

A Review of Fine-Grained Image Recognition Techniques Based on Deep Learning

Zhexuan Dong ^a

School of Electronic and Information Engineering, Hebei University of Technology, Tianjin, China

Keywords: Fine-Grained Image Recognition, Strong Supervision, Weak Supervision, Deep Learning, Transformer.

Abstract: Fine-grained image recognition aims to visually recognize different subcategories of traditional semantic categories in images at a fine-grained level. It holds significant scientific value and promising application prospects across various fields such as biological classification, security monitoring, smart retail, medical diagnosis, and industrial manufacturing. Although fine-grained image recognition has achieved significant results with the support of deep learning methods, its dependence on large-scale, high-quality fine-grained image data has become the main bottleneck limiting the promotion and popularization of this technology. This paper focuses on fine-grained image recognition, introduces the relevant classical public data sets in five fields, and then introduces the FGIR method based on strong supervision and the FGIR method based on weak supervision. It is concluded that the choice of backbone network has a significant impact on the performance of fine-grained image recognition. Finally, the emerging trends and future directions of fine-grained image recognition are analyzed and concluded.


1 INTRODUCTION

Fine-grained image recognition plays an important role in computer vision and pattern recognition research, aiming at visual recognition of different subcategories under a traditional semantic category, such as different kinds of birds and dogs in the field of biometrics, different kinds of cars and aircraft in the field of security monitoring, product recognition in smart retail, and different diseases in the field of medical diagnosis. It is widely used in the field of industrial manufacturing to identify whether there is a defect. Recently, research on fine-grained image recognition has focused on how to extract subtle but recognizable component level information from images and obtain image features that accurately represent fine-grained differences. With the continuous development of these methods, recognition accuracy and application effectiveness continue to improve. However, obtaining large-scale high-quality fine-grained image data remains a huge challenge. Especially in certain professional fields, data annotation often requires the participation of domain experts, which not only requires a lot of manpower and financial resources, but also becomes

an important obstacle to promoting the popularization of fine-grained image recognition technology. The research objective of this article is to explore the advantages and disadvantages of fine-grained image recognition technology based on deep learning. This exploration will be conducted on classic datasets in five major fields. The article will also compare the efficiency of fine-grained image recognition under strong supervision and weak supervision. Additionally, the article aims to explore the future development direction of fine-grained image recognition.

2 TYPES OF FINE-GRAINED DATASETS

As shown in Table 1, there are already multiple specialized datasets for birds and dogs in the field of biological classification, such as CUB-200-2011 and Stanford Dogs, which have shown good results. The iNaturalist dataset covers a wide range of species, but due to its numerous types, training for a specific species is not very convenient. Although it performs

^a <https://orcid.org/0009-0001-7834-6277>

well in multi species classification, it may not provide sufficient fine classification support for fine-grained discrimination, resulting in insufficient training performance on certain specific species. In the field of security monitoring, there are abundant datasets related to automobiles, such as Stanford Cars and VehicleID, which are widely used for vehicle recognition and tracking. In addition, monitoring data involving facial and sensitive information is often restricted by privacy protection regulations, which limits data sharing and cross scenario applications, becoming the main obstacle to data circulation and application. In the field of smart retail, the construction of datasets needs to cover complex shopping scenarios, such as customer behavior recognition, product display optimization, and customer flow analysis tasks. However, the diversity of the retail environment, frequent updates of product categories, and customer privacy protection requirements have led to a significant increase in data collection and labeling costs. Meanwhile, fine-grained action recognition (such as distinguishing between picking up products and browsing behavior) requires high-precision labeling, further increasing the difficulty of data preparation. In the retail goods field, there are large-scale datasets such as Products 10k, as well as small datasets such as RPC, while in the fashion field, there are datasets such as DeepFashion2. In the field of medical diagnosis, there are datasets such as ChestX-ray14 for the chest and

