2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

# Interpretable Fine-grained BI-RADS Classification of Breast Tumors

Yi Xiao[1], Kuan Huang[2], Sihua Niu[3], Jianhua Huang[1*]

*Abstract*— **Fine-grained classification of breast tumors is crucial for early diagnosis and timely treatment. Most fine-grained visual classification approaches focus on learning 'informative' visual patterns, which depend on the attention of the network, instead of 'discriminative' patterns, which interpretably contribute to classification. In this paper, we propose to extract discriminative patterns from informative patterns by utilizing the prior information of the dataset. The proposed method can detect the rough contour of the tumor area without boundary ground-truth guidance. At the same time, different masks are generated from the rough contour to reflect prior information on breast cancer. Moreover, a soft-labeling approach is utilized to replace the original BI-RADS label. Our model is trained using image-level object labels and interprets its results via a rough segmentation of tumor parts. Extensive experiments show that our approach achieves a significant performance increase on our BI-RADS classification dataset.**

## I. INTRODUCTION

Breast cancer has been the most common type of cancer. Due to the improvement of early diagnosis, which leads to timely treatment before the disease spreads in the whole body, breast cancer is now one of the cancers with the best curative effect. Starting from AlexNet [5] , deep neural networks [3], [11] has achieved great success in coarse-grained classification. However, subtle inter-class variance and large intra-class variance limit their performance in fine-grained settings. Furthermore, there is a lack of research on machines to recognize tumors at a fine-grained level like radiologists can do according to the Breast Imaging-Reporting and Data System (BI-RADS) criteria of the American College of Radiology [10].

Fine-grained visual classification (FGVC) has suffered from three main obstacles: (1) Large intra-class variance. Objects belonging to the same category usually exhibit significantly different postures and viewpoints. (2) Small inter-class variance. Objects belonging to different categories may be very similar, except for some subtle differences. (3) Limited training data. Constructing large fine-grained datasets is still not easy today. It requires not only time and money but also manual labeling generated by human experts. When it comes to fine-grained medical image datasets, apart from the difficulties mentioned above, even collecting data is a considerable challenge.

Existing deep learning-based FGVC approaches can be roughly divided into two branches: the first one is to enhance the detailed feature representation ability of the backbone network to achieve fine-grained feature learning [12]. The second one is to introduce part locations or object bounding box annotations as an additional optimization objective or supervision besides the primary classification network with extra modules [2], [4], [13], [15].

In [8], a feature representation enhancement method called Bilinear CNNs is proposed. Bilinear CNNs use two independent CNNs to capture local differences in images and calculate the relationship between their features. This type of method using high-level information has been proven to enhance the extraction of more meaningful information [6], [7] but suffers from enormous redundancy in the calculation process.

In [2], [15], two approaches based on part locations are designed, consisting of two parts: a detection sub-network and a classification sub-network. In [2], the RA-CNN [2] imitates the Region Proposal Network (RPN) network in [9] and proposes to use an Attention Proposal Network (APN) network to locate the informative area in the image. In [15], the amount of information in a target area has a positive correlation with its probability of being the target category, and a training method that enables a navigator to detect the area with an enormous amount of information is proposed.

Unlike previous methods, we believe that informative areas proposed by the network can be misleading and that discriminative areas should be selected to guide the network. As mentioned above, most previous methods focus on mining fine-grained 'informative' features, usually where the network has the peak response, aiming to gain better classification performances. However, not many have noticed that 'informative' is not equal to 'discriminative' that is genuinely beneficial to classification. One reason is that trusting the ability of the network to focus on correct informative regions without reasonable interpretable explanation is risky, especially in medical application scenarios. Moreover, some previous works use the information from fine-grained human annotations. Despite decent results, the expensive fine-grained human annotations make previous methods less applicable in practice.

