Transfer Learning Based Fruits Image Segmentation for Fruit-picking Robots

Yongfu He, Fangfang Pan, Baoyu Wang, Ziqing Teng and Jianhua Wu

April 23, 2020

Transfer Learning Based Fruits Image Segmentation for Fruit-picking Robots

Yongfu He, Fangfang Pan*
Department of Information Engineering
Gongqing College, Nanchang University
Jiujiang 332020, China
*Corresponding author: 155295223@qq.com

Baoyu Wang, Ziqing Teng, Jianhua Wu
Department of Electronic Engineering
School of Inform. Eng., Nanchang University
Nanchang 330031, China

Abstract—**It is an important prerequisite for a fruit-picking robot to accurately segment and locate the object in fruit images. However, image segmentation by manually selected features or deep learning-based approaches is a troublesome task. It requires a long time and a large number of annotated images for the model to be trained. In this study, transfer learning is used so that the learned parameters of a pre-trained convolutional neural network can be used as the initial settings in the new task. Three networks, Mobilenet_v2, Resnet_v1_50_beta and Xception_65, are used as backbone networks, which were used in the well-known semantic image segmentation model—DeepLab. The proposed transfer learning-based fruits image segmentation not only alleviates the stringent need of a large image dataset, but also saves much time for training. Experimental results show that the Xception_65 based network has the best performance in terms of the segmentation metric of mean intersection over union. A high-precision instance fruits segmentation guarantees subsequent accurate locations of fruit images for fruit-picking robots, which is of great significance for intelligent agriculture.**

Keywords-tranfer learning; semantic segmentation; instance segmentation; convolutional neural network; fruit image segmentation

I. INTRODUCTION

It is an important prerequisite for a fruit harvesting robot to accurately segment and locate the object in fruit images. Traditional image segmentation methods can be categorized as those based on the threshold, based on region and based on edges. Classic threshold-based methods include Otsu's algorithm [1] et al. Regions based image segmentation methods include K-means clustering, graph cut method and so on [2–5].

Early in 2008, Yuan proposed a near infra-red based fruit segmentation method for cucumbers in a greenhouse environment [6], where the shape difference between the background and fruits is used for two dynamic thresholding segmentations, alleviating the influence on image quality by poor environment and segmenting efficiently the cucumber object. However, it is necessary to pre-estimate the ratio *P* of the area of mature fruits over the total image according to the distance from the robot to the cucumber plants. This is not good for a fully automated harvest. In 2009, Li et al. proposed an image segmentation of cucumbers based on color and textual features [7], in which the color is used to remove irrelevant objects, then the grey-level co-occurrence matrix is used to extract textures, and finally the entropy and energy are obtained as two textural features to successfully the solve the segmentation problems of fruits and background which are similar in color. However, the discrimination of the entropy and energy with respect to fruits and background is not so high, poorly impacting the segmentation accuracy. Wang conducted an improved wavelet transform for fruit images to normalize the illuminance on the surface of fruits [8], performed the image enhancement based on retinex and finally obtained the segmentation results by means of K-means clustering. Experimental results show that the algorithm's robustness relative to light change in color fruits segmentation, as well as a super performance both in segmentation accuracy and efficiency.

The aforementioned algorithms require in common manually selected features such as *P* parameter, entropy, energy, grey-level co-occurrence matrix, etc. Theoretically, it is difficult to verify the optimal features and even enumerate them with experiments. Recent years, deep learning (DL) provides a new and effective way for automatic feature selection in image segmentation and recognition [9]. W. Ding, in his dissertation, performs the pepper detection and recognition by using convolutional neural networks (CNN) [10], but the thresholding is still used in the stage of segmentation, without totally giving up the manual classification features. An Indian researcher, R. Kexture proposed a deep semantic segmentation architecture (MangoNet) to detect and count mangoes in an open orchard [11]. The results demonstrate the robustness of detection for multiple factors such as scale, occlusion, distance and illumination conditions. Recently, S. Chen proposed a novel fully convolutional neural network model (Nv-Net) for low-resolution infrared image segmentation in weak illumination natural environments [12]. Experimental results show that the proposed method has the flexibility to segment the arbitrary input images on several public datasets, such as PASCAL VOC and ADE20K, and achieves the best segmentation performance in the low illumination environment.