ISIC for the skin. However there is a certain degree of annotation error, although there are expert annotations, the diagnosis of medical imaging itself has a certain subjectivity, especially for early diseases that may be difficult to distinguish. At the same time, Most medical image datasets are comparatively limited in size, especially fine-grained classification tasks that necessitate a substantial volume of labeled data, and medical datasets are often difficult to collect due to privacy issues. In the field of industrial manufacturing, tasks such as industrial defect detection rely on a large number of high-quality labeled samples, but the proportion of defective samples in actual production is extremely low, resulting the positive and negative samples are unbalanced. In addition, high-precision inspection of industrial equipment requires strong consistency in labeling, and equipment differences and process changes between different production lines make model generalization challenging. The data set mainly identifies surface defects such as MVTec AD (Anomaly Detection) and Severstal-steel-defect. To sum up, most of the fine-grained data sets still have certain limitations. Some types of samples are small, some data sets may have different image quality and lack detailed component annotation. The phenomenon of data imbalance may lead to the bias of the model, affect the accuracy, and lead to the degradation of the model performance.

Table 1: Fine grained image dataset

| field | Meta-class | Name | Year | Number of Images | Number of classes |
|---------------------------|--------------------|------------------|------|------------------|-------------------|
| Biological classification | Birds | CUB-200-2011 | 2011 | 11,788 | 200 |
| | | Birdsnap | 2014 | 49,829 | 500 |
| | | NABirds | 2015 | 48,562 | 555 |
| | Dogs | Stanford Dogs | 2011 | 20,580 | 120 |
| | | Dogs-in-the-wild | 2018 | 2,99,458 | 362 |
| | Flower | Oxford Flowers | 2008 | 8,189 | 103 |
| | Plants and Animals | iNaturalist | 2017 | 675170 | 5,089 |
| | | | 2018 | 461939 | 8,142 |
| | | | 2021 | 3,286,843 | 10,000 |
| Security monitoring | Airplanes | FGVC-Aircraft | 2013 | 10,000 | 100 |
| | Vehicles | Stanford Cars | 2013 | 16,185 | 196 |
| | | CompCars | 2015 | 30955 | 431 |
| | | VehicleID | 2016 | 221763 | 250 |
| smart retail | Fruits | Fru92 | 2017 | 69,614 | 92 |
| | Vegetables | Veg200 | 2017 | 91,117 | 200 |
| | Food Dishes | Food-101 | 2014 | 1,01,000 | 101 |
| | Retail Items | RPC | 2019 | 83,739 | 200 |
| | | Products-10k | 2020 | 1,50,000 | 10,000 |
| | | DeepFashion | 2016 | 800000 | 50 |
| | Clothes | FashionAI | 2018 | 357000 | 41 |
| | | DeepFashion2 | 2019 | 491000 | 13 |
| | Chest | ChestX-ray14 | 2017 | 112,120 | 14 |

| | | | | | |
|--------------------------|-----------------|------------------------------|------|--------|----|
| medical diagnosis | Skin | HAM10000 | 2018 | 10015 | 7 |
| | | ISIC | 2019 | 25331 | 10 |
| | | | 2020 | 33,126 | 23 |
| | | | 2021 | 35000 | 32 |
| Industrial manufacturing | Surface defects | MVTec AD (Anomaly Detection) | 2019 | 5,354 | 15 |
| | | DCT-Defect Dataset | 2021 | 4000 | 5 |
| | | Severstal-steel-defect | 2023 | 6,666 | 4 |

3 STRONG SUPERVISED FINE-GRAINED IMAGE RECOGNITION

Strong supervised learning is a machine learning method that trains models with a large amount of high-quality label data. Its main features are the integrity of labels and the high quality of training data. First, strongly supervised learning relies on complete label information, and each sample is accompanied by detailed annotations (such as classification labels, bounding boxes, or semantic segmentation masks, etc.), which enables the model to fully learn the mapping relationship between input and output. Second, the training data is of high quality and each sample is clearly labeled, thus ensuring that the model can accurately extract sample features and learn its categories. Under this framework, strongly supervised learning presents several advantages. The first is high-precision performance. By making full use of the complete label data, the model can learn fine-grained features, thus achieving excellent performance in tasks such as classification and detection. The second is the simplicity of the optimization process. The clear supervision signal makes the optimization direction of the model clear and the training process intuitive. However, strongly supervised learning also faces significant limitations. On the one hand, the cost of data annotation is high, especially when it involves fine-grained classification tasks, and the high standard requirements of manual annotation for categories and details lead to huge investment in manpower and resources. On the other hand, its scalability is poor, as the amount of data increases, the cost of label generation and collection also increases exponentially. Therefore, how to reduce the cost of labeling while maintaining model performance has become an important issue in the study of strongly supervised learning. Research has shown that under the strong supervision of various forms of manual annotations, FGIR achieves better

