MANGO-TERMINATOR: DEEP NETWORK FOR MANGO CLASSIFICATION

[1]*Sheng-Yu Huang,* [1]*Chiao-An Yang,* [2]*Chiou-Shann Fuh*

[1]Graduate Institute of Communication Engineering,
[2]Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan

E-mail: [1]{r08942095, joeyang}@ntu.edu.tw, [2]fuh@csie.ntu.edu.tw

ABSTRACT

How to apply machine learning researches into real world problems has been a really important issue. In this paper, we try to solve the mango grading problem of AI CUP2020 Mango Image Recognition Challenge. For the competition of mango image classification, dataset contains tens of thousands of images, including three grades A, B or C. A brief example is shown in Fig 1.

Instead of being a classic classification problem, this task has a dataset composed of very similar images between different grades, which lead to undesirable results of many common models like AlexNet [1], VGG [2] and ResNet [3].

In this paper, we introduce a new grading loss which is designed for grading tasks like this one. We would like to use some simple CNN model first (AlexNet [1], ResNet [3], VGG [2]), seeing how these baselines work, and then add our grading loss to show that it actually improves the performance of these models.

***Index Terms*—** Image Recognition, Pattern Recognition, CVGIP 2020.

1. INTRODUCTION

Image classification is a popular research area. It can extend to object detection, semantic segmentation, or even reconstruction and generation. After LeNet [4] was published, deep neural networks (NN) become the main area of research which aims to solve this problem. For the past decade, we have seen great improvements in NN, especially Convlutional Neural Network (CNN). AlexNet [1], VGG [2], GoogLeNet [5], and ResNet [3] are all famous models which show the world how powerful CNN is. More recently, U-Net [6]-like structure and usage of attention mechanism [7, 8] are more and more popular since they can handle many complicate settings such as segmentation and multi-object detection.

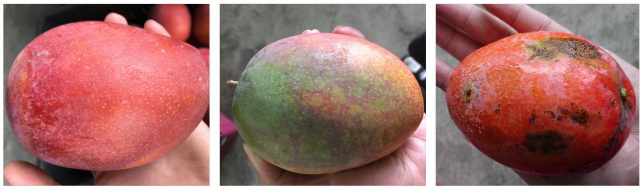

Fig. 1. **Example for each grade.** The left most figure is a example of grade A, the middle one is grade B, and the right most is grade C. It is obvious that grade A has uniformed red color and smooth texture, while grade B has some green area and grade C has trivial black spot.

This paper focuses on the mango grading task of AI CUP 2020 Mango Image Recognition Challenge. Grade classification includes three grades A, B or C, and specified defective categories.

In this paper, we introduce a new grading loss, which is designed for grading tasks. Our main contribution is listed as follows:

- We designed this loss that can fit for not only mango grading but also any grading related task.
- Our results show that this loss helps many backbone classifiers improve the classification performances.

2. RELATED WORK

2.1. Convolutional neural network.

When it comes to the discussion of Convolutional Neural Network (CNN), we always think of LeNet [4], which is the pioneer of all deep learning networks in recent years. At that time, what this CNN model could do was only recognizing number digits which are from MNIST dataset, since it is

