Journal of Computers Vol. 33 No. 1, February 2022, pp. 55-62
doi: 10.53106/199115992022023301006

A Modified Residual Network Based on Multi-scale Segmentation for Aerobics Motion Image Recognition

Xingxing Dai*

Department of Arts and Sports, Henan Technical College of Construction, Zhengzhou 450000, China
zxcvfdsa5024@foxmail.com
Received 27 September 2021; Revised 15 October 2021; Accepted 27 October 2021

Abstract. Image recognition is an important field in artificial intelligence, it makes use of the computer to conduct image processing, analysis and understanding to recognize a variety of different objects. And it uses a series of enhancement and reconstruction methods to effectively improve the image quality. Traditional deep Convolutional Neural Network (DCNN) not only improves the recognition accuracy, but also reduces the recognition speed. How to improve the speed while maintaining the accuracy has become an important direction in image recognition. In this paper, we propose a modified Residual network based on multi-scale segmentation for aerobics motion image recognition. The new residual network has the characteristics of shorter network length and faster recognition speed. First, it reduces the length of the network and gets a new residual network with seven layers. Then, combining with the multi-scale segmentation method, an image recognition residual network is obtained. Finally, experiments on the CIFAR10 dataset, the results show that the proposed new motion image recognition method has better recognition accuracy and faster recognition speed.

Keywords: Residual network, multi-scale segmentation, motion image recognition, DCNN

1 Introduction

The emergence of deep convolutional neural networks (DCNN) has made a great contribution to solve complex computer vision tasks, and it has been widely used in image classification, object detection and instance segmentation [1-3]. How to design an efficient network model becomes the key to improve the performance of deep convolutional neural networks. Traditional convolutional neural networks, such as AlexNet [4] and VGGNet [5] use a simple activation function and convolution structure to make multi-scale feature data-driven learning. The residual structure ResNet [6] could enable deeper networks to perform effective learning. WRN [7] improved the recognition accuracy by increasing the network width of the residual network. Res2Net [8] proposed a multi-scale segmentation method for the residual structure to improve the recognition accuracy. HS-ResNet [9] improved the multi-scale segmentation method, and proposed a network model with smaller parameters and faster recognition speed.

By increasing the depth and width of the network, rich feature information can be generated. By using multi-scale segmentation method, feature information can be fully utilized and redundant feature information can be reduced. However, increasing the depth of the network and multi-scale segmentation will also reduce the recognition speed of the network. How to improve the recognition speed while ensuring the recognition accuracy has become an important goal of designing an efficient network model [10]. In this paper, the multi-scale segmentation method s improved, and a simple multi-scale segmentation method is proposed. By reducing the number of network layers and increasing the width of the network, the recognition speed is faster under the condition that the recognition accuracy and parameters of the network are similar.

First of all, this paper combines the residual structure of ResNet-D [11] and WRN to improve the ResNet. Then, by reducing the number of network layers, the network model with seven layers is obtained. Finally, the multi-scale segmentation method of HS-ResNet is improved. And an image recognition residual network (SSRNet) based on multi-scale segmentation is obtained. The recognition accuracy of this network is similar to PyramidNet [12], but the recognition speed is faster than PyramidNet. The main contributions of this paper are summarized as follows:

A new multi-scale segmentation method is proposed, and the recognition accuracy and recognition speed of the network model are greatly improved.

The recognition speed is improved by increasing the number of channels and shortening the network length in the shallow layer network.

* Corresponding Author

2 Related Works

The emergence of residual networks makes deep learning widely used in all aspects of production and life. The residual network learns the fitting function by solving the difference between the predicted value and the observed value, so that the deep network can learn the parameters effectively. In recent years, the improved network based on the residual network has greatly improved the recognition accuracy of the network.

2.1 ResNet

In deep learning, the increase of network layers will generally consume more computing resources, the network model will appear over-fitting, gradient disappearance, and gradient explosion problems. As the number of network layers increasing, the network model will be degenerated, that is, as the number of network layers increasing, the training loss gradually decreases. Then it tends to saturation. If the depth of the network is increased, the training loss increases instead. When the network model degenerates, the shallow network model can achieve better training effect than the deep network model. At this time, if the low-level feature information is transformed to the high-level, the effect cannot be worse than the shallow network model. Residual network is born by using directly mapping to connect different layers of the network model.

