} \leftarrow \mathscr{M} \cup \{M\}$
    end for
end for
```

Learning Conditions

The model is learned under the following conditions.

- Loss Function: The loss function used in this paper is Tversky Loss [17]. Equation (1) is the formula for calculating Tversky Loss. p is the predicted label and t is the correct label . In this paper, α=0.7 and β=0.3.

$$\text{Tversky\_Loss}(p, t) = 1 - \frac{\sum pt}{\sum pt + \alpha \sum (1 - p)t + \beta \sum p(1 - t)} \tag{1}$$

- Optimizer: Adam [18]
- Learning Rate : 1e-4
- Batch Size: 4
- Epoch: 100

Evaluation index

The evaluation index used in this paper is IoU, Throughput, Params, and FLOPs.

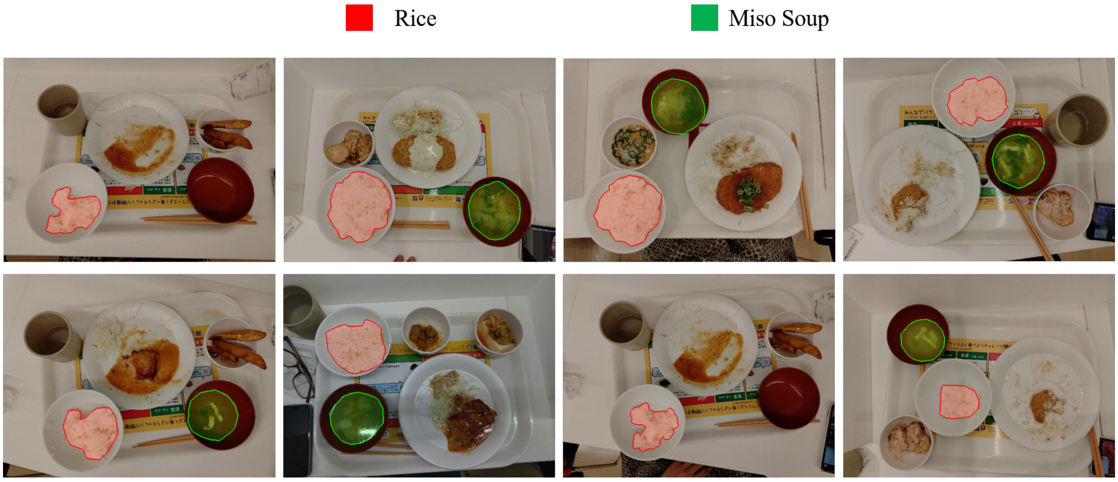


Figure 3: Dataset Examples

- IoU: Equation (2) is the formula for calculating IoU (Intersection over Union). p is the predicted label and t is the correct label .

$$IoU(p,t) = \frac{\sum pt}{\sum p + \sum t - \sum pt} \tag{2}$$

- Throughput: Throughput is a measure of how many images can be processed per second.
- Params: Params is the parameter quantity of the model
- FLOPs: FLOPs (FLoating-point OPerationS) is the computational complexity of the model.

4.3. Experimental Result

Table 1 shows the results of the models explored by Attention Search. As the comparison models, DeepLabv3 with MobileNetV3 small and large as the backbone are adopted. Segmentation results for each model type are as shown in Figure 3. A and B represent the case where the SE-Module is inserted into Attention Point 1 of TCB{2,5,6,7} and Attention Point 2 of TCB{3,7,9}, respectively. C and D represent the case where the CA-Module is inserted into Attention Point 1 of TCB{1,2,3,4,5,7,8} and Attention Point 2 of TCB{1,4,5,8,9}, respectively.

Type: B has the highest mIoU in the proposed models and the score is obviously superior for the baseline and comparison models. The FLOPs of proposed models differ little between the types and are less than the comparison models. For the number of parameters, Type: B of the proposed model increases the least from the baseline and is smaller than the comparison model. For the score of throughput, in the case of CPU and GPU, Type: B has the best score at 82 and 313 respectively. In the case of Jetson, Type: A and Type: B have better scores at 47 and 46. Note that the throughput of the baseline model is higher than other proposed U-net models because the baseline model has no additional modules.