The key advantage of DL based methods is the automatic feature extraction from the raw data and the capability to handle computation-intensive tasks. However, it requires a long time and a large number of annotated images for the model to be trained. For example, in [10], 11,096 image patches of size 200×200 are used for training and 1500 image patches are used for testing. In [11], the size of model weights is up to 375 M. The Nv-Net is trained by the TensorFlow framework on a single Nvidia GeForce GPU. The training requires approximately 79 h to learn a good model. Unfortunately, in practice, there is seldom such a large number of images included in a training set. The lack of annotated images often hinders the training procedure. Therefore, in this study, transfer learning is used to eliminate the need for a vast number of annotated fruit image datasets by transferring the learned features of a pre-trained CNN as initial weights and fine-tuning the whole network using our new dataset available. By the way, the model being trained converges much faster based on a relatively small training fruit image dataset. The proposed method achieves state-of-the-art segmentation results in terms of the mean intersection over union (MIoU) without need a very large image dataset.

The manuscript is organized as follows. In section II, some preliminaries are reviewed as fundamentals. Section III gives the proposed transfer learning-based fruit image segmentation model. Experimental results are given in section IV and finally, conclusions are drawn in section V.

II. Preliminaries

A. *Transfer learning*

Given a learning task in a source domain and a learning task in a target domain, transfer learning aims to help improve the learning of the target predive function using the knowledge in the source domain and the learning task in the source domain, when the two tasks are different or the two domains are not the same.

Early in 2010, Pan and Yang published a survey on transfer learning [13]. In many machine learning scenarios, it is assumed the training set and test set must follow the same distribution. However, this is not always true in practice. What is more, in many real-world applications, it is expensive or impossible to acquire sufficient training data. In such cases, transfer learning between task domains would be a solution. Pan categorized transfer learning into inductive transfer learning, transductive transfer learning and unsupervised transfer learning.

In the inductive transfer learning, the target task is different from the source task. Some labelled data in the source domain are needed to induce an object predictive model into the target domain. In the transductive transfer learning, the source and target tasks are the same, while the source and target domains are different. No annotated data in the target are available while a lot of annotated data in the source domain are available. Finally, in the unsupervised transfer learning, the target task is merely related to the source task. It aims to solve problems such as clustering in the target domain. There are no labelled data in both domains.

Transfer learning is feasible. In 2014, Yosinski showed how features are transferable in deep neural networks [14]. Although the transferability of features decreases as the distance between the source task and the target task increases, transferring features even from distant tasks can be better than using random initial weights. Donahue also verified that features extracted from the activation of a deep convolutional network trained in a fully supervised manner on a large set of object recognition tasks is able to be transferred to new general tasks [15].

B. *Convolutional neural networks*

Convolutional neural networks （CNN）have become very popular since AlexNet [16] won the ImageNet Challenge. Deeper and more complemented CNNs have been achieved high accuracy and efficiency in computer vision task including semantic image segmentation and object recognition. A lot of CNN architectures have been emerged, among which the following three networks are used in the study and briefed as follows.

- Resnet. K. He presented a residual learning framework to ease the training of deep networks [17]. The learning residual functions with reference to the layer inputs are formulated. Comprehensive experiments show the residual network is easier to optimize and can gain accuracy from the considerable residual depth.
- Xception. Xception is a CNN architecture based on entirely on depthwise separable convolution layers [18]. It has a linear stack of 36 convolutional layers for feature extraction. This makes the architecture very easy to define and modify. An open source apps module using Keras and TensorFlow can be used for implementation.
- Mobilenet. The general trend has been to make deeper and more complicated networks in order to achieve higher accuracy but often without a high efficiency in terms of size and speed. MobileNets are built to reduce the computation in the first few layers. Howard presented a class of efficient models called MobileNets [19]. Two simple global hyperparameters that efficiently trade off between latency and accuracy. The effectiveness of MobileNets is verified across a wide range of applications such as object detection, fine-grain classification and large-scale geo-localization.

