# GRVT: Toward Effective Grocery Recognition via Vision Transformer

Shu Liu[1,3], Xiaoyu Wang[1,3], Chengzhang Zhu[2,3(✉)], and Beiji Zou[1,3]

[1] School of Computer Science and Engineering, Central South University, Changsha 410083, China
[2] School of Literature and Journalism, Central South University, Changsha 410083, China
anandawork@126.com
[3] Hunan Engineering Research Center of Machine Vision and Intelligent Medicine, Changsha 410083, China

**Abstract.** Grocery recognition aims to classify items by visual features of the image. The intention is to improve retailing experience, manage inventory and help visually impaired people. It is an important task in computer vision. Most previous works utilize global image features with a unique decision rule to recognize groceries and products via convolutional neural network (CNN) models. Such methods work on different CNN architectures to explore more accurate and representative features. However, fine-grained characteristics are not considered in feature extraction. Recently, vision transformer (ViT) models achieve success in multiple computer vision tasks. And fine-grained visual categorization is leveraging self-attention mechanism of ViT to learn discriminative regions and features. In this paper, we propose a novel ViT based framework named grocery recognition vision transformer (GRVT). It integrates multiple granularity scales of patches by multi-scale patch embedding to introduce robust image representation without incurring excessive computation cost. The mixed attention selection module guides the network to choose these discriminative patches and crucial regions for fine-grained feature extraction. Our GRVT achieves the state-of-the-art performance on Freiburg Groceries Dataset and Grocery Store Dataset.

**Keywords:** Grocery recognition · Fine-grained visual categorization · Vision transformer · Multi-scale patch embedding · Mixed attention selection

## 1 Introduction

Grocery recognition is a novel and practical research topic in computer vision and deep learning area. It plays a vital role in many applications in retail. Automatic grocery recognition is beneficial for retail and on-shelf product management. On the one hand, grocery recognition based on deep learning can be used for an automatic checkout system and improves customers' shopping experience


N. Magnenat-Thalmann et al. (Eds.): CGI 2022, LNCS 13443, pp. 266–277, 2022.
https://doi.org/10.1007/978-3-031-23473-6_21

by reducing waiting time [1]. On the other hand, it is able to develop better replenishment and planogram by monitors and to improve turnover and profits. For customers who are visually impaired, grocery recognition on mobile devices can assist them to distinguish products and accomplish shopping independently in their daily life [2].

Generally, grocery recognition is a classification task of computer vision. Its target is to distinguish different images into corresponding labels correctly. A classification system includes image capturing, image preprocessing, feature extraction, feature classification and the output of recognition [3]. Traditional feature extraction needs hand-crafted features which are not suitable for products that are changing day by day and also suffer from low accuracy. With the development of deep learning, the convolutional neural network (CNN) based image classification method obtains great success in recent years [4]. And promoted the development of related research fields such as object detection [5] and segmentation [6]. Recently, vision transformer (ViT) [7] shows better performance in image classification. ViT and its variants also achieve great success in popular tasks and further exceeds CNNs in particular aspects [8–10]. That shows the potential of transformer in capturing global and local information.

However, there is still a blank for transformer in grocery recognition. In this paper, we explore the capabilities of ViT in such task. An effective grocery recognition framework based on ViT, named GRVT, is presented. To be specific, to enhance generalizability on the patches of ViT, we propose a multi-scale patch embedding (MSPE) module to enable the network to integrate multiple scales of input. Moreover, mixed attention selection (MAS) module calculates attention scores across different sets of patch embeddings to choose discriminative and crucial regions. We evaluate our model on popular datasets including Freiburg Groceries Dataset [11] and Grocery Store Dataset [12], and our GRVT outperforms existing methods on these benchmarks. The main contributions are summarized as follows:

- To the best of our knowledge, we are the first to implement ViT model on grocery recognition task and explore its transfer learning performance. And ViT produces competitive results with CNN models.
- We introduce GRVT, a novel and compact architecture for grocery recognition that fuses multi-scale patch embeddings and computes mixed attention to choose crucial local regions.
- We verify the effectiveness of our framework on grocery datasets, and the experiment results show our GRVT achieves better performance than existing public works.