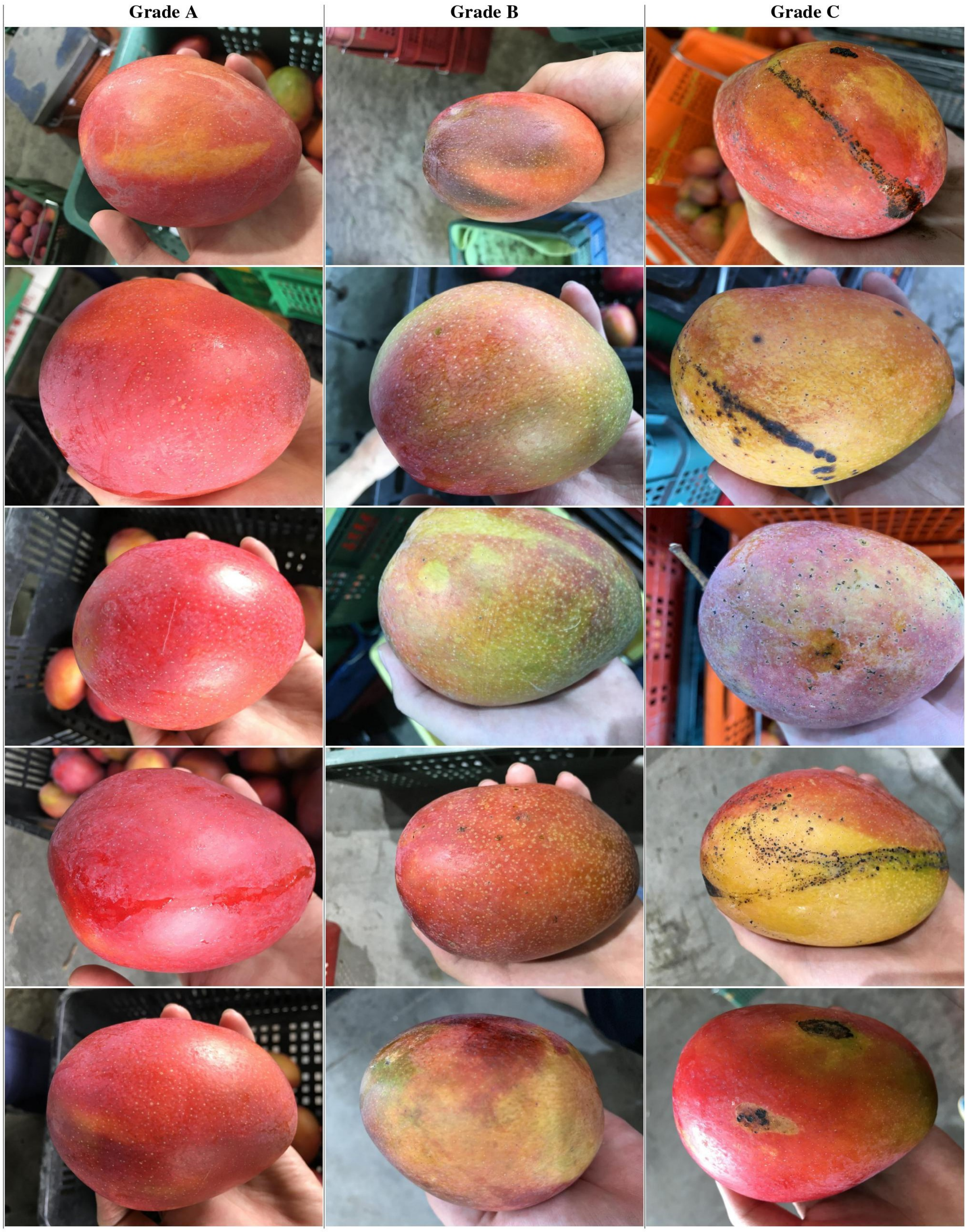

Fig. 2. **More examples for each grade.**

a very simple task. As larger and more complicated image datasets (CIFAR-10, CIFAR-100, ImageNet) are published, deeper and stronger CNN models are needed for better performance.

AlexNet [1] uses ReLu as activation function, and design a parallel structure to learn different feature for classification. VGG [2] and GoogLeNet [5] show that simply enlarge feature dimension and construct a deeper network improves the performance when the dataset is large enough. On the other hand, ResNet [3] make connections between different block, so that gradient vanish or explode can be avoided. All these models improve the performance and capability compared with LeNet and extend to a lot of applications for solving real-world problems. However, the feature they extract are mostly under single scale or easily affected by image background. This would restrict the model capability and lead to tragic failure when facing harder tasks.

2.2. Recent works about neural network.

To have better segmentation (i.e. pixel-wise classification) performance, U-Net [6] provides a method that concatenate features obtained during different scale of feature map when decoding from global feature to have more detailed reconstruction and segmentation.

As attention mechanism ([8, 7]) is introduced, object detection and image classification became much easier, since we can now get rid of unimportant background, and focus on what is really dominating the performance.

Although these mentioned methods seems well-performed, there are still some cases which would lead to failure. When the classification dataset is composed of targets which are basically similar (e.g. different grade of mangoes in our case), the model is require to be able to indicate very precise and detail differences between each class. Since all methods mentioned above are only supervised with Cross-Entropy Loss, we believe that by designing a proper grading loss, we can force the model to learn more relations between grades instead of simply regard them as several different classes.