C. *DeepLab*

DeepLab is a series of semantic image segmentation networks. In 2015, L. Chen proposed a semantic image segmentation with deep convolutional nets and fully connected conditional random field [20], aiming to the task of pixel-level classification. Afterwards, with VGG-16, DeepLabV1 obtained an MIoU of 71.6% in the image database PASCAL VOC-2012 [21]. DeepLabV2 [22] is an improved version of DeepLabV1 with the backbone network of ResNet101 [17], achieving an MIoU of 79.7% in PASCAL VOC-2012. To solve the problem of multiscale object segmentation, DeepLabV3 [23] uses a cascaded or parallel atrous convolution to adjust receptive field. DeepLabV3 obtains an MIoU of 85.7% without a dense post-processing.

In 2018, Chen proposed an encoder-decoder structure with atrous separable convolution for semantic image segmentation on the basic of DeepLabV3, known as DeepLabV3+ [24]. Its backbone network is Xception [22]. DeepLabV3+ achieves the MIoUs of 89.0% and 82.1% in the databases of PASCAL VOC-2012 and Cityscapes, respectively.

III. TRANSFER LEARNING BASED INSTANCE SEGMENTATION OF FRUITS IMAGES

This study uses transfer learning to segment fruit images with three kinds of backbone networks in target domain. As shown in figure 1, the specific implementation steps are described as follows:

- Acquirement of fruits images containing eggplant, luffa, pepper, using camera in a greenhouse environment.
- Pre-processing of the images, such as data augmentation and image denoising.
- Labelling of the image objects using an image annotation software, named LabelMe, to label the objects. The results annotated are: the background is set to zero, and other three kinds of fruits as 1, 2, 3.
- Partition of all images into the training set and the test set. The annotated images are divided accordingly into two parts, in which one part corresponding to the training set is used for training, and the other part corresponding to the test part is used for model evaluation.
- Load the weights of pre-trained DeepLabV3+ from TensorFlow as the initial settings, which are transferred to the Training module (as shown in figure 1).
- After 100,000 training steps for the model in target domain, the model is frozen and is fed into the Testing module, where the networks Mobilenet_v2, Xception_65 and Resnet_v1_50_beta are chosen as backbone nets.
- Evaluation of fruit segmentation effect. The performance is evaluated using the MIoU and training steps per second. Meanwhile, three different backbone networks, Mobilenet_v2, Xception_65 and Resnet_v1_50_beta, are used for comparison of test results and system's performance evaluation on representative test images.

IV. EXPERIMENTS AND ANALYSES

The fruits images used in the semantic instance segmentation experiments were acquired in the vegetable production centre in the county of De'an, Jiujiang, China. The model runs in a hardware environment of Intel(R) Core(TM) i7-8700 CPU@ 3.2GHz，with the memory of 16 GB and two graphic adapters of GeForce RTX 2080Ti. The workstation's operating system is Ubuntu 16.04. The program language and the deep learning symbolic library are Python 3.5 and TensorFlow 1.13.1, respectively. The image set includes 3716 images with details shown in table 1.