In this paper, we develop a straightforward approach to tackle problems of previous FGVC methods using a weakly supervised setting. We argue that recognizing an object can naturally be divided into two stages: roughly locating the whole extent of the object and roughly figuring out

*This work is supported by the National Natural Science Foundation of China (Grant No. 82071930)

[1]Yi Xiao and Jianhua Huang are with the Pattern Recognition and Intelligence System Research Center, Faculty of Computing, Harbin Institute of Technology, Harbin 150001, Heilongjiang Province, China xiaoyihit, jhhuang@hit.edu.cn

[2]Kuan Huang is with the Department of Computer Science, Utah State University, Logan, Utah, USA, 84341

[3]Sihua Niu is with the Department of Ultrasound, Peking University People's Hospital, Beijing 100044, China

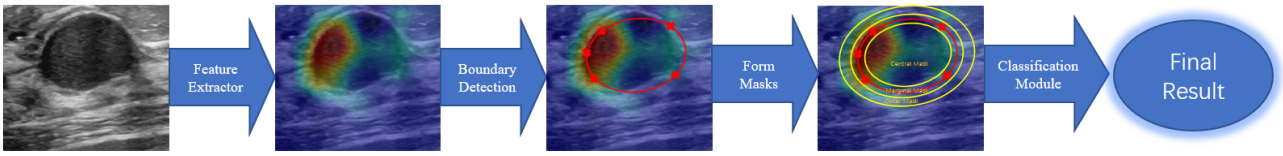


Fig. 1. Model overview. Proposal centers are firstly gained from a RPN. An ellipse is fitted as the rough position prediction of the tumor using the proposal centers. We then construct different interpretable masks by expanding and shrinking the ellipse and classify the tumor using the masks.

how the object of one category is different from objects in another category. Thus, we utilize a detection module to extract informative regions from the target ultrasound image without the human-generated detection annotations, after which we employ a novel approach to process the regions to form a mask that roughly fits the object, i.e., the tumor. A subsequent interpretable multi-branch classification module is attached to capture the features of different tumor areas and obtains fine-grained classification results for breast ultrasound images. Different tumor areas can provide useful medical knowledge for the final classification.

Our main contributions can be summarized as follows: 1) We propose a paradigm with three modules, namely feature extraction module, boundary detection module, and classification module, to perform the fine-grained BI-RADS classification. 2) A novel weakly-supervised approach to detect the regions to form a mask that roughly fits the object, different interpretable areas of the object are subsequently extracted, which can provide prior medical knowledge for the final classification. 3) A soft-labeling approach is utilized to handle label confusion in BI-RADS classification and make full use of prior knowledge of BI-RADS categories. 4) Our model can be trained end-to-end while providing accurate, fine-grained classification predictions as well as interpretable regions during inference.

## II. Method

Our method is based on the assumption that the informative regions from a neuron network can better characterize the object. In the proposed method, we transform informative regions into discriminative regions by using different parts of the region which contain the target object. As is shown in Fig. 1, our method consists of three continuous steps: Firstly, we utilize a feature extractor to extract the features of the image. Secondly, the outline of the object is extracted from the features. Finally, a classification module takes different parts of the object as input and outputs its final classification results.

### A. Feature Extraction Module

Consider a multi-class image classification task with $M$ classes. Define $I$ to be a training image with ground-truth label $l_J$, where $J = \{1, 2, ..., M\}$ is the label set containing all labels. Feature extractor takes $I$ as input, and outputs feature map $F_{full} \in \mathbb{R}^{C \times H \times W}$. Here, $C$, $H$, and $W$ refer to the number of channels, the height, and the width of the output activation maps. Our method uses a simple ResNet-50 without the final fully connected layer classification head as our extractor for a fair comparison with the literature.

### B. Boundary Detection Module

In order to guide the network to focus on correct regions, we first use a simple RPN with three scales. The RPN takes $F_{full}$ as input to propose a series of proposal centers $p \in \mathbb{R}^{2 \times n}$ as output, where $n$ represents the number of proposals. $p$ stores the coordinates of the centers of the proposals. Unlike previous object detection methods that take the proposals as the object position, we treat the proposals as the information areas used by the network to classify the entire input image. However, the information areas should not be treated as a whole part. Different regions in the information areas can provide different medical knowledge for final classification. The meaning of proposals can differ significantly under different dataset settings. In our BI-RADS classification problem for breast tumors, since malignant tumors usually show higher-margin coarseness and those benign tumors show the opposite, human experts practically judge tumors first by observing their margin. We believe that an interpretable classifier should share this experience. The proposals from a good classifier should focus on the information-rich margin of the tumor.

Based on this assumption, we predict a coarse margin of the object tumor from the proposals to provide medical knowledge to the classifier. We firstly store the centers of the proposals from the RPN (marked as the red crosses in Fig. 1). We then fit an ellipse using the convex hull of $n(n \geq 5)$ proposal centers. To minimize the error of fitting the ellipse, we limit the spatial distribution of proposals to prevent them from being too dense.