3. DATASET

For the competition of mango image classification, the dataset was created at three fruit collection facilities in Fangshan, Pingtung, containing tens of thousands of images. Current released data number are shown in Table 2. The images were made in well-lighted spaces using smartphones and video recordings with HDR camera. Each of the images contained a shot of one mango that was put on the data collectors' hand or the conveyor. Furthermore, each image was labeled with a mango grade and defective types by professionals. Grade classification includes three grades A, B or C, and specified defective categories. These labels are used as the gold standard of the AI CUP 2020 Mango Image Recognition Challenge. As mentioned in Fig 1, grade A is basically beautiful and well colored, with no black spot. Grade B are mangoes that are not that well colored or having very tiny black spots. If a mango has any trivial black spot on it, then it is classified as grade C. More examples are shown in Fig 2.

	Training	Validation
A	1792	243
B	2068	293
C	1740	264
Total	5600	800

Table 1. Number details of current dataset.

4. METHOD

4.1. Key idea.

Although designing a more complicated model might also improve the performance compared with AlexNet [1], VGG [2], and ResNet [3], we want to show that simply adding a well-designed loss is able to give better results as well. Our grading loss, denoted as $\mathcal{L}_{Gr}$, is specially designed for grading tasks.

4.2. Grading loss.

Given N images with their grades $\mathcal{G} = \{g_n \mid n = 1, 2, \ldots, N\}$ and their one-dimensional output score from the model $\mathcal{S} = \{s_n \mid n = 1, 2, \ldots, N\}$, our grading loss $\mathcal{L}_{Gr}$ is defined as:

$$\mathcal{L}_{Gr} = -\sum_{i=1}^{N}\sum_{j=1}^{N}\{\boldsymbol{Sigmoid}(-(s_i - s_j))\,|g_i > g_j\}. \quad (1)$$

The meaning of this loss is that we wish the output score g_n can represent the quality of our input image, and this quality is directly related to the grade of this image as shown in table 3. Therefore, for any pair of unequally graded data, we sum up the differences between scores of larger graded one and smaller graded one. The $\boldsymbol{Sigmoid}$ function is added to avoid extreme value, and we take a negative sign so that pairs which satisfy our suppose would have smaller loss value.

5. EXPERIMENTS

5.1. Model setting.

In the experiment section, we want to explore if a model's performance would be better by making a few adjustments