When the output dimension is not equal to the input dimension, ResNet needs to increase the dimension of the input dimension. In ResNet, a 1×1 convolution with step size 2 is directly used to increase dimensions. When the input image size is halved, the feature information will be lost. ResNet-D is based on ResNet. Before increasing the residual structure dimension operation, 2×2 uniform pooling with a step size 2 is added, and then 1×1 convolution with a step size 1 is used for increasing dimension. The recognition accuracy of network is greatly improved.

2.2 WRN

With the increase of network layers, training deep convolutional neural networks has problems such as gradient disappearance and gradient dispersion. The experimental results also show that the recognition accuracy improvement brought by the deeper network model is not obvious, but the recognition speed needs to be reduced [13]. Is the network deeper and narrower, the network is better? Or is it feasible to train a wider and shallower network as long as the network parameters are guaranteed? WRN proposed a convolutional neural network based on the extended channel number learning mechanism, by using the shallower network model, the recognition accuracy is comparable to that of the deep network model, and the recognition speed is faster.

WRN proves that the accuracy can be improved by increasing the network width. When the parameters are the same, the speed of WRN is faster. Like the residual network, when the parameter is too large, there will be problems such as gradient disappearance and gradient dispersion.

2.3 Res2Net

Multi-scale features presentation is very important for vision tasks. The latest developments of convolutional neural networks continue to show stronger multi-scale representation capabilities and achieve consistent performance improvements in a wide range of applications [14-16]. Most of the previous methods represent multi-scale features in a hierarchical feature pyramid [17]. Res2Net adds small residual blocks to the original residual unit structure to increase the receptive field range of each layer, and represents multi-scale features with smaller fine-grain sizes. The intermediate main convolution of residual element structure is changed from single branch to multi-branch.

The Res2Net module has a simple structure and excellent performance. It reveals a new scale dimension in addition to the three dimensions of convolutional neural networks (depth, width, and cardinality). The Res2Net module can be easily combined with other modules, the feature extraction ability is more powerful, and it does not increase the computational load.

2.4 HS-ResNet

HS-ResNet improves the convolution and connection of multi-scale segmentation of feature maps, which improves the recognition speed while improving the recognition accuracy. It mainly considers the following three issues: (1) How to avoid redundant information in feature graph; (2) How to let the network learn a stronger

feature expression without increasing the computational complexity; (3) How to get better recognition accuracy while maintaining a faster recognition speed. Based on the three problems, the designed HS-Block module is to generate multi-scale features.

The Res2Net method increases the computational complexity when the number of channels is large. So, are all feature maps necessary? The experimental results of GhostNet [18] show that some feature maps can be generated from existing feature maps. HS-ResNet partially connects the feature graph obtained by convolution of S2 group to S3 group, which realizes feature graph reuse and reduces computational complexity.

HS-ResNet proposes the HS-Block module, which can efficiently extract multi-scale features. It has achieved the most advanced recognition performance on multiple vision tasks (such as image classification, target detection and instance segmentation). HS-Block has the characteristics of plug and play, which can be easily embedded into the existing network and improve the recognition performance.

3 Proposed Motion Image Recognition Method

The multi-scale segmentation method of HS-ResNet enables different groups of feature information to enjoy different scales of receptive fields. In the feature information connected from the previous layer, the number of convolutions is less, the receptive field is smaller, and more attention is paid to detailed information. In the feature information connected in the later layer, the number of convolutions is more, the receptive field is larger, and the global information is more concerned. Through different sizes of receptive fields, the richness of feature information is increased. The proposed multi-scale segmentation method in this paper optimizes the multiple connection operations, and only the last connection operation is retained, which greatly improves the recognition speed.

WRN proposes a convolutional neural network based on the extended channels number learning mechanism, it will obtain the same accuracy as the depth network and faster recognition speed through the shallower network. In this paper, a network model with only seven layers is proposed by combining the WRN method for increasing the channel number and shortening the network length, which can ensure the recognition accuracy and further improve the recognition speed.

