Volume 6, Issue 1, 2022 ISSN: 2617-4537
DOI: https://doi.org/10.31058/j.data.2022.61003

Skin Segmentation Based on Improved DeepLabV3+

Taiyu Han[1], Ping Li[1,2], Enlin Yang[3], Xiaojin Liu[1], Shikai Feng[1], Hongliu Yu[1*]

[1] Institute of Rehabilitation Engineering Technology, University of Shanghai for Science and Technology, Shanghai, China
[2] Department of Biomedical Engineering, Changzhi Medical College, Changzhi, China
[3] Department of Industrial Engineering and Decision Analytics, The Hong Kong University of Science and Technology, Hong Kong, China

Email Address

yhl98@hotmail.com (Hongliu Yu)
*Correspondence: yhl98@hotmail.com

Received: 25 March 2022**;** **Accepted:** 14 April 2022**;** **Published:** 26 April 2022

Abstract:

In recent years, the number of elderly people and people with disabilities has been increasing. However, the intelligence of the bathing equipment used to assist the elderly and disabled lags far behind that of other medical equipment. The first task of developing an intelligent autonomous bathing aid system without human intervention is to achieve automatic segmentation and localization of skin. The traditional image segmentation model has low accuracy in the case of insufficient light, foam, and water mist occlusion, while the deep learning model exemplified by DeepLabV3+ has high accuracy and robustness. Replacing the backbone feature extraction network in DeepLabV3+ with the lightweight neural network MobileNetV2 can effectively solve the problem of computing power limitation when the deep learning algorithm is deployed to embedded devices, and the generalization performance ofthe improved model is optimized by five scenario-specific data enhancement methods and migration learning ideas. Experiments show that the proposed improved method reduces the weight file size by 93% and 87% with only 1.1% and 1.2% reduction in MIoU compared with the original and Unet models, respectively, achieving a better balance in accuracy and performance.

Keywords:

Skin Segmentation, Deep Learning, MobileNetV2, DeepLabV3+, Data Augmentation

1. Introduction

According to the seventh national census in 2020 and the second national sample survey of people with disabilities, China's population aged 60 and above was 264.02 million, accounting for 18.70% of the country's total population; the total number of people with various types of disabilities was 82.96 million, accounting for 6.34% of the country's population, of which 26.04 million were physically disabled [1,2]. However, at the same time, there is a large gap between the basic needs of the elderly and the disabled and the services provided by society. The intelligence of equipment to assist the elderly and disabled lags far behind the development of other medical equipment.

Bathing, for example, is an essential part of daily hygiene and reflects the basic ability to care for oneself in the home, as part of the Activities of Daily Living (ADL) assessment [3]. For elderly people and people with disabilities, bathing on their own is extremely draining. However, bathing with the assistance of caregivers involved may affect their self-esteem and cause mental burden[4]. With the increasing demand for bathing equipment for the elderly and disabled, a series of bathing systems have been developed at home and abroad [5,6,7], but these systems have a series of problems such as low intelligence, poor safety, and high prices. To truly free caregivers and realise autonomous bathing for elderly people and people with disabilities, the team proposes to develop the first international all-round intelligent bathing aid system based on visual positioning, filling the gap of intelligent bathing equipment. The system uses a camera to identify and locate the skin in the bathing scene. The position information obtained is used to control the robot arm to spray soap and then scrub.

The concept of image semantic segmentation was first introduced by Ohta et al. in 1978: each pixel in an image is assigned a predefined label indicating its semantic class [8]. Skin segmentation, the process of finding skin-coloured pixels and regions in an image or video, is the basis for many complex computer vision tasks, and the results of pre-processing an image or video are key to the success of many computer vision applications, in areas such as face recognition [9], medical diagnosis [10] and intelligent human-computer interaction [11].