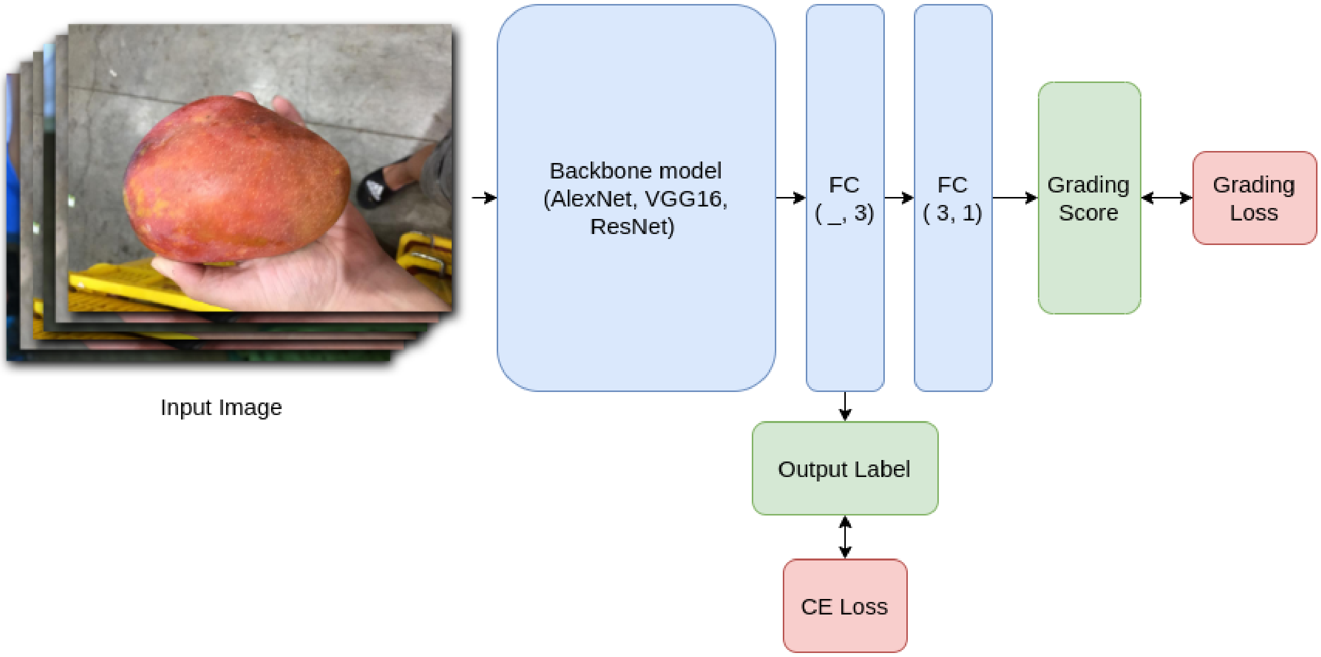

Fig. 3. **Illustration of model structure.** In the first FC layer, input channel equals to the output channel of the backbone model's former layer.

and simply adding our grading loss.

The model we are going to compare are AlexNet [1], VGG16 [2], ResNet18 and ResNet50 [3]. For each model, the baseline simply changes the final fully-connected layer's output channel from 1000 to 3 and then supervised by only Cross-Entropy loss. As for the grading loss related experiment, we add another fully-connected layer which input channel is 3 and the output channel is 1 so that this output can be supervised by our grading loss while the output of the previous layer supervised by Cross-Entropy loss as well. Therefore, instead of directly deciding the grade from the grading score, we still decide the grade from the three-dimensional score as shown in Fig 3.

5.2. Training details.

All the experiments are trained by GTX 1080 GPU and under PyTorch module. Data augmentation are applied as random shift between [-0.1, 0.1], random rotation between [-180, 180], and random scaling between [0.9, 1.1]. During training, learning rate is set to be 0.0001, and the loss weights are $\mathcal{L}_{Total} = \mathcal{L}_{CE} + 0.001 * \mathcal{L}_{Gr}$. To compare fairly, we train each model for 40 epochs and compare the results.

5.3. Results.

Our results are shown in Table 2. As we can see, for all backbone models we have tested, adding grading loss outperformed the origin one. This shows that model's performance can be improved by simply adding a proper loss function, and that out grading loss is what this task needed. Some examples of correctly and wrongly predicted cases are shown in 4.

	w/o grading loss	with grading loss
AlexNet [1]	0.80125	**0.81250**
VGG16 [2]	0.83125	**0.83750**
ResNet18 [3]	0.80625	**0.82625**
ResNet50 [3]	0.82750	**0.83750**

Table 2. **Results of models which are trained with and without our proposed grading loss.** The numbers in the table are accuracy. It is obvious that adding grading loss improves each model's performance.

5.4. Limitations.

We know that, for an image classification task, 5600 training image is quite small, so the performance would be some how limited by the size of dataset. However, the real main problem of this dataset is that the quality of label. As mentioned earlier, grade A are basically beautiful and well colored, with no any black spot. Grade B are mangoes which

	A	B	C
A		+	+
B	-		+
C	-	-	

Table 3. **A illustration of desirable grading score output for a 3 grades case.** For a three-graded task like this mango grading competition, we have A, B, C grades, where grade A is better than B and grade B is better than C. Therefore, the desirable output score should be that all grade A's score larger than B's and C's, and all grade B's score larger than C's.

are not that well colored or having very tiny black spots. If a mango has any trivial black spot on it, then it is classified as grade C. Unfortunately, as shown in Figure 5, we found many ambiguous cases where the mangoes looks fine but graded as B or C, or weird colored mangoes graded as A. We believe that this is an unneglectable issue for this task.

6. CONCLUSION

In this paper, we propose a novel grading loss to help many common models improve their ability of solving the grading task of AI CUP2020 Mango Image Recognition Challenge. Our results show that this loss is strong enough to improve all of them. We also take a look at the dataset's limitation, which might explain the reason why all methods' performance are stuck around 0.8.