3.1 Design of Multi-scale Segmentation Module

This paper proposes three kinds of multi-scale segmentation network models (SSRNet). The 3x3 convolution in ResNet is replaced by a new multi-scale segmentation module (SS-Block), as shown in Fig. 1. When the feature information is input, in the first step, the feature information input by 1×1 convolution is evenly divided into two equal parts according to the number of channels. In the second step, half of the feature information is directly sent to the end for fusion, while the other half is convolved. In the third step, it repeats the above steps until the last feature information. The fourth step is to combine the last feature information with the first half of the feature information obtained previously and output them together to 1×1 convolution. The multi-scale segmentation module of SSRNet-a is similar to the bottleneck module of ResNet. It just replaces the 3×3 convolution in the ResNet bottleneck module with SS-Block. Wherein, the convolution of SS-block is expressed as 3×3 convolution+batch regularization+ReLU activation function.

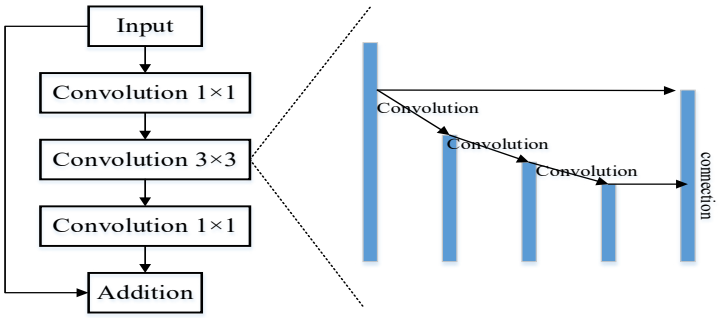

Fig. 1. Detailed SS-Block of SSRNet-a

The detailed structures of the three kinds of multi-scale segmentation network models on CIFAR dataset [19] are shown in table 1. The model a represents that only the 3×3 convolution in the ResNet bottleneck mod-

ule is replaced. Model b represents the removal of two 1×1 convolution models in ResNet bottleneck module on the basis of model a. Model c means that the number of feature channels in each sub-sampling is equal to the number of feature channels in the first group on the basis of model b. Compared with model b, model a has similar recognition accuracy, smaller number of parameters and slightly slower speed. Compared with model b, model c has lower recognition accuracy, smaller number of parameters and faster speed.

Table 1. Structure of three SSRNets

Group	Output size	SSRNet-a	SSRNet-b	SSRNet-c
1	32×32	[3×3,128]×2	[3×3,256]×2	[3×3,256]×2
2	161×6	[3×3,256]×2	[3×3,512]×2	[3×3,256]×2
3	8×8	[3×3,512]×2	[3×3,1024]×2	[3×3,256]×2
Average pooling	1×1	[8×8]	[8×8]	[8×8]

3.2 Design of Network Length and Segmentation Scale

Starting from AlexNet, deep convolutional neural networks learn multi-scale feature information by increasing the length of the network to improve network accuracy. While increasing the length of the network, the problem of reducing feature reuse will occur. WRN solves this problem by increasing the network width instead of the network length, making the recognition speed faster. If the network length is shorter, the recognition speed is faster. As the length of the network increases, the receptive field of the convolutional neural network increases. However, not all receptive fields have the same contribution to the output features, and the central area of the receptive fields has a greater impact on the output features [20]. Through rich experiments, this paper finds that When the network length is set to 1.5 to 2 times of the ratio of the receptive field size to the input image size, the recognition accuracy of the network is higher and the recognition speed is faster. A good balance between accuracy and speed is achieved. The receptive field of k-th layer is calculated as follows:

$$l_k = l_{k-1} + (f_k - 1) \times \sum_{i=1}^{k-1} s_i \cdot \quad (1)$$

Where l_{k-1} is the receptive field size of k-1 layer, f_k is the convolution kernel size of the current layer, and s_i is the step size of i-th layer. According to formula (1), the final output receptive field size of SSRNet-a is 51×51. After three kinds of scale segmentation, the receptive field is 67×67, which is 1.5 to 2 times of the input image size of 32×32. The length of the network model can achieve a good balance between recognition accuracy and recognition speed.

