Research on The Efficiency of Mixed Commodities Classification based on Deep Convolutional Neural Networks

Ruiqi Cao *

School of Cyber Science and Engineering, University of International Relations, Beijing, 100000, China

* Corresponding author: rqCao@uir.edu.cn

Abstract. With the wide application of deep learning in the field of image classification, the application of this technology in daily life has become more popular. In this context, self-service shopping based on deep learning makes people's lives more convenient. This paper extends the classification and identification of fruit and vegetable classes, which in turn investigates the classification of mixed commodities. This study aims to compare the performance of three widely used models, VGG, DenseNet, and ResNet, on a mixed commodities dataset. This research focuses on evaluating each model's performance on this dataset and determines which is the most suitable for this task. The study found that the accuracy levels of these three models varied. The training set's accuracy reached 0.91 for VGG16, 0.86 for Densenet201, and 0.74 for Resnet50. The validation set's accuracy reached 0.84 for VGG16, 0.87 for Densenet201, and 0.77 for Resnet50. In this study, DenseNet has performed relatively well and is more suitable for the mixed commodities dataset. The experiment has shown that to fulfill the high-efficiency requirements of self-checkout, the model shortcut can effectively save time and prevent the phenomenon of overfitting. Based on repeated experiments, it can be concluded that the data in the dataset used for self-service shopping model training should highlight its features and reduce the influence of background noise and other noises that may affect model training.

Keywords: Image classification, Mixed commodities, Densenet, VGG, Resnet.

1. Introduction

Image classification is a fundamental problem in computer vision, which aims to match different images with different category labels and has a wide range of applications. In recent years, deep learning has dominated the field of image classification, promoting the rapid development of neural networks and deep learning techniques, which have become the most advanced image classification methods. Based on the rapid development of such techniques, many aspects of daily life have become more intelligent.

For example, shopping has become much simpler and more efficient with the help of deep learning-based self-service shopping. This technology involves image classification that allows customers to only scan the products without having to scan specific barcodes. The system can recognize the products and calculate the price. Moreover, the system supports various electronic payment methods such as Alipay, WeChat, and others. Compared to traditional scanning methods, this deep learning-based shopping mode is more convenient and doesn't require manual inputs or calculations by salespersons, thus saving labor and time costs. In addition, this deep learning-based self-checkout model also provides convenience for people with vision problems compared to traditional scanning methods, such as barcodes or QR codes. Traditional barcodes are usually small in size and require a lot of time for visually impaired people to find the barcode to complete self-service shopping. Deep learning-based image recognition scanning relates to the direct recognition of the product regardless of its position, which is relatively friendly to the visually impaired.

The Belgian company Colruyt Group and Rob vision worked together to develop this deep-learning image recognition-based self-checkout technology and tested it in Colruyt's supermarkets focusing on the automatic recognition of fruits and vegetables [1]. According to the test results, this technique effectively enhanced checkout efficiency and had a high rate of accurate recognition [2].

The success of the test indicated the feasibility of incorporating deep learning-based image recognition in the area of self-service shopping, which significantly improved shopping efficiency.

However, Colruyt's test was limited to recognizing only fruits and vegetables. Therefore, this study aims to investigate whether deep learning models can efficiently classify a broader range of commodity samples. The focus of this paper is to compare the performance of three widely used image classification models, namely VGG, ResNet, and DenseNet, on a mixed commodity dataset in the same conditions. The purpose is to determine which model performs best in this kind of mixed commodity classification task. This research can provide a valuable reference for selecting models for deep learning-based hybrid goods classification and improve the efficiency of such tasks.

2. Dataset and methods

2.1. Dataset

The dataset selected in this paper consists of three different kinds of datasets Clothing & Models, Vegetable Image Dataset, and Fruits-262, recombined into a new dataset. All the above three datasets were obtained from the Kaggle website (https://www.kaggle.com/datasets). To improve the efficiency of the experiment, 90 categories of classes were randomly selected from the recombined dataset for training in this experiment. Among them, 6 clothes classes from Clothing & Models, 15 vegetable classes from Vegetable Image Dataset, and 69 fruit and vegetable classes from Fruits-262. To obtain a relatively balanced dataset and to make the experimental data more general, the dataset was further processed in this experiment. To be specific, 1000 photos were chosen as the training set. In case the number of images was less than 1000, training was done with the original data size and the minimum number of data images was not less than 800. For both validation and test, 50 images were used.