7. FUTURE WORK

In the future, we would like to explore more properties of the grading loss. For example, for every single example, the defect of any part of the target is definitely less than the whole. Using this information, we can see the grading loss as a semi-supervise loss. For targets with ground truth labels, we can compare their score between different targets. As for targets do not have ground truth labels, compare scores between the whole target and its parts is also acceptable. We believe that this property is able to help all the model tested before to achieve better performance.

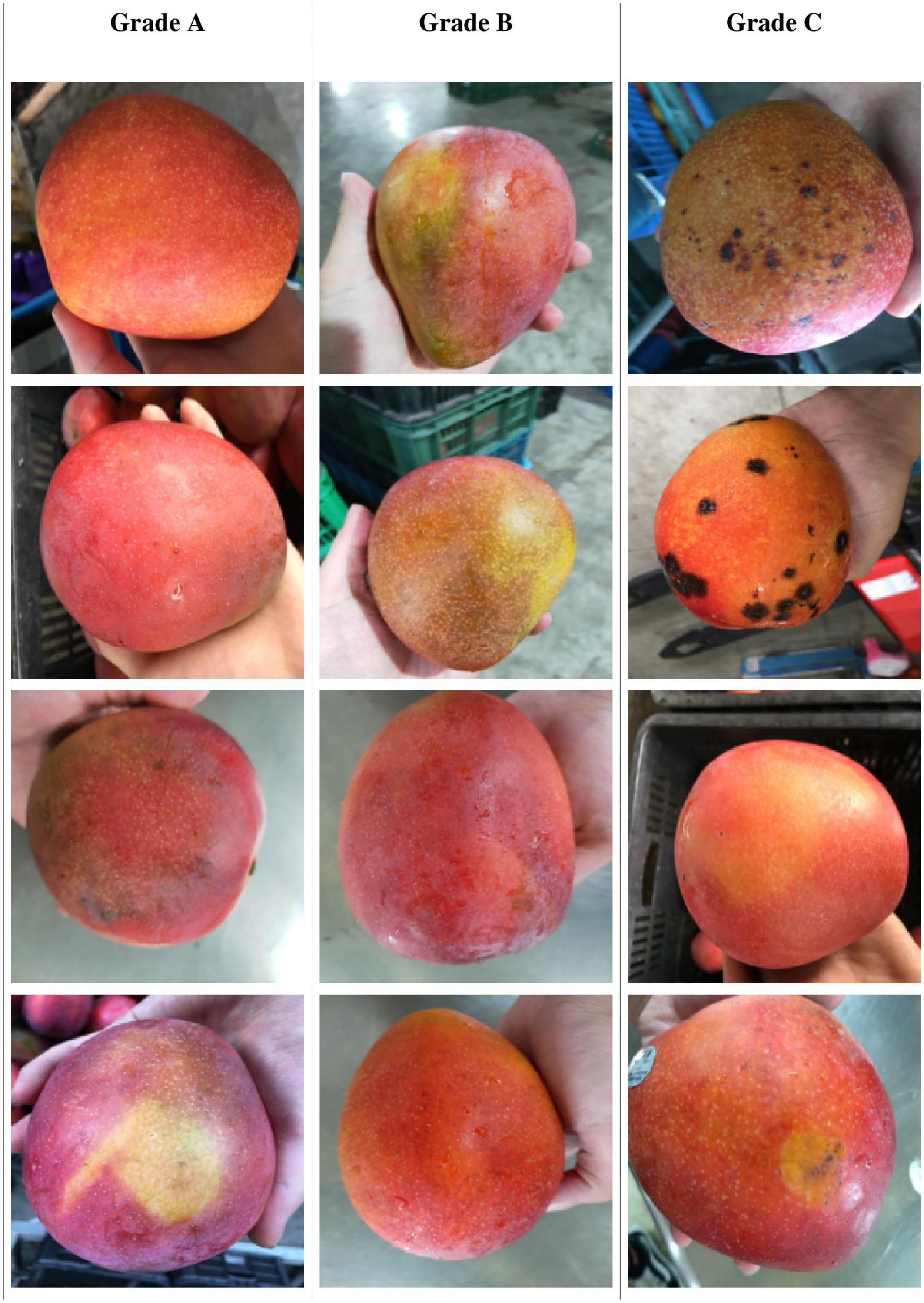

Fig. 4. **Examples of correct and wrong predicted data.** The first two rows are correctly predicted examples of grade A, B, and C, while the last two rows are failed cases. It is obvious that failed cases are due to the ambiguity of those data, where appearance between different grades are similar.