3.3 Design of Network Width and Down-sampling Rate

By increasing the network width (that is, the number of feature channels), the recognition accuracy can be improved while maintaining the network recognition speed. But increasing the network width will increase the number of parameters, so the network width needs to be set according to the specific situation. When the computing power is sufficient, it can set the network width of each group to be the same, and the network widths of different groups increase sequentially (for example, model a and model b). When the computing power is insufficient, the network width of each group can be set to be the same, and the number of the network model parameter can be reduced by sacrificing the accuracy (such as model c).

When the input image is too large, the model computation can be reduced by down-sampling. When the input image is down-sampled to small size (6×6, 7×7, 8×8), it is more appropriate [21]. The proposed multi-scale segmentation network model reduces the size of the input image from 32×32 to 8×8 by twice double down-sampling on the CIFAR dataset, and then it uses average pooling to change the output feature size to 1×1 for image recognition.

4 Experiment and Analysis

In the network model training, the proposed algorithm in this paper is implemented using the flying paddle

deep learning framework. The used data sets are cifar 10 and cifar 100. Each image uses random edge filling and random horizontal flipping for data enhancement. Finally, the mean value is normalized. The training environment is Baidu AI Studio, and the GPU is Tesla V100.

4.1 Datasets

CIFAR 10 and CIFAR 100 are labeled as subsets with 80 million small image datasets. They were collected by Alex Krizhevsky, Vinod Nair and Geoffrey Hinton.

CIFAR 10 data set consists of 60000 color images with 32×32pixels and 10 classes, each class has 6000 images. There are 50000 training images and 10000 test images. The 10 categories are: airplane, automobile, bird, cat, deer, dog, frog, horse, boat and truck.

The CIFAR100 dataset is similar to the CIFAR10 dataset. It has 100 classes, each class contains 600 images. Each category has 500 training images and 100 testing images. There are 50000 training images and 10000 testing images. The 100 classes in CIFAR100 are divided into 20 superclasses. Each image has a fine label and a rough label.

4.2 Training Strategy

The random gradient descent method with momentum is used to train the network model, and 300 rounds of training are conducted on the training sets (CIFAR10 and CIFAR100). The initial learning rate is set as 0.00001, and 10 rounds of linear warm-up training are carried out. Then, 290 rounds of cosine decay training are started with a learning rate of 0.1. The weight parameter of convolution kernel is initialized by MSRA [22], the weight attenuation coefficient is set to 0.00005, the momentum coefficient is set to 0.9. The data size of each batch is set to 128. The label smoothing strategy with the coefficient of 0.05 is used.

4.3 Results and Analysis

Four images are randomly selected from CIFAR10 data set, and the recognition results of SSRNet are shown in Fig. 2. The original image is composed of a 32×32 pixel color image, and the upper left corner is the category label corresponding to the recognition result.

Fig. 2. Recognition result of SSRNet

The comparison of recognition speed and recognition accuracy of SSRNet, ResNet and HS-ResNet trained on CIFAR10 data set with the same training parameters is shown in Fig. 3. The three circles in the upper left corner represent the multi-scale segmentation network model (SSRNet) in this paper, the three circles in the lower right corner represent the residual network model (ResNet), and the middle circle represents the multi-scale segmentation module (SS-Block), which is replaced by HS-Block. The recognition accuracy of SSRNet-c is similar to that of ResNet-56, and the speed is faster than that of ResNet-20. Compared with SSRNet-a and SSRNet-b, the recognition accuracy and speed of HS-ResNet obtained by shortening network length method are slightly lower. Thus, compared with HS-Block, the multi-scale segmentation module SS-Block has higher recognition accuracy, faster recognition speed and better performance.