2.2. Methods and Models

This experiment will focus on the performance of the three models in this environment and evaluate their ability to categorize the differences in category features in different categories and the similarity of category features in the same category. The three models are VGG, Resnet, and Densenet, which are widely used and powerful in this field.

Visual Geometry Group (VGG) is a deep convolutional network architecture proposed by the Visual Geometry Group of the University of Oxford in the 2014 ILSVRC classification task [3]. VGG can be seen as a deepened version of AlexNet [4]. The VGG model consists of five convolutional layers followed by three fully connected layers with a softmax output layer, which are separated by max-pooling layers, and each hidden layer uses ReLU activation functions. In this experiment, VGG16 is used, and the network as a whole uses a 3*3 size convolutional kernel and 2*2 size max-pooling, so that each convolutional layer fights to keep the same size of the previous layer, a total of 16 layers of conv+fc layers are stacked. The model performs well in image classification work on different categories of datasets [5, 6].

Dense net was introduced by Huang et al. at CVPR 2017[7]. It builds on the ResNet architecture, using shortcuts to a greater extent. This approach reduces the number of model parameters and speeds up model training, while also improving accuracy. Unlike ResNet, which shortcuts only in the front and back layers, Densenet enriches the DenseBlock module. This means that each layer receives inputs from all previous layers, combining the features of higher and lower layers and reducing the number of channels between each layer. By including shorter connections between layers close to the input and output, convolutional networks can be trained more deeply, accurately, and efficiently [8]. Using these shortcuts allows for the loss function to be directly passed down to the bottom layer, which speeds up model convergence and mitigates the problem of vanishing gradients.

ResNet was introduced by Kaiming He and his team in 2015 [9] and further improved in 2017 [10]. It was designed to address the problem of deep neural network degradation. To achieve this, ResNet introduces a residual learning module, which adjusts the number of channels of the feature

map. This module allows for a constant mapping of the neural network through shortcut connections when the dimension of the output feature map increases. With this method, the number of parameters and computations does not increase, making it easier to optimize the network. As the depth of the network increases, the performance of ResNet improves.

2.3. Evaluate Method

In this task, two metrics will be used for comparing the results of the experiments. These two metrics are Loss and Accuracy. The loss function will be used to measure the difference between the predicted results of the model and the true results. Accuracy represents the number of samples correctly categorized for a given data as a proportion of the total number of samples. In addition, image prediction will be introduced in this experiment to assist in analyzing the experimental results. After 10 epochs of training, the experimental results of the three models will be compared and analyzed. Evaluation will be done by validation set and prediction will be done by test set. When the Loss function of the validation set is lower, the accuracy of the validation set is higher, and the image prediction match is better, it indicates that the model has a stronger effect on the given task and is more suitable for this task.

3. Results

The study gained results by conducting multiple experiments using three distinct models. In this task, the effectiveness of the models will be evaluated in terms of accuracy, loss function, and time cost.

Fig 1(a), (b), (c) displays the changes in the accuracies of VGG16, Densenet201, and Resnet50 during the training process, respectively, and Fig1 (a) shows that the accuracies of VGG16 can grow smoothly to more than 0.9. Fig 1 (b) shows that the accuracies of Densenet201 can grow smoothly to close to 0.9. Fig 1 (c) shows that the accuracies of Resnet50 can grow smoothly to close to 0.8, which indicates that the three models are effective in this task.

Fig 2 (a)(b)(c) depicts the changes in the loss functions of three models individually. During the training process, the loss values of all three models decrease steadily and approach the fit. These indicate that all three models can learn successfully and perform well on this dataset, thus demonstrating the feasibility of these models for this task.