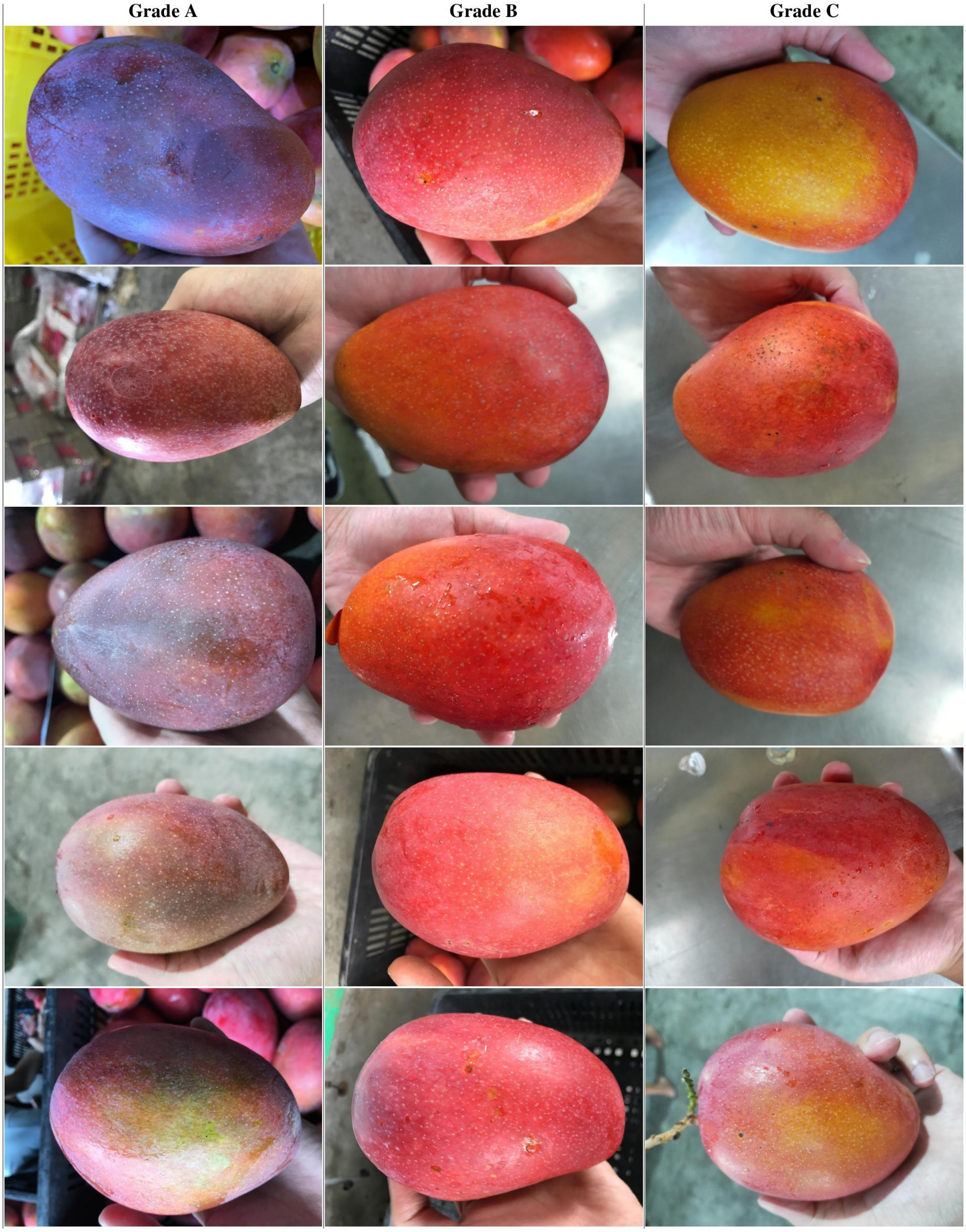

Fig. 5. Examples of ambiguous data.

REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.

[2] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015.

[7] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3156–3164.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.