Table 1
Result of mIoU, FLOPs, Params, and Throughput

Model	Type	Attn.	mIoU	FLOPs	Params	Throughput (ips)		
			(%)	(G)	(M)	CPU	GPU	Jetson
MobileNetV3	Small		81.1	1.105	6.127	21	201	37
	Large		82.9	1.908	11.025	31	177	21
U-Net (Ours)	Baseline	—	84.1	0.702	0.488	83	359	49
	A	+SE	86.4	0.702	0.510	45	300	**47**
	B	+SE	**87.1**	0.702	**0.490**	**82**	**313**	46
	C	+CA	86.1	0.704	0.496	70	192	41
	D	+CA	86.8	0.704	0.494	78	214	42

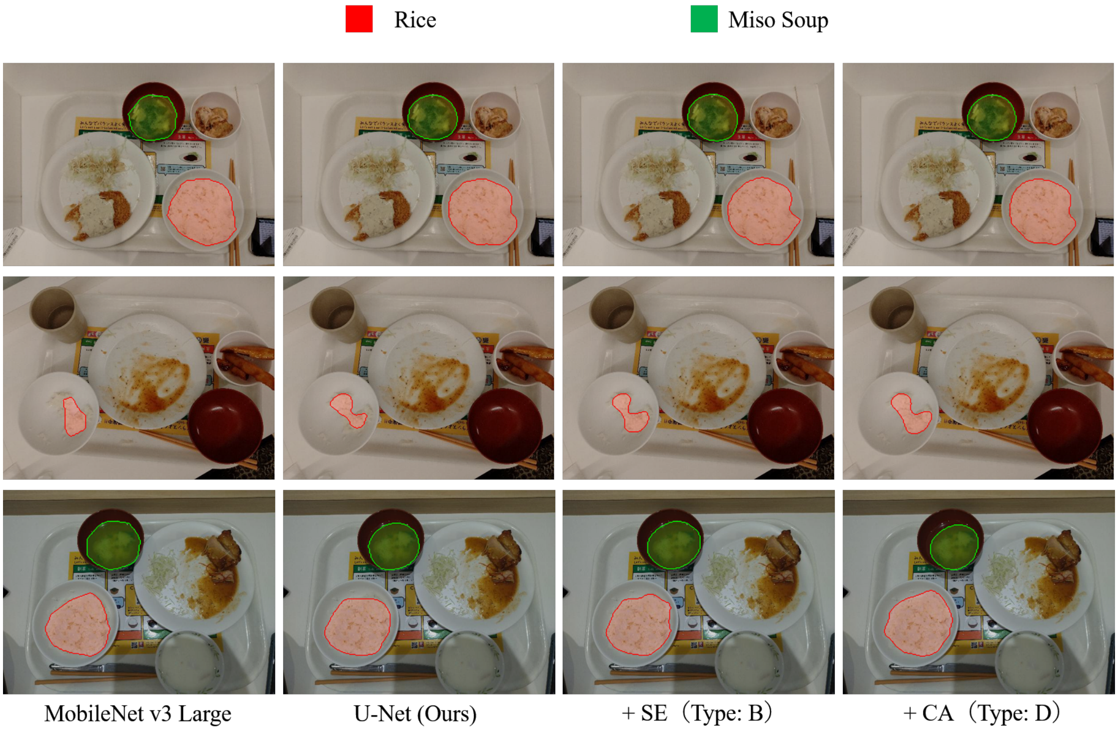


Figure 4: Segmentation Results

4.4. Discussion and Future Work

From the results, inserting Attention module is an effective way to improve mIoU. However, depending on the number of Attention module insertions, the number of parameters and FLOPs increases, and the throughput decreases. As a result of the Attention Search, it is shown that mIoU is improved by inserting in Attention Point 2 rather than Attention Point 1 in both SE and CA cases without much increase in the parameter amount. In addition, SE is more effective compared to CA in Attention Point 1 and 2 because mIoU is higher and the number of parameters is similar. From the results of the throughput (Table 1), it is considered that the throughput corresponds to the number of parameters because the number of cores is not sufficient when executing on CPU, and the throughput corresponds to FLOPs because the number of cores is sufficient when executing on GPU and Jetson. As shown in Figure 3, segmentation image accuracy is improved corresponding to mIoU. According to the results of Table 1, Type: B of the proposed model has a higher score in all the evaluation indexes, which shows Type: B of the proposed model superiority. Based on the discussion so far, it can be shown that the model without the number of channels increased and added the Attention module has an advantage over the DeepLabv3 models with MobileNetV3 as the backbone for the task of this paper with the small dataset.