With the development of deep learning and computing power, skin segmentation algorithms based on deep learning have become the mainstream algorithms in the field of image segmentation, with accuracy and efficiency far exceeding traditional algorithms [12]. According to the different classification granularity, deep learning skin segmentation can be divided into two categories: Image semantic segmentation based on the regional classification first generate a large number of candidate regions for the original input image, extract semantic information and features for each candidate region, and output the results after classification. In contrast, Image semantic segmentation based on the pixel classification directly classifies each pixel in the image, and the original image is directly outputted after an end-to-end model, which not only improves the learning efficiency but also effectively increases the segmentation accuracy [13].

However, most of the research on skin segmentation has been focused on applications such as face recognition and medical diagnosis. Less research has been conducted on skin segmentation in bathing scenarios where there is insufficient illumination, water mist, and foam occlusion. The improved DeepLabV3+ model proposed in this paper can be deployed on embedded devices and provides an effective and lightweight method for skin segmentation, which promotes the development of age-appropriate home cleaning personal care devices for the elderly and disabled.

2. Methodology

2.1. MobileNet

Deep learningbased image segmentation systems firstly extract features from the original input image, and secondly, classify the target by the extracted features.

Common backbone feature extraction networks include Inception, VGGNet, ResNet, AlexNet [14] and so on. With the continuous iterative development of deep learning, the structure of neural networks is becoming more complex and larger, the performance of the models is constantly being improved. However, the hardware resources required for training and prediction are also becoming more and more demanding. Mobile and embedded devices are limited by computing power, which makes it difficult to train and predict complex deep learning models. In practical engineering applications, embedded devices are often used for algorithm deployment. To solve the problems of computing power limitation and insufficient training, deep learning models are constantly developing towards lightweight, including NasNet-A, ShuffleNet, Xception [15], MobileNet [16], etc.

The MobileNet family is a lightweight deep neural network proposed by Google for embedded devices. Compared to other lightweight networks, the MobileNet family maintains a high accuracy rate with a low number of parameters. The accuracy of the MobileNet model, the number of parameters number, computation size, and time consumed running with google Pixel phones compared to several other common lightweight neural networks are shown in Table 1 [17].

Table 1. *Comparison of common lightweight neural networks.*

Network	Accuracy	Params	MAdds	pixel
NasNet-A	74%	5.3M	564M	192ms
ShuffleNet(1.5)	69%	2.9M	292M	-
ShuffleNet(x2)	70.9%	4.4M	524M	-
MobileNetV1	70.6%	4.2M	575M	123ms
MobileNetV2	71.4%	3.4M	300M	149ms

Xception is used as the backbone feature extraction network in DeepLabV3+. Compared to other lightweight neural networks, Xception and MobileNet series are characterised by the use of Depthwise Separable Convolution, while Xception increases the number of parameters to improve performance through Depthwise Separable Convolution, MobileNet compresses and reduces the number of parameters to improve speed through Depthwise Separable Convolution. The actual comparison of Xception and MobileNetV2 is shown in Table 2 [18].

Table 2. *Comparison ofXception and MobileNetV2 Parameter Numbers.*

Network	Xception	MobileNetV2
Size	88MB	14MB
Accuracy	0.79	0.71
Depth	126	88
Number of trainable parameters	2049	1281
Total number of parameters	20863529	2259265

Depthwiseseparabl convolution can be divided into tw parts, Depthwise Convolution, which convolves each channel of the input and then stacks the results, and Pointwise Convolution, which performs a 1×1 convolution of the stacked results. By using Depthwise separable convolution, MobileNet significantly reduces the number of parameters in the model and speeds up the convergence of the network.

The computational quantities for Depth wise separable convolution and Depth convolution are shown in equation (1)(2).

$$\begin{aligned} C_1 &= C_{DW} + C_{PW} \qquad (1) \\ &= D_k \times D_K \times \mathrm{M} \times \mathrm{D_f} \times \mathrm{D_f} + \mathrm{M} \times \mathrm{N} \times \mathrm{D}_f \times D_f \end{aligned}$$

$$C_2 = D_k \times D_k \times M \times N \times D_f \times D_f \qquad (2)$$

Where C1 represents the Depthwise separable convolution computation, C2 represents the Depth convolution computation, Dk represents the size of the convolution kernel, Df represents the height and width of the input matrix, M and N represent the Depth of the input and output feature matrices respectively. Theoretically, the Depthwise separable convolution computation is 1/9 to 1/8 of the Depth convolution.