## 2 Related Work

Fine-grained visual categorization (FGVC) is a challenging task because of high intra-class variances and low inter-class variances. Datasets are mainly weakly-supervised with only class labels. Methods on FGVC are focusing on local regions to capture discriminative features as a categorization basis. Hu et al. [13] proposed

weakly supervised data augmentation network (WS-DAN) to generate attention maps with discriminative parts, and then utilized data augmentations to reinforce the learning procedure. Chen et al. [14] proposed a destruction and construction learning (DCL) by region confusion mechanism and injected more discriminative local details into the classification network. Muktabb et al. [15] integrated a local concepts accumulation layer emphasize local features and showed effect gains. Ji et al. [16] proposed an attention convolutional binary neural tree which characterizes the coarse-to-fine hierarchical feature learning process, and used the attention transformer module to enforce the network to capture discriminative features. He et al. [17] firstly introduced ViT to fine-grained recognition, transformer architecture for fine-grained recognition (TransFG) selects discriminative image patches and adopts contrastive loss to enlarge the distance between sub-classes.

Grocery recognition aims at classifying objects in supermarket scenarios, like fruits, milk, and snacks. It could be considered a FGVC task. There is low inter-class variance, for example, apples come in a great many varieties, and their color and shape are familiar even if customers can not distinguish them easily. Furthermore, there is also high intra-class variance, a Golden-Delicious apple in different angles or lighting conditions can vary much in vision. Besides, it has some other characteristics which make the task more challenging. In grocery shops, products can be put on the shelf or piled up in a container. The object scale is not definite, the background could be other products and environment conditions are unconstrained. In recent studies, grocery recognition methods exploit deep neural networks (DNN) models as a feature extractor. The unique decision rule is adopted to classify low-dimensional vectors such as support vector machines (SVM) to identify retail goods [12]. Ciocca et al. [18] proposed a multi-task learning network to leverage hierarchical annotations based on the CNN feature extractor. Noy et al. [19] and Nayman et al. [20] adopted neural architecture search (NAS) on grocery product recognition by introducing expert advice and architecture pruning. Wei et al. [1] proposed a retail product checkout (RPC) dataset and an automatic checkout (ACO) task. Wang et al. [21] proposed self attention based DCL to learn crucial region information to classify retail product images in the laboratory environment. Leo et al. [22] systematically study DNNs and ensemble DNNs on grocery recognition and found that model ensemble shows significant improvement.

## 3 Method

Here we explore the model performance in the application of grocery recognition. To better elaborate our framework, we first reviewed the ViT model design, and the overall GRVT architecture and its two modules are then introduced.

### 3.1 Vision Transformer

Transformer [23] has achieved outstanding performance in natural language processing due to its capacity and superiority in multi-head attention mechanism.

Due to the potential of transformer-based vision models, the number of ViT variants is increasing, such as data-efficient image transformer (DeiT) [24], pyramid vision transformer (PVT) [25], Swin Transformer [26].

Vision transformer applies a standard transformer [23] directly into images with least modifications [7]. For an image of $X \in \mathbb{R}^{H*W*C}$, where $H, W$ are the original resolution of image and $C$ is the number of channels, flatten it into patches $X_p \in \mathbb{R}^{N*C}$, where $P$ denotes the size of each patch and $N = HW/P^2$. The patches as a sequence are linearly projected $E$ into a $D$-dimention latent embedding space. The linear projection can be written as

$$T = [X_p^1 E, X_p^2 E, \cdots, X_p^N E] \tag{1}$$

In bidirectional encoder representations from transformers (BERT) [27] model, the first token is $[cls]$, it is a randomly initialized vector and represents classification result of the whole sequence after the encoder layer. Similar to BERT's class token, a learnable embedding vector $X_{cls}$ and position embedding $E_{pos}$ are extended. In Eq. 2, the final sequence $Z_0$ serves as the input of the transformer encoder.