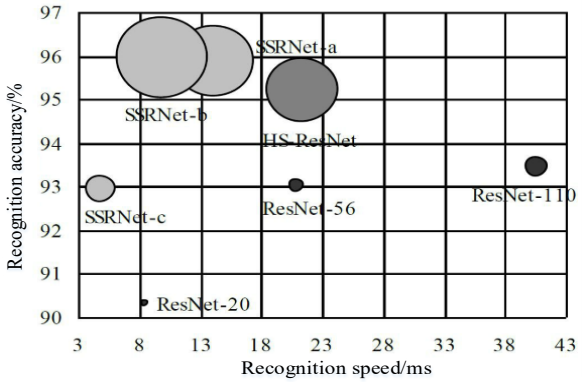

Fig. 3. Comparison of speed and accuracy with different networks

Under the training environment and strategy of this paper, the experimental results with different network models on CIFAR dataset are shown in table 2. The recognition speed of multi-scale network model SSRNet-c is the fastest, which is four times as ResNet-56. The recognition error rates of SSRNet-a and SSRNet-b are similar, which are 4.14% and 4.04% respectively. The former has smaller model parameters, while the latter has faster recognition speed. When the recognition speed is high, SSRNet-c can get the fastest recognition speed. When the recognition accuracy and model parameters are required to be high, SSRNet-a can obtain high recognition accuracy and low model parameters. SSRNet-b is a trade-off between the performance of SSRNet-a and SSRNet-c, it achieves a good balance between recognition accuracy, recognition speed and model parameters.

Table 2. Comparison of speed(ms) and error rates (%) with different networks on CIFAR datasets

Network model	Parameter size	Recognition speed	CIFAR 10	CIFAR 100
ResNet-20	0.2M	117	9.70	-
ResNet-56	0.9M	47	6.99	-
ResNet-110	1.7M	25	6.57	29.68
HS-ResNet-20	0.2M	37	9.71	-
HS-ResNet-56	0.8M	12	7.14	-
HS-ResNet-110	1.5M	6	6.52	29.68
HS-ResNet-a	17.1M	47	4.81	20.50
SSRNet-a	20.2M	71	4.14	21.00
SSRNet-b	26.8M	100	4.04	20.87
SSRNet-c	3.1M	200	7.07	30.69

Table 3 shows the different network models on the CIFAR data set. When the recognition error rate of the new network model in this paper is similar to that of other network models, the number of network layers is the smallest. Most of the neural networks currently used for image recognition are based on deep convolutional neural networks, and the number of network layers is much larger than the number of network layers in this paper. The fewer network layers present the faster recognition speed. When the recognition error rate is similar, the recognition speed of the network model in this paper is faster than other network models.

Under the training environment and training strategy in this paper, the results of SSRNet-b ablation experiment on CIFAR data set are shown in table 4. After the residual module in ResNet is replaced by the multi-scale segmentation module SS-Block, the error rate is similar, and the recognition speed decreases from 25ms to 11ms. After the improvement of ResNet by shortening the network length method, the error rate is similar and the recognition speed is increased from 25ms to 192ms. The error rate of the obtained SSRNet is greatly reduced by 2.53% and 8.81% on CIFAR10 and CIFAR100 datasets. The recognition speed is improved from 25ms to 100ms. That is a threefold increase. It can be seen that shortening the network length method can greatly improve the recognition speed, while combining it with multi-scale segmentation method can greatly improve the recognition accuracy.

Table 3. Error rates (%) with different networks on CIFAR datasets

Network	Parameter size	Layer number	CIFAR 10	CIFAR 100
ResNet	1.7M	110	6.57	29.68
WRN	36.5M	28	4.17	20.50
PyramidNet	1.7M	110	4.58	23.12
HS-ResNet	1.5M	110	6.52	29.68
SSRNet-a	20.2M	7	3.82	19.38
SSRNet-b	26.8M	7	3.87	20.87
SSRNet-c	3.1M	7	6.75	30.69

Table 4. Comparison of speed (ms) and error rates (%) of SSRNet-b ablation experiment on CIFAR datasets

Network	Recognition speed	CIFAR 10	CIFAR 100
ResNet-110	25	6.57	29.68
SSRNet-110	11	6.84	30.54
ResNet-b	192	6.76	30.79
SSRNet-b	100	4.04	20.87

5 Conclusions

The proposed multi-scale segmentation method in this paper does not increase the model parameters and computation when it represents multi-scale feature information. The method of proposed shortening the network length in this paper can greatly improve the recognition speed while maintaining the recognition accuracy. Compared with other network models, SSRNet based on multi-scale segmentation is faster when the recognition accuracy is similar. In future work, the network model will be applied to object detection and other fields to further improve the recognition speed in other fields.