### C. Classification Module

As shown in Fig. 2, the proposed classification module consists of three parts: 1) capture medical features based on the prior masks constructed with the rough boundary from the previous subsection, 2) an attention module to predict the importance of each part in the medical knowledge features, and 3) a classifier with soft-labeling for BI-RADS classification, which will be illustrated in Subsection II-D.

Firstly, we capture novel features with medical information. We construct different masks with the ellipse from the detection step to better identify the tumor. Three masks are generated: marginal mask, central mask, and outer mask. The marginal mask aims to capture the morphological features such as circularity, margin spicules, margin coarseness, margin indistinctness, margin lobulation, etc.; the central mask

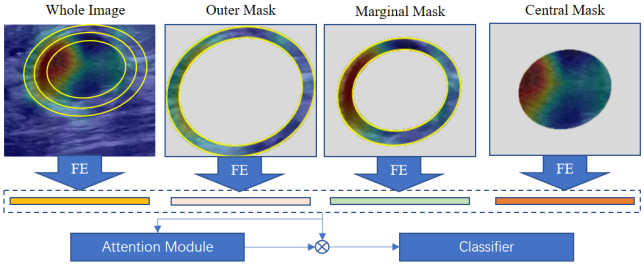


Fig. 2. Training method of classification module. After gaining the masks, we crop the regions from the whole image, resize them to the pre-defined size and feed them into feature extractor(FE) and global average pooling to gain features $F_{margin}$, $F_{outer}$, $F_{center}$, respectively. All three feature maps are in $C \times 1$ dimension. We then concatenate those features with a feature generated from the input image by feature extractor and global average pooling and utilize an attention module to describe the importance of each region.

aims to capture features like internal calcification, internal echo heterogeneity, and other texture features; the outer mask aims to capture features of surrounding areas.

In order to construct these three masks, we firstly expand and shrink the ellipse by a controllable parameter $k$ to create the boundaries of these masks. In our settings, marginal mask represents area between ellipses expanding and shrinking by $k_1^m = 1.2$ and $k_2^m = 0.6$, for outer mask, $k_1^o = 1.4$ and $k_2^o = 1$, and for central mask, we use the original ellipse as the output, i.e. $k^c = 1$. By using different interpretable masks, the network can better capture the differences between classes instead of focusing on the most informative areas.

After obtaining the masks, we clip different tumor parts according to the masks and calculate their features with the feature extractor introduced in Subsection II-A and global average pooling with shared parameters for simplicity. Finally, $F_{margin}$, $F_{outer}$, $F_{center}$ are captured for margin mask, outer mask, and center mask, respectively. They all have $C \times 1$ dimension.

Then, we concatenate those features with the feature generated by global average pooling of $F_{full}$, which is represented by $F_{ave} \in \mathbb{R}^{C\times 1}$, and we name the obtained feature as $F = [F_{ave}, F_{margin}, F_{outer}, F_{center}] \in \mathbb{R}^{C\times 4}$.

In the second step, an attention module is attached on top of $F$ to predict the importance of each region. This is realized by a sub-network $f_{att}$, given by $att = \text{Soft-max}(f_{att}(F^{\mathrm{T}}))$, where $f_{att}$ consists of two bottleneck blocks with three $1\times 1$ convolutions with batch normalization and ReLU in-between. The result $att \in \mathbb{R}^4$ is further used for classification.

Finally, we re-weight the transformed region features $F$ using the attention vector $att$, followed by a linear classifier. Therefore, the final prediction is given by:

$$y = \text{Soft-max}(W \times F \times att) \tag{1}$$

where $W \in \mathbb{R}^{M\times C}$ is the weights of a linear classifier for $M$-way classification and $y \in \mathbb{R}^M$ represents the final prediction result. The attention serves as a modulator of the different features in $F$. Thus, large values suggest a more important feature for classification.