$$Z_0 = [X_{cls}, T] + E_{pos} \tag{2}$$

The transformer encoder consists of a series of multi-head self-attention (MSA) and multi-layer perceptron (MLP) blocks. Before every block, layer normalization (LN) is applied. MSA consists of several attention layers in parallel to learn from different spaces, and the number of layers in MSA is $K$. The encoding procedure can be described as Eq. 3 and Eq. 4, where $Z_l$ denotes the encoded image representation. After multi layers' encoding, $X_{cls}$ is the global feature representation for final classification.

$$Z_l^{'} = MSA(LN(Z_{l-1})) + Z_{l-1}, l = 1, \cdots, L \tag{3}$$

$$Z_l = MLP(LN(Z_l^{'})) + Z_l^{'}, l = 1, \cdots, L \tag{4}$$

### 3.2 GRVT Architecture

The pretrained ViT model on high resolution and large datasets could facilitate to better transfer learning results on middle and small benchmarks [7]. In this sense, ViT could achieve promising performance in grocery recognition. However, the granularity of the patch scale is single, and it is unable to adapt the requirements for scenarios of stores where objects can vary a lot in images. To this end, we propose a simple but effective GRVT architecture that utilizes multi-scale patch embedding and mixed attention selection module to filter out high-confidence patches for final encoding. The overview of our framework is illustrated in Fig. 1.

### 3.3 Multi-scale Patch Embedding

Multi-scale feature representation can help networks better detect and recognize objects [8,25]. Chen et al. [28] proposed a dual-branch transformer called

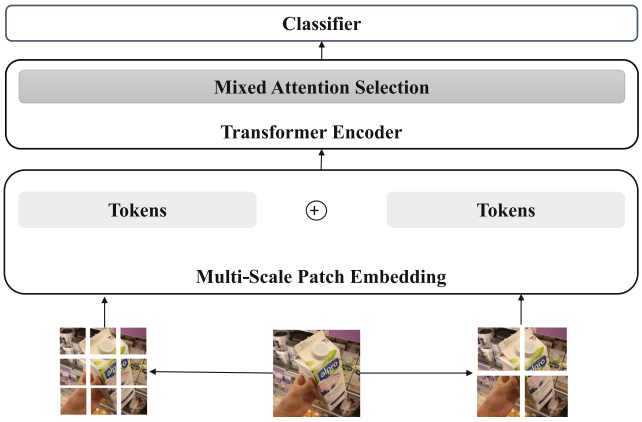



**Fig. 1.** The architecture of GRVT framework. Images are firstly split into a different scale of patches which are projected with their position embeddings and class embeddings. Image tokens are concatenated subsequently and input to the transformer encoder. Crucial and discriminative regions used for final classification are filtered by calculated attention weights from transformer encoder layers.

CrossViT that integrates different sizes of patches. Because the patch sizes are different in resolution, image patches are processed by separate encoder branches. Our MSPE module aims to integrate different granularity scales of patches as the input of the transformer encoder, with the same backbone architecture and very low computational cost. As shown in Fig. 2, let an image firstly be resized to $H_S \times W_S$ and $H_L \times W_L$. Patch number in each branch is $N_S = H_S W_S / P^2$, $N_L = H_L W_L / P^2$. For each patch embedding, we extend linear projection with their corresponding $X_{cls}$ and $E_{pos}$ following Eq. 1 and Eq. 2. Then we concatenate $Z_0^S$ and $Z_0^L$ as the final embedding of the encoder as Eq. 5. Dual granularities of token representations are complemented with each other.

$$Z_0 = [Z_0^S, Z_0^L] \tag{5}$$

### 3.4 Mixed Attention Selection

ViT feeds forward all patches across transformer encoder layer. Patch usually plays a different role in an image, it could be a part of the object, or invalid background. Noises can be harmful to the result if background patches are not filtered. The discriminative and crucial parts can guide the model to achieve better performance [17]. Here we propose a mixed attention selection module that fuses dual tokens' information and utilizes attention results across layers to filter out important parts. Figure 3 illustrates the design of MAS. The input of the encoder layers remains unchanged except for the last one. Multi-head self-attentions are calculated layer by layer, and $N$ means the number of patches and