recognition accuracy, exceeding the performance of earlier methods relying on manual features. A partial region-based convolutional neural network (R-CNN) algorithm proposed by Zhang et al (Zhang,2014). It is based on the classic target detection framework R-CNN and has become a basic tool for fine-grained recognition tasks (Girshick, Donahue, Darrell, 2014). It uses a selective search algorithm to generate object candidate boxes in fine-grained images (Uijlings, van de Sande, Gevers, 2013). Huang et al designed a Part-Stacked CNN (PS-CNN) to extract semantic features that effectively represent image categories and capture discriminative information within the same semantic part (Huang, 2016). This method captures the features of the entire object and rich feature information from different semantic parts.

4 WEAKLY SUPERVISED FINE-GRAINED IMAGE RECOGNITION

Weakly supervised learning means that models use partial and incomplete labeling during the training process, rather than using fully labeled data as in strongly supervised learning. As an important method to solve machine learning problems under the condition of scarce label data, weakly supervised learning has significant characteristics and advantages. First of all, its core feature lies in the limited nature of label information. Unlike strongly supervised learning that relies on large-scale accurate labeling data, weakly supervised learning uses a small amount of label data to combine unlabeled or partially labeled samples to infer, which can effectively reduce the need for labeling resources; In addition, its learning method is flexible and can establish an inference mechanism for unlabeled data under limited labeling conditions. According to the scarcity and incompleteness of labels, weakly supervised learning can be divided into three main types: one is image-level labeling, which only provides image category

information (such as "cat" or "dog"), but does not contain the exact location or part of the object information; The second is partial labeling, which only annotates the attributes of some samples or objects (such as the bounding boxes of some objects), rather than complete labeling; The third is sparse labeling, which refers to the fact that only a few samples in the training data have labeled information, while the majority of samples are unlabeled. In practical applications, the main advantages of weakly supervised learning include significantly reduced labeling costs and broad applicability, especially in scenarios where labeling is difficult to obtain or expensive. However, its drawbacks cannot be ignored: due to the incompleteness or inaccuracy of labels, the model may have difficulty learning fine-grained features, leading to a decrease in prediction accuracy; Meanwhile, its performance is highly dependent on the quality of labeled data, and noise or erroneous information in weak labels may significantly affect the learning performance of the model. Therefore, weakly supervised learning needs to balance label cost and model performance in practical applications, and design more robust algorithms based on domain characteristics.

4.1 Fine-grained Recognition Paradigm Based on Local Regions

The fine-grained recognition paradigm based on local regions refers to the use of local detection or segmentation techniques to locate key regions corresponding to fine-grained images, such as bird heads, tails, and wings. Local detection techniques are detected by bounding box annotations. Zhang et al. first proposed component-level bounding box annotations, and then trained a regional convolutional neural network (R-CNN) model as a critical region detector (Zhang, 2014). Local segmentation technology enhances the fine-grained recognition ability of classification subnets by segmenting local information such as masks, as segmentation replaces rough bounding box annotations and is completed at a more detailed pixel level. However, traditional fine-grained image recognition requires a large number of component level labels for training, which greatly limits the application scenarios of such methods. Therefore, an increasing number of methods using image level labels for training have emerged in subsequent research.