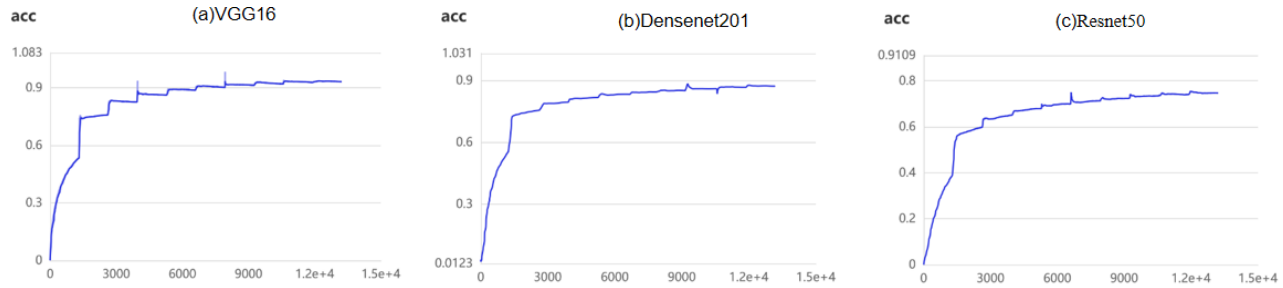

Figure 1. Changes in the accuracy of the three models: (a) is the VGG 16; (b) is the Densenet201; (c) is the Resnet50.

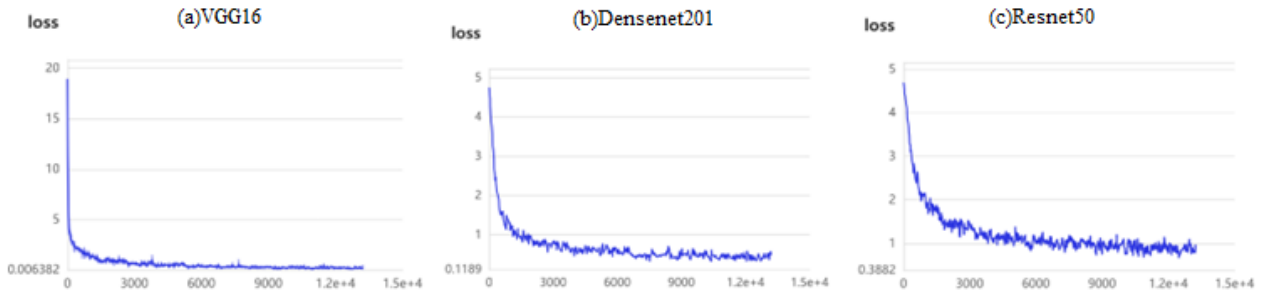

Figure 2. Changes in the accuracy of the three models: (a) is the VGG 16; (b) is the Densenet201; (c) is the Resnet50.

Table 1. Three models' best performance in replications

Model name	Best training Accuracy	Best training Loss	Best validation Accuracy	Accuracy comparing	Best validation Loss	Loss comparing
VGG16	0.91	0.015	0.84	-3.5%	0.03	0
DenseNet201	0.86	0.45	0.87	0	0.07	+0.04
ResNet50	0.74	1.22	0.77	-11.5%	0.3	+0.27

From Table 1, according to the consequence of the 10 epochs training under the same conditions, VGG16 can achieve an accuracy of 0.91 and a loss function of 0.015; Densenet201 can achieve an accuracy of 0.86 and a loss function of 0.45; Resnet50 can achieve an accuracy of 0.74 and a loss function of 1.22. In comparison, VGG16 will perform slightly better than Densenet201 on the training set and significantly better than Resnet50. However, in terms of performance on the validation set, VGG16 can achieve an accuracy of 0.84 with a loss of 0.03, Densenet201 can achieve an accuracy of 0.87 with a loss of 0.07, and Resnet50 can achieve an accuracy of 0.77 with a loss of 0.3.

In terms of performance on the validation set, desnet201 will perform slightly better than VGG16 and significantly better than Resnet50. Combining the performance on the training set and validation set, VGG16 and Densenet201 have a significant advantage over Resnet50 and perform better on this dataset. These two better models will be explored further in this paper in conjunction with the image prediction visualization.

For both models, repeated experiments were conducted in this experiment to remove the chance factor. The results of multiple experiments show that VGG16 has 7% higher accuracy and 80% lower loss compared to densenet201 accuracy during model training, indicating that VGG16 performs better on the training set. However, in the performance on the validation set, densenet201's accuracy is 3.5% higher compared with VGG16, indicating that densenet201 is relatively better on the validation set, but the loss of VGG16 is still significantly lower than that of densenet201.

After analyzing multiple visualization results, it can be concluded that the difference in the prediction of this task between densenet201 and VGG16 is not significant. However, these two models have some similarities, and differences between the images cannot be certainly classified. Taking the kiwifruit group as an example as shown in Fig. 3 and Fig. 4. As the Fig shows, both models cannot effectively classify this image in the third row of the fourth column correctly. It shows that both models cannot classify the pictures with vague features accurately.