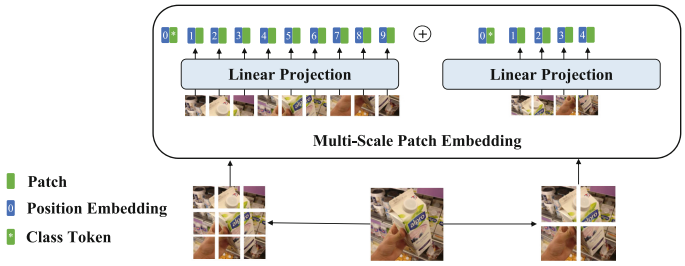



**Fig. 2.** Illustration of MSPE module. Images are divided into different scales of patches which are projected with position embedding and class embedding subsequently. Embeddings are concatenated into a sequence of tokens for the input of transformer encoder.

class tokens in Eq. 6. The attention weights $\boldsymbol{a}$ in every layer are accumulated by $K$ attention heads in Eq. 7 and Eq. 8.

$$N = 1 + N_S + 1 + N_L \tag{6}$$

$$a_l^i = [a_l^{i_1}, a_l^{i_2}, \cdots, a_l^{i_N}], i \in 0, 1, \cdots, K \tag{7}$$

$$\boldsymbol{a}_l = [a_l^0, a_l^1, \cdots, a_l^K], l \in 1, 2, \cdots, L-1 \tag{8}$$

Different tokens get mixed step by step across layers of the encoder, and the class token can guide the model to capture global representation. We follow TransFG [17] to obtain unified weights $\boldsymbol{a}_{final}$ by recursively multiplying raw attention matrix. The calculation can be written as

$$\boldsymbol{a}_{final} = \prod_{l=0}^{L-1} \boldsymbol{a}_l \tag{9}$$

The part selection of TransFG [17] selects index of maximum value in $\boldsymbol{a}_{final}$ for each attention heads, which does not adapt for the characteristics of multi-scale patch embedding. The MSPE introduces two sets of patch tokens. $N_S$ and $N_L$ are not quantitatively equal and single attention selection is not robust. To solve these problems, we develop the MAS module to handle mixed attention and conduct a balanced selecting strategy to choose crucial patches. As Eq. 10, $I$ denotes the balanced sampled top $M$ indexes of $K$ attention heads in different sets of patches respectively.

$$I = A_1^1, A_2^1 ... A_M^1, \cdots, A_1^K, A_2^K ... A_M^K \tag{10}$$

Then we concatenate corresponding patches in $Z_{L-1}$ as the final tokens in Eq. 11. GRVT focuses on the principle part to discover subtle differences with selected tokens. After the final layer, we integrate $X_{cls}^S$ and $X_{cls}^L$ as the final class token.

$$Z_{mix} = [X_{cls}^S, X_{cls}^L | Z_{L-1}^i], i \in I \tag{11}$$

$$Z_{cls} = mean(X_{cls}^S, X_{cls}^L) \tag{12}$$

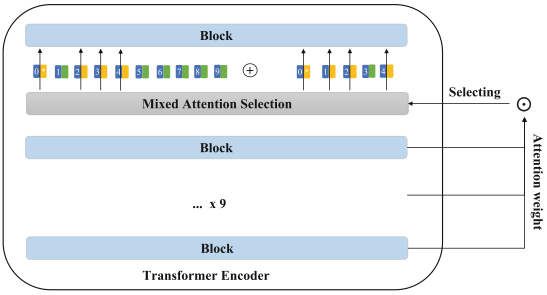



**Fig. 3.** Illustration of MAS module. Attention weights are gathered together and are calculated to filter out discriminative patches with a higher overall score.

## 4 Experiments

### 4.1 Datasets and Implementation Details

Freiburg Groceries Dataset (FGD) [11] is a challenging dataset due to the variety of objects and consists of 25 classes of groceries. The dataset mainly considers real-life scenes which include multi viewpoints of individual objects and packed shelves. Photos are taken in sophisticated lighting conditions with reflections and shadows at different stores. Grocery Store Dataset (GSD) [12] contains natural grocery items and refrigerated product images. The labels include coarse-grained and fine-grained classes with a hierarchical structure. It mainly includes vegetables, fruits, and packages. The coarse-grained class number is 43 and the fine-grained class number is 81. Furthermore, there are product descriptions with nutrition values to help visually impaired users in shops.