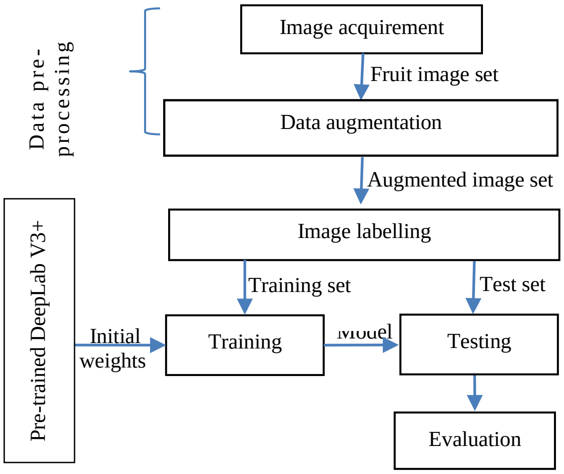

Figure 1. Transfer learning-based instance segmentation of fruits images

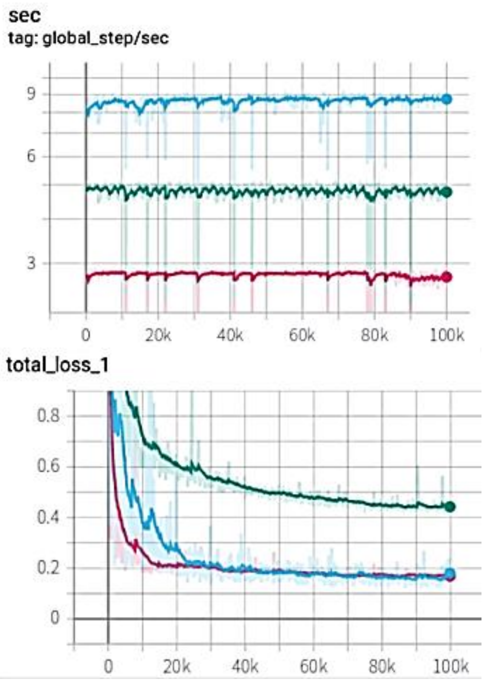

Figure 2. Training process of three networks.

Three representative networks, Mobilenet_v2, Xception_65, and Resnet_v1_50_beta, are chosen as the backbone nets with the official pre-trained weights as initial weights for training in our experiments. The total number of training steps is set to be 100,000, the base learning rate is set to be 0.0007, 'Ploy' is chosen as the strategy of adjustment of the learning rate, the weight decay parameters are set to be 0.00004 and 0.0001 for Mobilenet_v2, Xception_65 and Resnet_v1_50_beta, respectively. The atrous rate for atrous spatial pyramid pooling and the ratio of input to output spatial resolution (output stride) are set to default values for Mobilenet_v2; they are respectively set to be [6, 12, 18] and 16 for Xception_65, Resnet_v1_50_beta. For all the three networks, the stride for decoder output is set to be 4. The image crop sizes are set to be 513×513 and 1525×1200 during training and evaluation, respectively. The batch size is set to be 4 [24].

Table 1. Statistics for image data

	Pepper	Eggplant	Luffa
Training set	714	606	500
Test set	758	614	524
In total	1472	1220	1024

Table 2. Segmentation for different backbone nets.

Network	MIoU (%)	Steps/second	Training time (h)
Mobilenet_v2	72.3	8.7	3.19
Xception_65	94.9	2.8	9.92
Resnet_v1_50_beta	93.9	4.8	5.79

In DeepLab, different network structure can result in different MIoU and different time consumptions. The intersect over union (IoU) is defined as $\text{IoU} = A_r \cap A_m / A_r \cup A_m$ where A_r is the region regarded as the object by the proposed algorithm, and A_m is the region of ground truth. The mean intersection over union (MIoU) is the average of all IoUs over the test image set and reflects the accuracy of the used semantic segmentation architectures. Table 2 shows the comparison of segmentation results with the three backbone networks used in this article, where the MIoU reflects the accuracy of the used semantic segmentation structures, and the training steps per second can reflect the speed of training process. The Training time of three backbone are shown in table 2. Thanks to the use of transfer learning, the model can reach a high MIoU in only about ten hours of training. It is known from table 2 that Xception_65 has the longest training time in three networks, but the highest MIoU value.

The training process visualized in TensorFlow is shown in figure 2, in which the blue, red and green represent the Mobilenet_v2, Xception_65 and Resnet_v1_50_beta networks, respectively. It can be seen that Xception_65 (in red) has the least number of training steps per second, meaning the most time-consuming, but the total loss decreases the fastest, thus the final total loss is the smallest.