According to the consequence of the experiments, it is hard for models to classify an image with vague features. In addition, there are also some differences in the characteristics of the images for which the two models fail to make a correct classification. According to repeat experiments, Densenet201 is relatively easy to recognize part of the image feature as an item that also has this feature. As shown in Fig. 5, the kiwi with a handle is recognized as a bottle gourd.

After analyzing the results of multiple rounds of VGG16 experiments, observed that the model has relatively more cases of classification errors due to image color. As shown in Fig 6, the same situation did not occur during the training of densenet201, which may indicate that the VGG model is more

sensitive to color features. It has also been shown in others' studies that VGG performs well in tasks with high chromatic aberration.

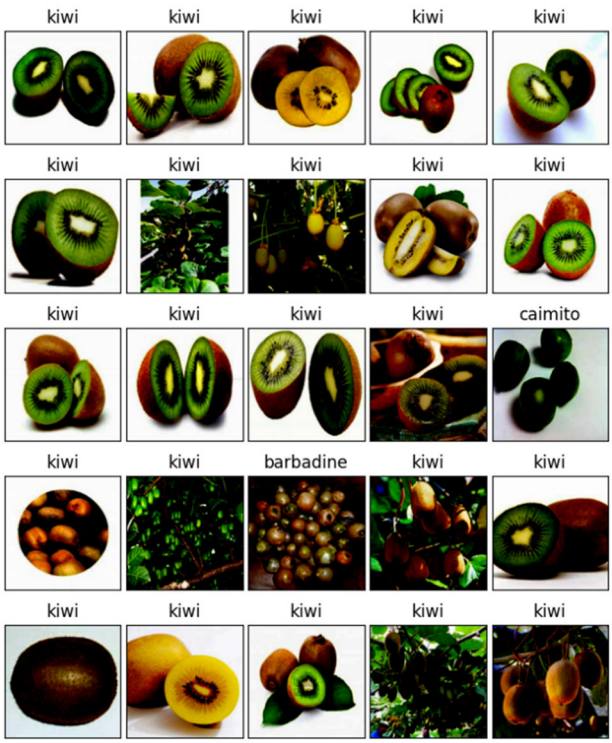

Figure 3. densenet201 kiwi images prediction

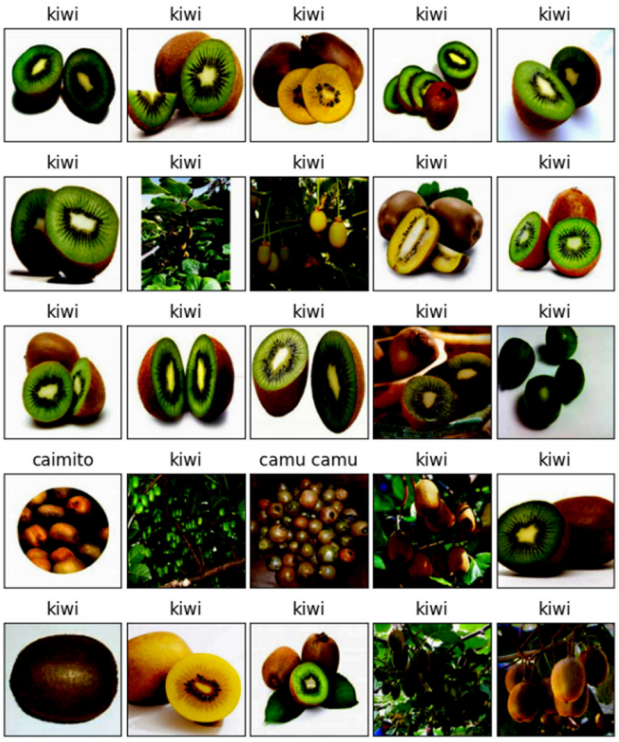

Figure 4. VGG16 kiwi images prediction

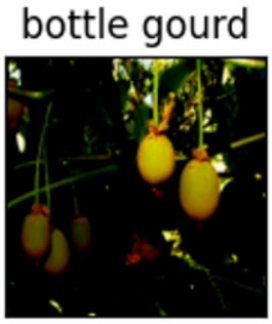

Figure 5. Kiwi but recognized as a bottle gourd.