MobileNetV2 [17] has two improved structures compared to MobileNetV1 [16]: Inverted residuals and Linear bottlenecks, as shown in Figure 1. The backbone part of the inverted Residuals structure first expands the number of channels by 1×1 convolution of the input, followed by 3×3 Depthwise separable convolution for feature extraction. Finally the number of channels is reduced by 1×1 convolution and the output of the residual edge part is directly connected to the input. The linear bottleneck structure uses a linear activation function, ReLU6, instead of the normal ReLU function. The ReLU6 activation function loses less information for tensors with lower channel counts and can reduce the accuracy and feature corruption of low-dimensional spatial information.

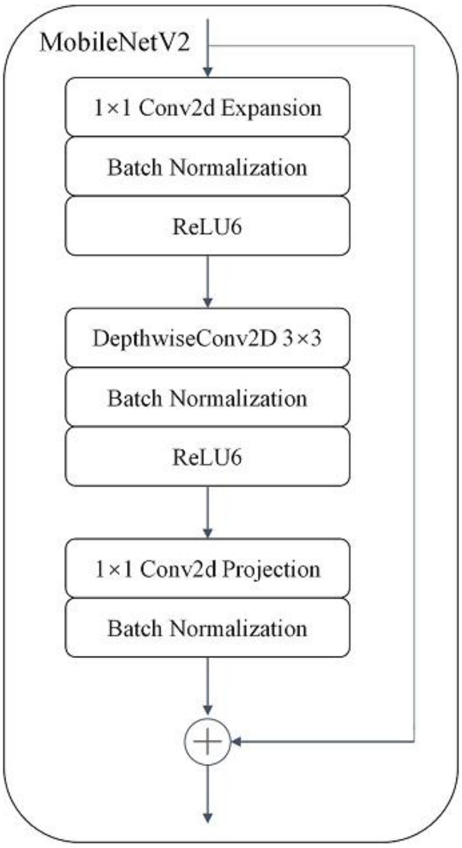

Figure 1. *MobileNetV2 network architecture.*

When using an embedded platform to deploy algorithms, it is necessary to minimize the computational effort and reduce the hardware computing power requirements while ensuring accuracy. Considering the hardware computing power limitation, training time, and accuracy, this experiment uses MobileNetV2, which has a good balance of accuracy and performance, instead of Xception as the backbone feature extraction network of DeepLabV3+.

2.2. DeepLabV3+

In semantic segmentation, there are disadvantages such as reduced resolution due to signal downsampling and spatial “insensitivity”, etc. Chen et al. proposed a series of DeepLab models based on Full Convolutional Networks to address this problem, as shown in Table 3 [19,20,21,22]. The DeepLabV3+[22] has been called the new peak of semantic segmentation, with better detail restoration capability in the delineation of object edges.

Chen et al [19] proposed DeepLabV1, which consists mainly of Deep Covolutional Neural Netsworks (DCNNS) and a cascade of Conditional Random Fields (CRFs). DeepLabV1 removes the final fully connected layer and the two pooling layers, while expanding the size of the perceptual field through Atrous convolution. To address the impact of the multi-scale problem of the target on the segmentation results, Chen et al [20] proposed DeepLabV2 in 2016, which sampled the Atrous spatial pyramid pooling (ASPP) with different sampling rates assigned to a given input feature to obtain multi scale image context information, refining the boundaries of the segmentation results.