GRVT models are trained on $224 \times 224$ resolution images for fair comparison. Firstly, images are resized to $300 \times 300$ and then cropped to $224 \times 224$ for model inputs. For the MSPE module, images are resized $112 \times 112$ for another scale. The patch size $P = 16$ as the standard ViT-B-16 model. The learning rate is set to 0.03 with cosine annealing and the optimizer is stochastic gradient descent with a momentum of 0.9 and 0.0001 weight decay. The attention dropout rate is 0.1 and attention head number $K = 12$. The loss function is cross entropy loss and training iterations are 15,000. All experiments are performed with two Nvidia RTX 3090 GPUs, using Pytorch 1.9.0 with CUDA 11.1 and APEX with FP16 training.

### 4.2 Comparison with the State of the Art

We compare our GRVT against recent works, which are mostly based on CNN models that are trained end-to-end to classify images, or used as feature extractors for the subsequent decision rule. Table 1 shows the comparison results on FGD. ViT [7] performs better than public works' best result [20], and is closed to that of fine-grained approaches [14,17]. GRVT improves 0.63% accuracy than the standard ViT and exceeds the fine-grained approaches, which indicates the

**Table 1.** Comparison of different methods on FGD.

| Method | Backbone | Acc |
|---|---|---|
| Baseline [11] | AlexNet | 78.9 |
| ASAP [19] | NAS | 89.3 |
| XNAS [20] | NAS | 93.7 |
| ResNet50 [4] | ResNet | 91.94 |
| ResNet101 [4] | ResNet | 92.04 |
| WS-DAN [13] | Inception | 91.26 |
| DCL [14] | ResNet | 94.70 |
| ViT [7] | ViT | 94.53 |
| Swin-B [26] | ViT | **95.25** |
| PVT-M [25] | ViT | 93.02 |
| TransFG [17] | ViT | 94.92 |
| GRVT | ViT | 95.16 |

effectiveness of our method. As for comparison on GSD in Table 2, ViT outperforms the ensemble CNN models [22] and CNN based fine-grained methods [13,14] in both label structures. Nevertheless, our GRVT further achieves 0.66% and 0.4% improvement. Compared to TansFG [17], ours also shows the superiority. GRVT brings multi-scale patch representation and makes the model learn more diverse, robust patches from the different granularity of tokens. Besides, the MAS module makes regions with discriminative information can be focused on the last transformer encoder layer. Although Swin Transformer [26] performs best on FGD, but falls behind our GRVT on GSD, mainly because the GSD category is more closed to a FGVC task. PVT [25] needs fewer parameters and computation but suffers from accuracy results.

**Table 2.** Comparison of different methods on GSD.

| Method | Backbone | Acc/Fine | Acc/Coarse |
|---|---|---|---|
| DenseNet169 [12] | DenseNet | 85.0 | 85.2 |
| DN+MTL [18] | DenseNet | 89.13 | 94.33 |
| ResNet50 [22] | ResNet | 90.58 | 93.61 |
| ResNet101 [22] | ResNet | 92.55 | 94.87 |
| Ensemble [22] | CNNs | 93.48 | 95.84 |
| WS-DAN [13] | Inception | 87.67 | 91.43 |
| DCL [14] | ResNet | 93.36 | 95.25 |
| ViT [7] | ViT | 94.59 | 96.70 |
| Swin-B [26] | ViT | 93.12 | 95.58 |
| PVT-M [25] | ViT | 91.50 | 94.19 |
| TransFG [17] | ViT | 94.85 | 95.78 |
| GRVT | ViT | **95.25** | **97.10** |

Table 3 tabulates the model efficiency, measured by the number of parameters (#Params) and computational costs (FLOPs). Generally, the ViT-based models have more parameters, with better performance than ResNet models. TransFG [17] introduces an overlapping patch sampling method to avoid information loss around patch edges, however, the total number of the patch is increased a lot. Compared to ViT [7], for the same image resolution $224 \times 224$, TransFG brings 65% more patches to calculate, while GRVT needs 25% more computational cost.