6 Acknowledgement

The authors express their appreciation for the anonymous review.

References

[1] S. Karim, Y. Zhang, S.-L. Yin, I. Bibi, A Brief Review and Challenges of Object Detection in Optical Remote Sensing Imagery, Multiagent and Grid Systems 16(3)(2020) 227-243.
[2] J. Yu, L. Zhao, A Novel Deep CNN Method Based on Aesthetic Rule for User Preferential Images Recommendation, Journal of Applied Science and Engineering 24(1)(2021).
[3] R. -W. Bello, A. S. A. Mohamed, A. Z. Talib, Contour Extraction of Individual Cattle From an Image Using Enhanced Mask R-CNN Instance Segmentation Method, IEEE Access 9(2021) 56984-57000.
[4] A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet Classification with Deep Convolutional Neural Net-works, in: Proc: Neural Information Processing Systems (NIPS), 2012.
[5] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, (2015). https://arxiv.org/pdf/1409.1556.pdf
[6] K. He, X. Zhang, S. Ren, S. Jian, Deep Residual Learning for Image Recognition, in: Proc: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[7] S. Zagoruyko, N. Komodakis, Wide Residual Net-works, in: Proc: British Machine Vision Conference 2016, 2016.
[8] S. Gao, M. Cheng, K. Zhao, X.-Y. Zhang, Res2Net: A New Multi-scale Backbone Architecture, in: Proc: IEEE Transactions on Pattern Analysis and Ma-chine Intelligence, 2019.
[9] P. Yuan, S. Lin, C. Cui, Y. Du, S. Han, HS-ResNet: Hierarchical-Split Block on Convolutional Neural Network, (2020). https://arxiv.org/pdf/2010.07621v1.pdf
[10]D.-F. Yuan, C.-F. Chen, F.-M. Dong, Research on Residual Network of Image Recognition Based on Multiscale Split, Computer Engineering, (2021) 1-7. (in Chinese)
[11]T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, M. Li, Bag of Tricks for Image Classification with Convolutional Neural Networks, (2018). https://arxiv.org/pdf/1812.01187.pdf
[12]D. Han, J. Kim, J. Kim, Deep Pyramidal Residual Networks, in: Proc: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[13]X. Gong, T. Zhang, C. L. P. Chen, Z. Liu, Research Review for Broad Learning System: Algorithms, Theory, and Applications, IEEE Transactions on Cybernetics (2021), doi: 10.1109/TCYB.2021.3061094.
[14]A.-O, Karali, S. Cakir, T. Ayta, Multiscale contrast direction adaptive image fusion technique for MWIR-LWIR image

pairs and LWIR multifocus infrared images, Applied Optics 54(13)(2015) 4172-4179.
[15]S. Yin, H. Li, Hot Region Selection Based on Selective Search and Modified Fuzzy C-Means in Remote Sensing Images, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13(2020) 5862-5871.
[16]W. Ma, X. Wang, J. Yu, A Lightweight Feature Fusion Single Shot Multibox Detector for Garbage Detection, IEEE Access 8(2020) 188577-188586.
[17]T.-L, Lin, P. Dollar, R. Girshick, et al., Feature Pyramid Networks for Object Detection, in: Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[18]K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, C. Xu, GhostNet: More Features From Cheap Operations, in: Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[19]A. Krizhevsky, G. Hinton, Learning Multiple Layers of Features from Tiny Images, Handbook of Systemic Autoimmune Diseases 1(4)(2009).
[20]W. Luo, Y. Li, R. Urtasun, R. Zemel, Understanding the Effective Receptive Field in Deep Convolutional Neural Networks, (2017). https://arxiv.org/pdf/1701.04128v1.pdf
[21]X. Cao, A Practical Theory for Designing Very Deep Convo-Lutional Neural Networks, (2015). http://pdfs.semantic-scholar.org/7922/2fad9f671be142bd7e42cd785a2cb 06a1d30.pdf
[22]K. He, X. Zhang, S. Ren, J. Sun, Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, in: Proc. 2015 IEEE International Conference on Computer Vision (ICCV), 2015.