The amount of leftover food is detected by calculating the volume corresponding to the obtained area. The food intake is estimated from the volume obtained from photographs taken before and after meals. However, this method can only be used if the plate size does not change between the photographs taken before and after meals. Future work includes an object detection [19, 20] component for detecting the plate, allowing us to make estimations even if the plate size changes. Additionally, future work covers not only rice and miso soup but also other dishes. Furthermore, we constructed a model with high mIoU using Attention Search; however, brute force is not efficient. For this reason, it is necessary to search for a model using NAS in future work.

5. Conclusion

The aging society is progressing worldwide, and the labor shortage in the nursing care industry is becoming serious in Japan. This paper proposes the model using semantic segmentation to measure food intake, which is one of the tasks in the nursing care industry. The proposed model with Attention module improves the IoU while suppressing the increase in the number of parameters and FLOPs of the model. In conclusion, our proposed method is suitable for measuring leftover food areas due to the limitations of computational resources. Future work implements model with addition of object detection to measure food intake without being affected by plate size. Additionally, pruning models using NAS is an interesting direction for future work. Further, this work is expected to be applied in food service on robot-based food industry automation.

Acknowledgments

The dataset used in this work is provided by MOUantAI, ”Meshi-Pasha Project”.

References

[1] U. Nations, Population by broad age groups, 2022. https://population.un.org/wpp/Graphs/DemographicProfiles/Line/900.

[2] C. O. of Japan, Annual report on the ageing society, 2023, 2023. https://www8.cao.go.jp/kourei/whitepaper/w-2023/html/zenbun/index.html.

[3] H. Li, L. Meng, Hardware-aware approach to deep neural network optimization, Neurocomputing 559 (2023) 126808.

[4] X. YUE, L. MENG, Yolo-sm: a lightweight single-class multi-deformation object detection network, IEEE Transactions on Emerging Topics in Computational Intelligence (2024).

[5] X. YUE, L. MENG, Yolo-msa: A multi-scale stereoscopic attention network for empty-dish recycling robots, IEEE Transactions on Instrumentation and Measurement (2023).

[6] Y. Ge, Z. Li, X. Yue, H. Li, Q. Li, L. Meng, Iot-based automatic deep learning model generation and the application on empty-dish recycling robots, Internet of Things 25 (2024) 101047.

[7] S. Hayashi, Q. Li, L. Meng, Facial expression recognition with face mask using attention mechanism and metric loss, in: 2023 International Conference on Advanced Mechatronic Systems (ICAMechS), IEEE, 2023, pp. 1–6.

[8] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, Springer, 2015, pp. 234–241.

[9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[10] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv:1706.05587 (2017).

[11] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE transactions on pattern analysis and machine intelligence 37 (2015) 1904–1916.

[12] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[13] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 13713–13722.

[14] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 (2017).

[15] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., Searching for mobilenetv3, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1314–1324.

[16] B. Zoph, Q. V. Le, Neural architecture search with reinforcement learning, arXiv preprint arXiv:1611.01578 (2016).

[17] S. S. M. Salehi, D. Erdogmus, A. Gholipour, Tversky loss function for image segmentation using 3d fully convolutional deep networks, in: International workshop on machine learning in medical imaging, Springer, 2017, pp. 379–387.

[18] P. K. Diederik, Adam: A method for stochastic optimization, (No Title) (2014).

[19] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer, 2016, pp. 21–37.