4.2 Fine-grained Recognition Paradigm Based on Attention Mechanism

In the scene of fine-grained image recognition, the fine-grained recognition paradigm of local areas has been widely used, but in many practical applications, some parts of some objects may be difficult to identify. Therefore, using attention mechanisms as to selectively process regions of interest without manual annotation has become a popular direction. As is well known, the attention mechanism plays a crucial role in the human perceptual system. Based on this characteristic, Fu et al. were the first to use attention mechanisms to improve the accuracy of fine-grained object recognition (Fu, 2017). Peng et al. proposed a multi-level attention model to obtain hierarchical attention information at the object level and component level (Peng, 2018). In recent years, the classic model Transformer for natural language processing, which combines mechanism, has become increasingly popular. This model can be trained in parallel and can capture global information (Ma, 2022). Compared to CNN, Transformer's image serialization is a completely new form. He et al. introduced TransFG as the first framework to introduce the visual Transformer architecture in fine-grained image recognition (He, 2022).

4.3 Fine-grained Recognition Based on High-Order Feature Coding

In the early stages of deep learning, image representation typically uses fully connected layer features. Later, due to the richer information contained in the feature mapping of the top convolutional layer, people gradually began to use top convolutional features for fine-grained image recognition. In fine-grained image recognition, the encoding technique of convolutional neural networks achieves better performance compared to fully connected ones. However, these encoding techniques are partially derived from high-order statistical encoding of features. The representation based on covariance matrix is a representative high-order feature interaction technique. A commonly employed approach is the bilinear convolutional neural network, which encodes an image through two deep convolutional networks and subsequently decodes second-order statistical representations, leading to a notable enhancement in the accuracy of fine-grained recognition. However, this method is prone to overfitting on large-scale datasets. In order to mitigate this challenge, Gao et al. used Tensor Sketch to reduce the feature dimensions (Gao, 2016). Yu et

al. used dimensionality reduction projection before bilinear mapping to alleviate the problem of dimensionality explosion (Yu, 2018).

4.4 Fine-grained Recognition Paradigm Based on Highly Ordered Spatial Relationship Recognition

The spatial positions and relationships between objects are important components of image features. The term "spatial relationship" refers to the relative positions of objects, including adjacency, overlap,

and containment. Different from existing localization methods that rely on local information, Krause et al. used ordered spatial information of local regions to extract distinguishable features for FGIR (Krause, 2015). This method is suitable for fine-grained images that only use category labels, minimizing the need for additional annotations. Qi et al. proposed a two-part method for FGIR (Qi, 2019). Firstly, they select locally distinguishable regions based on the spatial relationships between targets capture information, then perform FGIR based on the extracted distinguishing features.

Table 2: Fine-grained recognition results are compared in caltech-UCSD-birds-200-2011(birds)、Stanford Cars(cars)

| | Method | Skeleton | Accuracy | |
|--------------------|---------------------------|------------|----------|--------|
| | | | birds | cars |
| Strong supervision | Bilinear CNN | VGG-16 | 72.50% | |
| | Mask-CNN | VGG-16 | 85.4% | |
| | dropout & Batch-Normalize | VGG-16 | | 82.26% |
| | Huang et al. | ResNet-152 | | 89.80% |
| | PS-CNN | Alexnet | 76.20% | |
| | FCN Attention | GoogLeNet | 84.30% | |
| Weak supervision | Two-level Attention | VGG-16 | 77.90% | |
| | RA-CNN | VGG-19 | 85.30% | 92.50% |
| | NTS-Net | ResNet-50 | 87.50% | 93.90% |
| | MTAL | ResNet-101 | 88.90% | 95.20% |
| | FCN Attention | GoogLeNet | 82.00% | |
| | HGTrans | ViT | 91.60% | |

Table 2 summarizes the representative algorithms in the FGIR field, which are divided into strong supervised paradigm and weak supervised paradigm, and introduces their performance on bird, dog, automobile and other data sets in detail. This table describes neural networks such as AlexNet、VGG、ResNet and GoogleNet. In the same dataset and using the same backbone network, the strongly supervised paradigm typically outperforms the weakly supervised paradigm in fine-grained recognition performance, as can be seen in fully convolutional network FCN and similar methods. In weakly supervised methods without manual annotation, the choice of backbone network has a significant impact on recognition performance. Huang et al. obtained different FGIR performance using VGG-16 and ResNet-152 under the same experimental setup, indicating the possibility of exploring multiple convolutional layer networks to achieve potentially better results (Huang, 2023). HGTrans, combined with the Visual Transformer (ViT) model, has the highest recognition accuracy in the bird dataset in Table 2. It can be concluded that the automatic extraction of detailed features from images through

self attention mechanism effectively improves the accuracy and efficiency of fine-grained recognition.