Figure 6. Errors of VGG16 on similar color mages during training.

However, in terms of time cost, the overall training time cost of VGG16 reaches 4427 seconds, with each training step about 320ms and each evaluation step about 270ms. The overall training time cost of Densenet201 reaches 3727 seconds, with each training step about 265ms and each evaluation step about 220ms. Comparing these two models, Densenet201 is more advantageous in time cost, being able to save 700 seconds over the 10 epochs of this task. This advantage expands with the increase of the training step and the increase of epoch due to the expansion of the dataset.

4. Discussion

4.1. Model Problem

During the training process, there is not much difference in the accuracy of Densenet and VGG models on the validation set. However, the loss of VGG16 is 80% lower compared with the loss of Densenet, and the difference that exists between the two models on this task should not reach this interval. This paper argues that the experiment suffered from overfitting, which can be attributed to the structural differences between the two models used. The VGG model achieved deeper layers

simply by stacking layers, which, although effective, led to an increase in the model's parameters and complexity as it became deeper with more layers, which might be the feature that caused overfitting during this task. However, the Densenet model includes DenseBlock modules that enhance the transfer of information and gradients within the network. This dense connection improves the efficiency of resource utilization, making the network easier to train [11]. Deep supervision enables each Densenet layer to use the loss function gradient and front inputs, facilitating deep network training. At the same time, dense connections have a regularizing effect that reduces overfitting in tasks with fewer training sets [5].

Since this task is a study based on a local small-scale deep learning model, and the period of the task is limited, it is impossible to gain a huge amount of data and hardware conditions such as those for large model training, so this overfitting reduction feature of Densenet makes the model able to reach a better outcome and less likely to be overfitted in this task. However, this also opens up new avenues for potential improvements in future work. By expanding the dataset, operating cross-validation techniques, or applying more preprocessing methods to the data to increase its complexity, it is possible to achieve better results with the improved model.

4.2. Feature Disturbance

Thoughts based on Fig 5, similar features and unfamiliar backgrounds can disturb the model's classification. According to previous research in this field, when an item is placed in an unusual context, it can make it difficult for the model to recognize it [12]. When models focus on such multiple and complex features, image segmentation may treat a part of the environment as the target object [13].

Under the influence of such distinctive features of both the target and the environment, it becomes difficult for traditional deep-learning models to make correct classifications.

As shown in Fig 7. In the left image, the bag is recognized by models as mashed potatoes InceptionV3 (55%), VGGNet16 (60%), and ResNet-50 (68%). In the right image, the chair is recognized as a guillotine, forklift, and turnstile InceptionV3 (74.4%), VGGNet16 (9.7%), and ResNet-50 (53.6%). It means that errors in image classification are context-based judgments. Obviously, mashed potatoes are more relevant to food, and forklifts are more relevant to cars. Based on the available data, it can be concluded that the model incorrectly classified a kiwi as a bottle gourd because they share similar features. The model identified the shank as a feature that is more likely present in other classes, leading to the misclassification. In this type of task, the appearance of similar features is a common situation, especially in vegetable and fruit classes. Because under the same broad category, many vegetables and fruits with similar biological features may impact the model's classification. Therefore, how to make the model classify correctly with background interference is a feasible direction for future work.