**Table 3.** Comparison of computational efficiency.

| Method | #Params | FLOPs |
|---|---|---|
| ResNet50 [4] | 26M | 4.1G |
| ResNet101 [4] | 45M | 7.9G |
| ViT [7] | 87M | 17.6G |
| Swin-B [26] | 88M | 15.1G |
| PVT-M [25] | 44M | 6.7G |
| TransFG [17] | 87M | 27.8G |
| GRVT | 87M | 21.0G |

### 4.3 Ablation Study

In order to investigate the effectiveness of MSPE and MAS modules, as well as the impact of parameter $M$ in MAS, we conduct ablation studies in the following.

**Effect of MSPE.** MSPE introduces dual granularities of image patches. As shown in Table 4, it improves 0.31% and 0.42% evaluation result on FGD and GSD, respectively. The diversity of image patches improves the model's robustness and generalization capability.

**Effect of MAS.** MAS calculates attentions in encoder layers and samples those with high attention scores in a balanced way. Without considering noisy backgrounds, GRVT can focus on informative regions and capture important local features. Table 4 shows that MAS further achieves 0.32% and 0.24% improvement on two datasets.

**Impact of Parameter $M$.** MAS selects Top $M$ patch indexes according to the result of $\boldsymbol{a}_{final}$, where $M$ refers the number of patches to be chosen in each attention head. As shown in Table 5, the best accuracy is achieved when $M = 3$. If $M$ is too small, only a few patches can be selected for final encoding and predicting, which are not robust enough. If $M$ is too big, each attention head produces too many indexes which would be redundant and the result becomes inconspicuous.

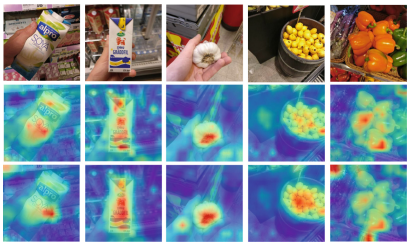
**Table 4.** Ablation studies of MSPE and MAS.

| ViT | MSPE | MAS | FGD | GSD |
|---|---|---|---|---|
| ✓ | | | 94.53 | 94.59 |
| ✓ | ✓ | | 94.84 | 95.01 |
| ✓ | ✓ | ✓ | **95.16** | **95.25** |

**Table 5.** Quantitative results of parameter $M$ in MAS module.

| Method | FGD | GSD |
|---|---|---|
| GRVT (M = 1) | 94.68 | 95.17 |
| GRVT (M = 3) | **95.16** | **95.25** |
| GRVT (M = 5) | 94.91 | 95.21 |

(a) Attention visualization on FGD



(b) Attention visualization on GSD

**Fig. 4.** Examples of visualization results on (a) Freiburg Grocery Dataset and (b) Grocery Store Dataset. The first row is the original image, the second row is the single branch attention map, and the third row is the attention map fused from two branches. Attention maps are overlaid on raw images for better visualization. (Color figure online)

### 4.4 Visualization

Figure 4 shows our visualization results on randomly selected images from both datasets. Attention maps from backbone encoder are transformed into the input space for better visualization. It can be observed that GRVT successfully captures discriminative regions such as product packages, edges, logos, and patterns, and also recognizes fruits and vegetables' shape, corners, and color. The background and shelf are purple and blue, while the important parts of products get more attention. For comparison of single branch and fused multi-scale attention maps, the latter shows more red and yellow regions on body of products, indicating the effectiveness of our GRVT.

## 5 Conclusion

In this work, we investigate the effectiveness of ViT on grocery recognition task, and demonstrate it can get better performance than CNN models. Furthermore, we propose a novel fine-grained grocery recognition framework named GRVT without introducing much computational cost, which outperforms recent works on grocery datasets. As GRVT achieves encouraging results, we believe that the transformer-based models have great potential for computer vision tasks, especially on grocery recognition. However, ViT models require more memory and computational resources, which may not suitable for lightweight needs and reality

applications. In the future, we will further study transformer-based lightweight and efficient models that are possible for deployment on mobile devices.