DeepLabV3 [21] removes the CRFs structure which does not improve the accuracy much, introduces Multi-grid to set the expansion coefficients more reasonably, and adds a batch normalization (BN) layer to ASPP to capture the image context information. DeepLabV3 upgrades the backbone feature extraction network, ResNet-101, to Xception and proposes an Encoder-Decoder architecture. High-level features in the Encoder stage capture longer distance information, and the Decoder stage provides information to help recover the spatial dimension and detail of the target, optimising the boundary segmentation ofthe image.

Table 3. *DeepLab models.*

Model	DeepLabV1	DeepLabV2	DeepLabV3	DeepLabV3+
Main features	1: Expanding the receptive field through Atrous convolution 2: Improved access to detailed information through CRFs	1: Replacing VGG16 with ResNet-101 2: Multi-scale capture of image context information via ASPP structures	1: Improved ASPP structure with BN layer 2: Removed CRFs	1: Replacing ResNet-101 with Xception 2: Use of Encoder-Decoder architecture

The structure of DeepLabV3+ is shown in Figure 2. DeepLabV3+ uses Encoder to obtain the feature information of the input image, and Decoder to recover the details of the target and obtain the prediction results. The input image is passed through the deep convolutional neural network to obtain two feature layers, and the feature layer with higher semantic information is subjected to parallel Atrous convolution with a dilation rate of Rate=6, Rate=12, and Rate=18 for multi-scale information fusion, and is stacked with the feature layer with shallower semantic information entering the Decoder to complete the fusion and extraction of features.

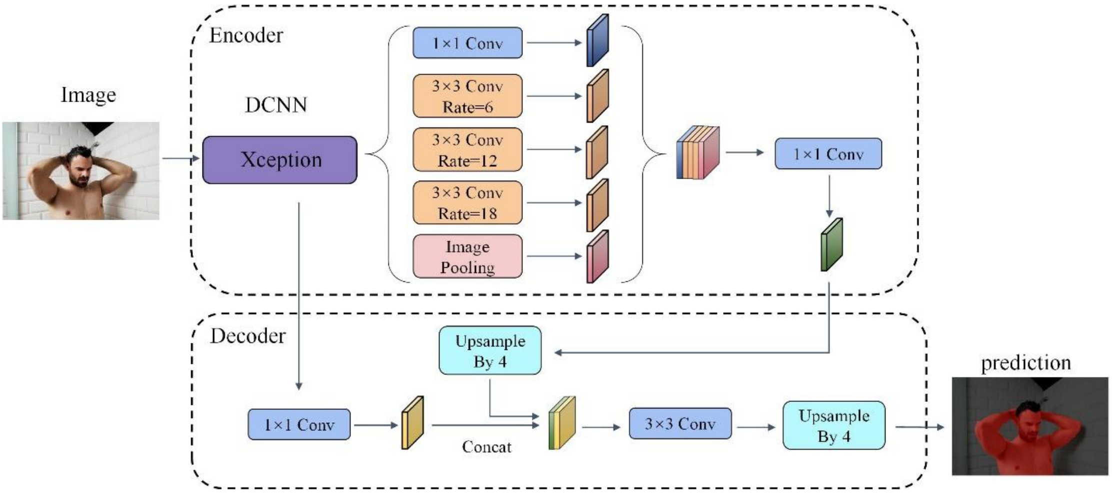

Figure 2. *The overall structure of DeepLabV3+.*

3. Test and Results

3.1. Experimental Data Setpreparation

In the field of deep learning-based image segmentation, large-scale annotated data is often required as training samples to prevent underfitting of experiments. However, in bathing scenes, the number of images containing skin is relatively limited. In this experiment, a total of 1104 images of skin involving different races, genders, ages, and shooting angles in a bathing scene with insufficient lighting, foam and water mist occlusion were collected manually and divided into a training validation set and a test set according to a 9:1 ratio. To overcome the shortage of experimental data, this experiment will augment the above dataset. In the bathing scene, multiple diffuse reflections and refractions of light occur, and the skin is obscured by water mist and bathing foam. In order to make the enhanced images better reproduce the original experimental scene, some of the images with higher definition are enhanced in five different ways according to the characteristics of the bathroom scene. Figure 3 shows the original image, horizontal flip, equal scale scaling, random brightness, unequal scale scaling, and average pooling in that order, expanding the database to 5520 images through data augmentation.