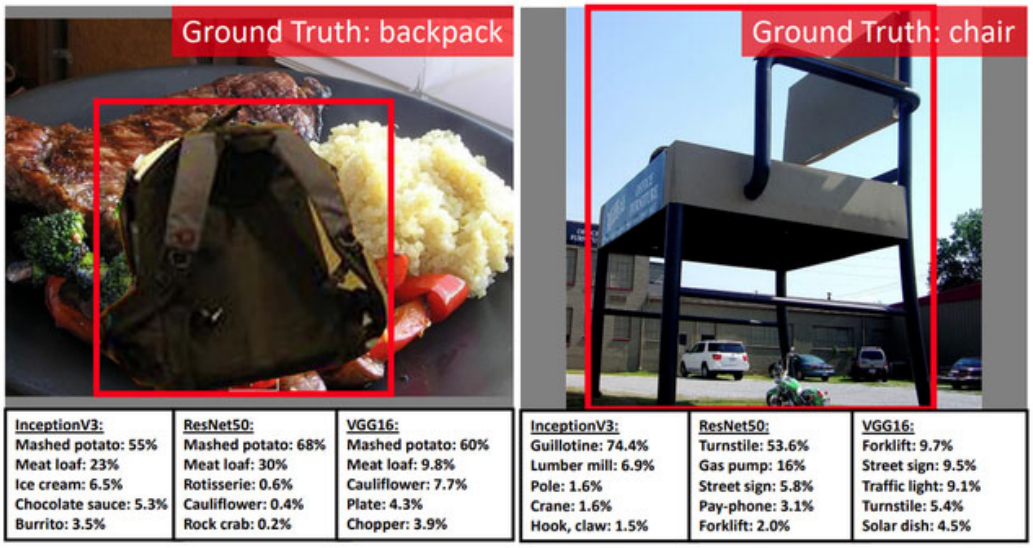

Figure 7. Misclassifications occur when InceptionV3, ResNet50, and VGG16 encounter unfamiliar contexts. Misclassifications occur when InceptionV3, ResNet50, and VGG16 encounter unfamiliar contexts. The bottom of the page displays the top 5 labels and their corresponding confidence levels for each model [13].

5. Summary

This paper compares the effectiveness of different models in classifying mixed commodities. The experiments found that out of the three models, Densenet and VGG perform significantly better than Resnet. Further comparisons reveal that Densenet performs better than VGG in this dataset. Densenet201 achieved a prediction effect of 0.87 on this dataset 3.5% higher compared with VGG on the test set. Meanwhile, in multiple repeat experiments, VGG shows abnormally low loss values without a substantial increase in accuracy during training. This indicates that based on this dataset, the VGG model, due to its model features, leads to a large number of parameters during the training process, which generates an overfitting phenomenon during the training process and hurts the final prediction effect of the model. This phenomenon appeared in the early epoch, so it is difficult to reduce the overfitting by simply reducing epochs, which means that the VGG model is not a good match in this experiment. However, this overfitting phenomenon is not obvious during the training process of the Densenet model. In this model, the accuracy increases steadily while the loss decreases in the training process. Meanwhile, compared with the VGG model, the Densenet model saves about 700 seconds (0.2 hours) in training and validation time in 10 epochs. The goal of this task is to improve the accuracy and efficiency of product identification during self-service shopping. The model should be able to predict with high accuracy and speed, thus saving the customer's time, which means that the model should quickly identify the products that the customer needs to check out with the correct match. In conclusion, among the three models compared, the Densenet model has some advantages such as time cost, and prediction efficiency, which means it is more suitable for the mixed commodities classification task based on this dataset than VGG & Resnet model.

References

[1] Colruytgroup. Colruyt is the first Belgian supermarket to test automatic fruit and vegetable recognition, Colruytgroup, 2023, https://press.colruytgroup.com/colruyt-is-the-first-belgian-supermarket-to-test-automatic-fruit-and-vegetable-recognition#.

[2] Robovision. The Colruyt Group Speeding up check-out times with vision AI, Robovision, 2023, https://robovision.ai/case-study/smart-scales-for-faster-check-out/.

[3] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv. 20141409. 1556.

[4] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012, 25.

[5] Hassannejad H, Matrella G, Ciampolini P, et al. Food image recognition using very deep convolutional networks. Proceedings of the 2nd international workshop on multimedia assisted dietary management. 2016: 41 - 49.

[6] Sitaula C, Hossain M B. Attention-based VGG-16 model for COVID-19 chest X-ray image classification. Applied Intelligence, 2021, 51: 2850 - 2863.

[7] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700 - 4708.

[8] Huang G, Liu Z, Pleiss G, et al. Convolutional networks with dense connectivity. IEEE transactions on pattern analysis and machine intelligence, 2019, 44 (12): 8704 - 8716.

[9] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770 - 778.

[10] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 1492-1500.

[11] Huang G, Chen D, Li T, et al. multi-scale dense networks for resource efficient image classification. arXiv preprint arXiv: 1703. 09844, 2017.

[12] Bomatter P, Zhang M, Karev D, et al. When pigs fly: Contextual reasoning in synthetic and natural scenes. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 255 - 264.

[13] Zhang M, Tseng C, Kreiman G. Putting visual object recognition in context. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020: 12985 - 12994.