Figure 3 gives an example of segmentation for an image containing an eggplant, a pepper and a luffa, tested for all the three networks. The image crop size is set to be 513×513; the scale to resize images for inference is set to be 1.0. The original image is shown in figure 3. The segmentation results are shown in figure 4, figure 5, and figure 6 for Mobilenet_v2, Resnet_v1_50_beta and Xception_65, respectively. In figures 4, 5, and 6, the green, yellow and red represent the eggplant, luffa and pepper, respectively. From these figures, it can be seen that Mobilenet_v2 and Resnet_v1_50_beta mistake a small portion in the middle eggplant as background, several pixels at the bottom of the luffa are erroneously assigned to pepper by Mobilenet_v2, and Resnet_v1_50_beta mistakes the bottom tip of the luffa as background. These results also verify the best segmentation performance of Xception_65, which not only classifies all pixels of the objects correctly, but also assigns the pixels of a little eggplant behind a big eggplant correctly, which is easily ignored by human. This is in agreement with the above conclusion that Xception_65 performs the best in terms of MIoU.

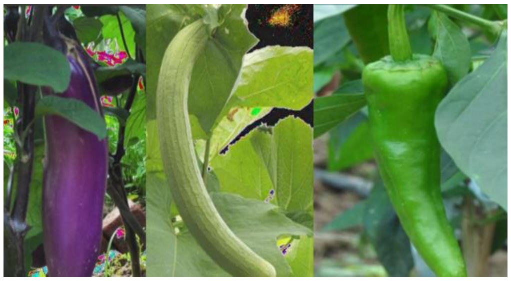
Figure 3. Original image to be segmented

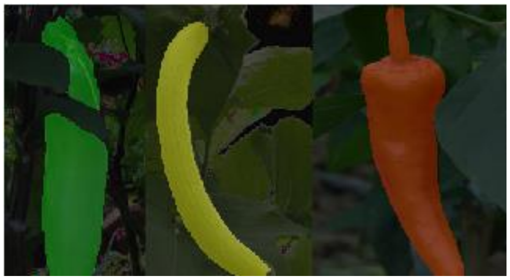
Figure 4. Segmentation result with Mobilenet_v2

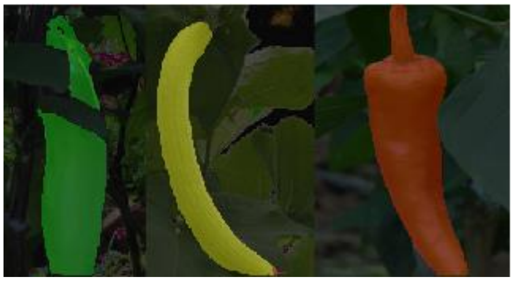
Figure 5. Seg. result with Resnet_v1_50_beta

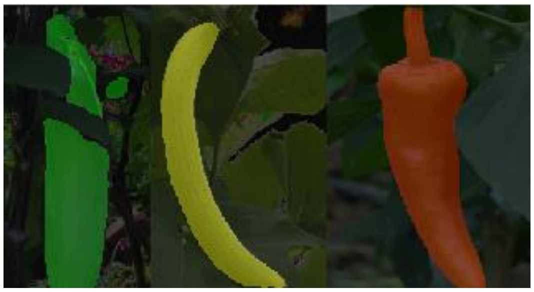
Figure 6. Segmentation result with Xception_65

V. CONCLUSION

In this study, transfer learning-based fruit image segmentation is implemented, which not only alleviates the stringent need of a large image dataset, but also saves much time for training. Experimental results show that the instance fruit image segmentation achieves state-of-the-art results in terms of the mean intersection over union. Furthermore, it is verified that the Xception_65 based structure obtains the highest MIoU value in three backbone networks. The model can reach a high MIoU in only about ten hours of training. The segmentation results with Xception_65 are satisfactory and apparently superior to results with other two.

A high-accuracy semantic segmentation ensures an accurate localization and efficient harvest for robots, which is of great significance to smart agriculture.

References

[1] N. Otsu, "A threshold selection method from gray-level histogram," IEEE Transactions on Systems, Man, and Cybernetics, 1979, vol. 9, no. 1, pp. 62–66.

[2] L. Zhu, Y. Gao, A. Yezzi and A. Tannenbaum, "Automatic segmentation of the left atrium from MR images via variational region growing with a moments-based shape prior," IEEE Transactions on Image Processing, 2013, vol. 22, no. 12, pp. 5111–5122.