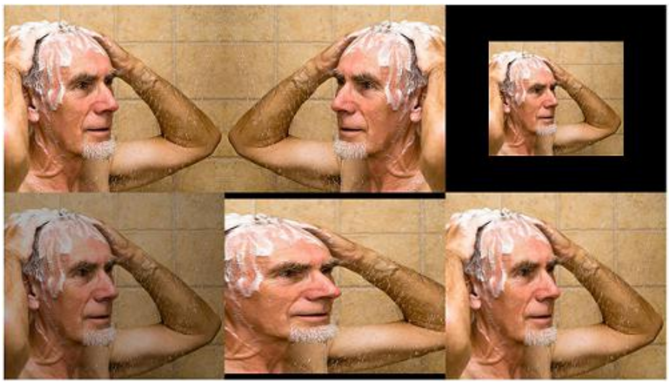

Figure 3. *Label visualisation images.*

Accurately annotated image samples can provide a large amount of detail and local features, which can significantly improve the segmentation accuracy and training efficiency of neural networks. Therefore, this experiment uses Labelme to accurately annotate the images at the pixel level after data augmentation, generating json files and converting them to label files in batch. The dataset annotation used in this model is divided into two parts, and consists of both the original dataset and the label file. The visualisation of the original image and the label file is shown in Figure 4.

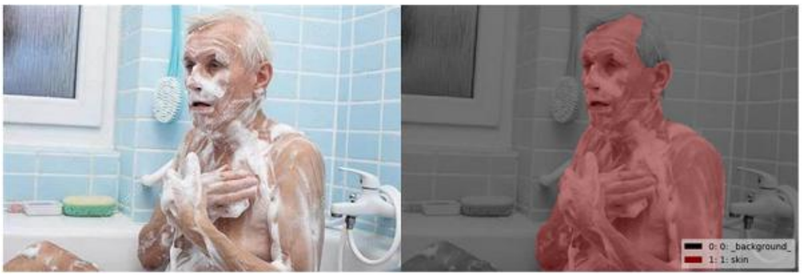

Figure 4. *Original image with labelfile visualization image.*

3.2. Experimental Settings

The hardware environment for this experiment uses Nvidia Geforce GTX TITAN Black, the software environment is Tensorflow-GPU2.2.0, and the language is python3.6.13. To measure the difference between the experimental prediction and the true value, the loss function was composed of the sum of the Cross-Entropy Loss Function and the Dice Loss Function. The Mean Intersection over Union (MIoU) and mean Average Precision (mAP) were calculated on the segmentation results of the self-built skin dataset to compare the precision and accuracy of the model, and the weight file size was compared to measure the memory footprint ofthe model.

$$Loss = Loss_C + Loss_D \qquad (3)$$

For binary classification experiments, the cross-entropy loss function is defined as shown in Equation (4), where y denotes the label of the sample, 1 for the skin region, and 0 for the non-skin region; $\hat{y}$ denotes the probability that the sample is predicted to be a skin region. Dice Loss is usually used to detect the similarity between samples and is defined as shown in Equation (5), where X is the set of sample predictions and Y is the set of sample true sets, with the numerator being the intersection of X and Y.

$$Loss_c = -(y \times \log(\hat{y}) + (1-y) \times \log\ (1-\hat{y})\ ) \qquad (4)$$

$$Loss_D = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \qquad (5)$$

MIoU and mAP are the more commonly used evaluation metrics in image segmentation currently. MIoU represents the degree of overlap between the predicted and true values, and mAP represents the proportion of correctly classified pixels within a category, as defined in equations (6) and (7), respectively.

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ij}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \qquad (6)$$

$$mAP = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \qquad (7)$$

Where: p_{ij} denotes False Positive (FP), i.e. the number of pixels where the true value of i is predicted to be j. Similarly, p_{ii} and p_{ji} denote True Positive (TP), False Negative (FN) respectively.