5 CONCLUSIONS

Fine-grained image recognition constitutes a significant area of research within the field of computer vision. With the explosion of various deep learning models today, fine-grained image recognition has made great progress. This paper summarizes the relevant common data sets in five fields and discusses the fine-grained image recognition methods of strong supervision and weak supervision. We compared the research results of previous researchers in the fields of biological classification and safety monitoring. In fields such as medical diagnosis and industrial manufacturing, the problems of small sample size and high annotation cost in datasets have always constrained the training and application of models. Traditional supervised learning methods often rely on a large amount of manually annotated data, which is particularly difficult and expensive in certain professional fields.

Therefore, exploring visual Transformer models based on zero sample learning and unsupervised learning has important research value. In addition, developing efficient and lightweight fine-grained recognition algorithms is particularly crucial for future development, especially in situations where computing resources are limited. Building flexible and powerful recognition systems within the framework of a few sample and unsupervised learning will have a profound impact on the future.

REFERENCES

- Fu, J., Zheng, H., Mei, T., 2017. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.4438-4446.
- Gao, Y., Beijbom, O., Zhang, N., et al., 2016. Compact bilinear pooling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.317-326.
- Girshick, R., Donahue, J., Darrell, T., et al., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.580-587.
- He, J., Chen, J.N., Liu, S., et al., 2022. TransFG: a transformer architecture for fine-grained recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 36(1), pp.852-860.
- Huang, S., Xu, Z., Tao, D., et al., 2016. Part-stacked CNN for fine-grained visual categorization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.1173-1182.
- Huang, Y., 2023. Fine-grained vehicle recognition algorithm based on deep learning and its optimization analysis. Integrated Circuit Applications, 40(3), pp.270-273.
- Krause, J., Jin, H., Yang, J., et al., 2015. Fine-grained recognition without part annotations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.5546-5555.
- Lin, T.Y., RoyChowdhury, A., Maji, S., 2015. Bilinear CNN models for fine-grained visual recognition. In Proceedings of the IEEE International Conference on Computer Vision, pp.1449-1457.
- Liu, X., Xia, T., Wang, J., et al., 2016. Fully convolutional attention networks for fine-grained recognition. arXiv:1603.06765.
- Ma, Y., Zhi, M., Yin, Y., et al., 2022. A review of the application of CNN and Transformer in fine-grained image recognition. Computer Engineering and Applications, 58(19), pp.53-63.
- Peng, Y., He, X., Zhao, J., 2018. Object-Part Attention Model for Fine-Grained Image Classification. IEEE Transactions on Image Processing, 27(3), pp.1487-1500.
- Qi, L., Lu, X., Li, X., 2019. Exploiting spatial relation for fine-grained image classification. Pattern Recognition, 91, pp.47-55.
- Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., et al., 2013. Selective search for object recognition. International Journal of Computer Vision, 104(2), pp.154-171.
- Yang, Z., Luo, T., Wang, D., et al., 2018. Learning to navigate for fine-grained classification. In Proc. Eur. Conf. Comput. Vis., pp.420-435.
- Yu, C., Zhao, X., Zheng, Q., et al., 2018. Hierarchical bilinear pooling for fine-grained visual recognition. In Proceedings of the European Conference on Computer Vision (ECCV), pp.574-589.
- Yu, Y., Wang, J., 2023. Hybrid granularities transformer for fine-grained image recognition. Entropy, 25(4), p.601.
- Zhang, N., Donahue, J., Girshick, R., et al., 2014. Part-based R-CNNs for fine-grained category detection. In Computer Vision – ECCV 2014, Cham: Springer International Publishing, pp.834-849.