**Acknowledgements.** This work was supported in part by the National Natural Science Foundation of China under Grant 61902435, in part by the International Science and Technology Innovation Joint Base of Machine Vision and Medical Image Processing in Hunan Province under Grant 2021CB1013, and in part by the Fundamental Research Funds for the Central Universities of Central South University. We are grateful for resources from the High Performance Computing Center of Central South University.

## References

1. Wei, X.S., Cui, Q., Yang, L., Wang, P., Liu, L.: RPC: a large-scale retail product checkout dataset. arXiv preprint arXiv:1901.07249 (2019)
2. Leo, M., Furnari, A., Medioni, G.G., Trivedi, M., Farinella, G.M.: Deep learning for assistive computer vision. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11134, pp. 3–14. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11024-6_1
3. Wei, Y., Tran, S., Xu, S., Kang, B., Springer, M.: Deep learning for retail product recognition: challenges and techniques. Comput. Intell. Neurosci. **2020**, 23 (2020). https://doi.org/10.1155/2020/8875910. Article ID: 8875910
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
5. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 28. Curran Associates, Inc. (2015)
6. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
7. Dosovitskiy, A., et al.: An image is worth $16 \times 16$ words: transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
8. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2021)
9. Zheng, S., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6881–6890 (2021)
10. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12873–12883 (2021)
11. Jund, P., Abdo, N., Eitel, A., Burgard, W.: The freiburg groceries dataset. arXiv preprint arXiv:1611.05799 (2016)
12. Klasson, M., Zhang, C., Kjellström, H.: A hierarchical grocery store image dataset with visual and semantic labels. In: 2019 IEEE Winter Conference on Applications of Computer Vision, pp. 491–500. IEEE (2019)

13. Hu, T., Qi, H., Huang, Q., Lu, Y.: See better before looking closer: weakly supervised data augmentation network for fine-grained visual classification. arXiv preprint arXiv:1901.09891 (2019)
14. Chen, Y., Bai, Y., Zhang, W., Mei, T.: Destruction and construction learning for fine-grained image recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5157–5166 (2019)
15. Srivastava, M.M.: Bag of tricks for retail product image classification. In: Campilho, A., Karray, F., Wang, Z. (eds.) ICIAR 2020. LNCS, vol. 12131, pp. 71–82. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-50347-5_8
16. Ji, R., et al.: Attention convolutional binary neural tree for fine-grained visual categorization. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10468–10477 (2020)
17. He, J., et al.: TransFG: a transformer architecture for fine-grained recognition. arXiv preprint arXiv:2103.07976 (2021)
18. Ciocca, G., Napoletano, P., Locatelli, S.G.: Multi-task learning for supervised and unsupervised classification of grocery images. In: Del Bimbo, A., et al. (eds.) ICPR 2021. LNCS, vol. 12662, pp. 325–338. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68790-8_26
19. Noy, A., et al.: ASAP: architecture search, anneal and prune. In: International Conference on Artificial Intelligence and Statistics, pp. 493–503. PMLR (2020)
20. Nayman, N., Noy, A., Ridnik, T., Friedman, I., Jin, R., Zelnik, L.: XNAS: neural architecture search with expert advice. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
21. Wang, W., Cui, Y., Li, G., Jiang, C., Deng, S.: A self-attention-based destruction and construction learning fine-grained image classification method for retail product recognition. Neural Comput. Appl. **32**(18), 14613–14622 (2020). https://doi.org/10.1007/s00521-020-05148-3
22. Leo, M., Carcagnì, P., Distante, C.: A systematic investigation on end-to-end deep recognition of grocery products in the wild. In: 2020 25th International Conference on Pattern Recognition, pp. 7234–7241. IEEE (2021)
23. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)
24. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357. PMLR (2021)
25. Wang, W., et al.: Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578 (2021)
26. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
27. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
28. Chen, C.F.R., Fan, Q., Panda, R.: CrossViT: cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 357–366 (2021)