3.3. Privacy Risk Avoidance

Considering that the bathing aid system involves the identification and positioning of human skin, which may cause users' concerns regarding privacy issues, the system will adopt the following methods for privacy risk avoidance.

(1) The processing of the relevant video and RGB images is in the local embedded device, without the need to transfer user data to the server, eliminating the risk of data leakage from the source.

(2) The processing of the input video or images by the local embedded device is offline without networking, and the models adequately trained under the public skin dataset are migrated to the real application scenario.

3.4. Results

Although the data-enhanced database expands the data sample size for the experimental training, it still has a certain gap from the data volume of a standard large classical database. To further enhance the training effect of the experimental model, using the migration learning idea, the weight file of a large database with sufficient data volume and sufficient training is migrated to the present bathing application scenario with less data volume as pre-training weights, which can further increase the generalization ability of the model.

The loss curves of DeepLabV3+, the improved DeepLabV3+ model, and the Unet model, which is also based on MobileNet, are shown in Figure 5, Loss_X, Loss_M, and Loss_U denote the loss functions of the training set, Validation loss_X, Validation loss_M, and Validation loss_U respectively. Validation loss_X, Validation loss_M, Validation loss_U denote the loss functions of the three validation sets respectively. Through the experimental continuous adjustment of the deep learning parameters, all three models were stopped early within the set Epoch, ensuring the generalization ability of the models while preventing overfitting. The loss of both training and validation sets decreases continuously with increasing Epoch until it stabilizes and approaches 0, indicating that all three models basically converge.

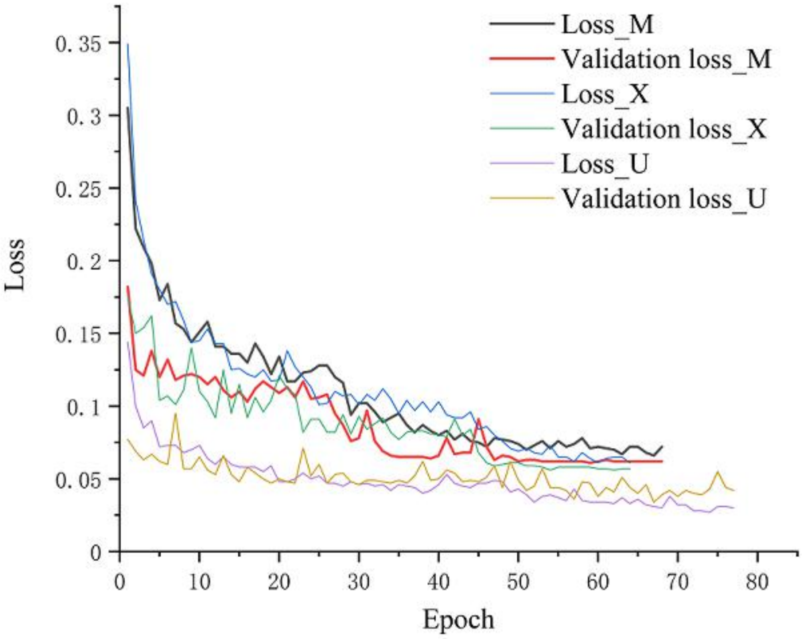

Figure 5. *Training set and validation set loss curves.*

The experimental results are shown in Table 4. In terms of several evaluation metrics for segmentation accuracy, the improved DeepLabV3+ has an MIoU of 91.57% and an mAP of 95.06% in the self-built skin dataset. This is 1.1% and 1.2% less MIoU and 0.79% and 1.59% less mAP than the original DeepLabV3+ and Unet, respectively. However, the improved DeepLabV3+ reduces the weight file size by 93% compared to the original DeepLabV3+ model and 87% compared to the Unet model, provided that the segmentation results are close to the true segmentation boundary. The reduction in model weight file size can improve the computational speed and lead to a wider range of practical applications in embedded devices.