[3] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman and A. Wu, "An efficient k-means clustering algorithm: analysis and implementation," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, vol. 24, no. 7, pp. 881–892.

[4] J. Shi, J. Malik, "Normalized cuts and image segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, vol. 22, no. 8, pp. 888–905.

[5] D. Hochbaum, "Polynomial time algorithms for ratio regions and a variant of normalized cut," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, vol. 32, no. 5, pp. 889–898.

[6] T. Yuan, W. Li, Y. Tan, "Research on information acquiring technology for cucumber harvesting robots in a greenhouse environment," Proc. of Annual Conf. of Chinese Agricultural Mechanical Society, 2008, pp. 898–901, Jinan, China. (in Chinese)

[7] E. Li, Z. Li and W. Liu, "Image segmentation of cucumbers based on color and textual features," Optical Technology, 2009, vol. 35, no. 4, pp. 529–31. (in Chinese)

[8] C. Wang, Y. Tang, X. Zou, et al. W. Situ and W. Feng, "A robust fruit image segmentation algorithm against varying illumination for vision system of fruit harvesting robot," Optik, 2017, vol. 131, pp. 626–631.

[9] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," Nature, 2015, vol. 521, no. 7553, pp. 436–444.

[10] W. Ding, "Detection and recognition of pepper based on convolutional neural network and machine vision," Dissertation for Master Degree, 2017, Tianjin University, Tianjin, China. (in Chinese)

[11] R. Kestur, A. Meduri and O. Narasipura, "MangoNet: A deep semantic segmentation architecture for a method to detect and count mangoes in an open orchard," Engineering Application of Artificial Intelligence, 2019, vol. 77, pp. 59–69.

[12] S. Chen, Z. Chen, X. Xu, N. Yang and X. He, "Nv-Net: Efficient infrared image segmentation with convolutional neural networks in the low illumination environment," Infrared Physics and Technology, 105 (2020) 103184.

[13] S. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359.

[14] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" In Advances in Neural Information Processing Sustems 27 (NIPS'2014), NIPS Foundation, 2014.

[15] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," In ICML, 2014. http://arxiv.org/abs/1310.1531v1 [cs.CV] 6 Oct 2013.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Communications of the ACM, June 2017, vo. 60, no. 8, pp. 84–90.

[17] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," IEEE International Conference on Computer Vision and Patter Recognition, vol. 1, pp. 770–778, 27–30 June 2016, Las Vegas, USA.

[18] F. Chollet. "Xception: Deep learning with depthwise separable convolution," IEEE International Conference on Computer Vision and Patter Recognition, vol. 1, pp. 1800–1807, 26–27 July 2017, Honolulu, USA.

[19] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto and H. Adam, "MobileNets: efficient convolutional neural networks for mobile vision applications," arXiv: 1704.04861v1 [cs.CV] 17 Apr 2017.

[20] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "Sematic image segmentation with deep convolutional nets and fully connected CRFs," Proceedings of International Conference on Learning Representations, May 7–9, 2015, San Diego, USA.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556v6 [cs.CV] 10 Apr 2015.

[22] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, vol. 40, no. 4, pp. 834–848.

[23] L. Chen, G. Papandreou, F. Schroff and H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv:1706.05587v3 [cs.CV] 5 Dec 2017.

[24] L. Chen, Y. Zhu, G. Papandreou, F. Schroff and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," arXiv:1802.02611v3 [cs.CV] 22 Aug 2018.

Authors' background (This form is only for submitted manuscript for review)

Your Name	Title	Affiliation	Research Field	Personal website
Yongfu He 何勇福	Lecturer	Nanchang University	Signal Processing and Patter recognition	
Fangfang Pan 潘芳芳	Lecturer	Nanchang University	Signal Processing and Patter recognition	
Baoyu Wang 汪保玉	Master student	Nanchang University	Image Processing and Machine learning	
Ziqing Teng 滕梓晴	Master student	Nanchang University	Image Processing and Machine learning	

Jianhua Wu 吴建华	Professor	Nanchang University	Patter Recognition and Machine learning	