Table 4. *Comparison of skin segmentation results of three models.*

Model	Regions	Backbone	size	MIoU	mAP
DeepLabV3+	Skin	Xception	158MB	92.76%	95.85%
	Background				
Unet	Skin	MobileNetV2	95MB	92.87%	96.65%
	Background				
Improved DeepLabV3+	Skin	MobileNetV2	10.9MB	91.57%	95.06%
	Background				

In addition to the quantitative experimental evaluation metrics, the results of skin segmentation were visualised and qualitatively analysed in this experiment by randomly selecting segmentation results from three test set images and partially zooming in to show them. The results are shown in Figure 6. The traditional skin tone model has low robustness of segmentation results as the light changes due to the presence of water mist and hair occlusion. Although the accuracy of the improved DeepLabV3+ model is reduced in many evaluation indicators, the overall segmentation boundary is smoother and does not show jaggedness, which restores the real boundary of the skin region to a greater extent. It can thus be demonstrated that the improved DeepLabV3+ model proposed in this paper achieves a good balance in terms of accuracy and performance.

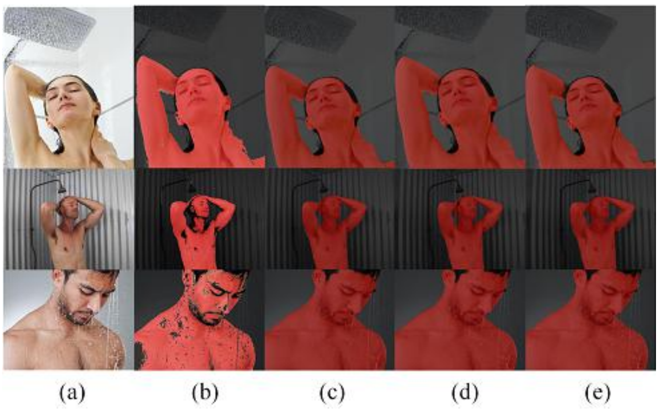

Figure 6. *Comparison of the skin segmentation result(a:The segmentation results of the original image; b:The skin tone model; c: DeepLabV3+; d: the Unet model; e: The improved DeepLabV3+ model).*

4. Conclusions

Most of the current skin segmentation and localisation by traditional computer vision methods are less robust and more sensitive to noise. The skin in the bathroom scene is often obscured by water mist and foam, there are background areas that are similar in colour to the skin, all of which can cause greater interference to the traditional algorithm. There is still much room for improvement in the detection results.

In this study, the background and skin regions were segmented at the pixel level using the DeepLabV3+ algorithm in an environment with a lot of noise interference, and the segmentation results far exceeded traditional learning methods. In the self-built skin database, by using migration learning ideas and data augmentation methods, the DeepLabV3+ model obtains better segmentation results with MIoU as high as 92.76%, and the segmentation results have clear boundaries without jaggedness and have better segmentation ability in details.

With the continuous development of deep learning and high performance servers, the number of parameters and computation of neural networks is becoming increasingly large, and the network structure of the models is towards complexity. The accuracy of the theoretical models is constantly refreshed, but it also poses certain difficulties for the deployment of embedded devices in practical engineering applications. In this experiment, by lightening the DeepLabV3+ neural network, the weight file size is greatly reduced, compared with the original DeepLabV3+ and Unet models, the weight file size is reduced by 93% and 87% respectively, reducing the number of neural network parameters and the time required for training, and solving the problem of computing power limitation of embedded devices. This study bridges the gap of skin segmentation in bathing scenarios, advances the development of age-appropriate personal cleaning devices and intelligent devices for the elderly and disabled, and empowers deep learning algorithms with more possible application spaces. In future work, by improving the DeepLabV3+ model for segmenting skin areas, the location information of the skin areas will be extracted and provided to the robotic arm as control information for application to the intelligent control of autonomous bathing, and we plan to continue to enrich the database for bathing scenarios, make improvements in the structure of the algorithm model, and continuously reduce the performance required by the hardware, so that deep learning can be applied to a wider range of applications.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

This work was supported by the National Natural Science Foundation of China, grant number 62073224.

References

[1] Lin, B. Active response to population aging: connotation, goals and tasks. *Chinese Journal of Population Science,* 2021, 3, 42-55.

[2] Liu, H.L. Trends of Population Aging in China and the World as a Whole. *Scientific Research on Aging,* 2021, 9(12), 1-16.

[3] Anita, A.; Colin B.; Parikh, Toral, M. Associations Between Activities of Daily Living Independence and Mental Health Status Among Medicare Managed Care Patients. *Journal of the American Geriatrics Society,* 2020, 6.

[4] Phillip, W.J.; Miriam, G.R. The lived experience of bathing adaptations in the homes of older adults and their carers (BATH-OUT): A qualitative interview study. *Health & social care in the community,* 2019, 6.

[5] Manti, M.; Pratesi, A.; Falotico, E. Soft assistive robot for personal care of elderly people. In *IEEE International Conference on Biomedical Robotics and Bi*omechatroni*cs (BioRob),* 2016, 833-838.

[6] He, Z.M; Yuan, F.; Chen, D.S. Ergonomic Design of Multi-functional Bathing Robot. In *IEEE International Conference on Real-time Computing and Robotics (RCAR),* 2019, 580-585.

[7] Dong, Q.; Miao, X.G. Research on the improvement of reclining bathing robot. *Technique and application,* 2017, 4, 28-32.

[8] Csurka, G.; Perronnin, F. An efficient approach to semantic segmentation. *International Journal of Computer Vision,* 2011(2).

[9] Zhang, X.Y.; Zhao, H.T. Hyperspectral-cube-based mobile face recognition: A comprehensive review. *Information Fusion,* 2021, 74, 132-150.

[10] Rehman, A.; Khan, M.A.; Mehmood, M. Microscopic melanoma detection and classification: A framework of pixel-based fusion and multilevel features reduction. *Microscopy Research and Technique,* 2020, 83(4), 410-423.

[11] Suliman, A.S.; Omran, A.A. Computer vision assisted human computer interaction for logistics management using deep learning. *Computers & Electrical Engineering,* 2021, 96, 107555.

[12] Li, P.; Yu, H.L. Survey of object detection algorithms based on two classification standards. *Application Research of Computers,* 2021, 38(9), 2582-2589.

[13] Tian, X.; Wang, L.; Ding, Q. Review of image semantic segmentation based on deep learning. *Journal of Software,* 2019, 30(2), 440-468.

[14] Krizhevsky, A.; Sutskever, I.; Hinton, E.G. ImageNet classification with deep convolutional neural networks. *Association for Computing Machinery,* 2012, 60, 84-90.

[15] Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR),* 2017, 1800-1807.

[16] MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. Available online: https://arXiv:1704.04861 (accessed on 17 April 2017).

[17] Sandler, M.; Howard, A.; Zhu, M.L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2018, 4510-4520.

[18] Barboza, D.; Silva, C.; Alexander, N.; Silva, A.; Barroso, G. Convolutional Neural Networks Using Enhanced Radiographs for Real-Time Detection of Sitophilus zeamais in Maize Grain. *Foods,* 2021. 10(4), 879.

[19] Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. Available online: https://arxiv.org/abs/1412.7062 (accessed on 22 December 2014).

[20] Chen, L.C.; Papandreou, G.; Kokkinos, I. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans on Pattern Analysis and Machine Intelligence,* 2016, 40(4), 834-848.

[21] Rethinking Atrous Convolution for Semantic Image Segmentation. Available online: https://arxiv.org/abs/1706.05587 (accessed on 5 December 2017).

[22] Encoder-Decoder with Atrous Separable Convolution for Semantic Image Available online: https://arxiv.org/abs/1802.02611 (accessed on 22 August 2018).