### D. Soft-labeling

Soft-labeling is a method of label embedding. In the past few decades, deep neural networks have greatly improved image classification performance, but they only focus on a single objective of accuracy. These methods treat all errors the same, which leads to a situation: the possibility of making mistakes is indeed less than before, but when the mistakes happen, they can also be absurd or catastrophic and more challenging to explain. This problem is even more critical in medical image processing. Therefore, severity for neural network errors should be introduced. We should give different levels of punishment for different errors. Another intuition shows that it is not appropriate to use one-hot labels on some images. For example, one single case can be classified as class BI-RADS 4b; however, we cannot say BI-RADS 4a or BI-RADS 4c is 100% wrong. In this case, the ordinary one-hot representation of labels is inappropriate.

Soft-labeling usually uses a mapping function $S(l_J)$ to associate classes with a representation that encodes class relationship information. In this paper, we utilize a mapping function $S_{soft}(l_J)$, which outputs the categorical distribution on the class to replace one-hot labels. The function can be given componentwise by:

$$S^a_{soft}(l_J) = \frac{\exp(-\beta d(a,c))}{\sum_{b\in l_J}\exp(-\beta d(b,c))} \tag{2}$$

where $\beta$ is a parameter controlling the 'softness' of the label; $a, c$ represent two classes in $l_J$ (in this research, we have 6 BI-RADS classes); a distance function $d(a,c)$ is defined on BI-RADS classes as: $d(a,c) = 1$ for two adjacent BI-RADS classes, $d(a,c) = 2$ for two non-adjacent BI-RADS classes with one interval BI-RADS class, $d(a,c) = 3$ for two non-adjacent BI-RADS classes with two interval BI-RADS classes, etc. Here we give some examples in Eq. (3):

$$\begin{aligned} d(\text{BI-RADS 2}, \text{BI-RADS 3}) &= 1 \\ d(\text{BI-RADS 2}, \text{BI-RADS 4a}) &= 2 \\ d(\text{BI-RADS 2}, \text{BI-RADS 4b}) &= 3 \\ d(\text{BI-RADS 3}, \text{BI-RADS 4a}) &= 1 \end{aligned} \tag{3}$$

## III. Experiments

### A. Dataset and Implementation

We conduct extensive experiments on our BI-RADS classification dataset with 1061 breast ultrasound images. The breast ultrasound images in our dataset are obtained from the patients in Peking University People's Hospital, Southeast University Zhongda Hospital, the First Affiliated Hospital of the Guangxi University of Chinese Medicine, and the First Affiliated Hospital of Zhengzhou University. The ethics committees of the four hospitals approved this study. Written informed agreements were obtained from all participants. All the doctors participated in the ultrasonic examinations. According to the ACR BI-RADS® Atlas Fifth Edition, two physicians who were blind to the pathological results and with more than 10 years experience in breast ultrasound diagnosis evaluated the suspicion for malignancy for all the

TABLE I

DETAILED RESULTS ON OUR BI-RADS DATASET

| | BI-RADS 2 | BI-RADS 3 | BI-RADS 4a | BI-RADS 4b | BI-RADS 4c | BI-RADS 5 |
|---|---|---|---|---|---|---|
| sample amount | 14 | 346 | 75 | 266 | 216 | 144 |
| $p_{mal}$(Ground Truth) | 0.00% | 0.58% | 12.00% | 64.66% | 90.28% | 97.92% |
| $p_{mal}$(Ours) | 10.53% | 7.42% | 25.93% | 65.96% | 84.62% | 94.53% |
| $p_{mal}$(ACR BI-RADS® Atlas) | $\approx$ 0.00% | $\leq$ 3.00% | 3.00 $\sim$ 30.00% | 30.00 $\sim$ 60.00% | 60.00 $\sim$ 95.00% | $\geq$95.00% |
| precision | 68.42% | 83.63% | 70.37% | 76.17% | 69.66% | 66.41% |
| recall | 92.86% | 94.51% | 50.67% | 67.29% | 75.46% | 59.03% |

TABLE II

RESULTS ON OUR BI-RADS DATASET

| method | acc |
|---|---|
| DFL-CNN [14] | 73.80 |
| NTS-Net [15] | 73.70 |
| DCL [1] | 72.38 |
| LIO [16] | 74.18 |
| ResNet-50 [3] | 73.70 |
| Ours | **75.87** |

TABLE III

ABLATION STUDY ON OUR BI-RADS DATASET

| method | acc |
|---|---|
| w/o object detection | 74.93 |
| w/o soft-labeling | 74.65 |
| ResNet-50 | 73.70 |
| full model | **75.87** |

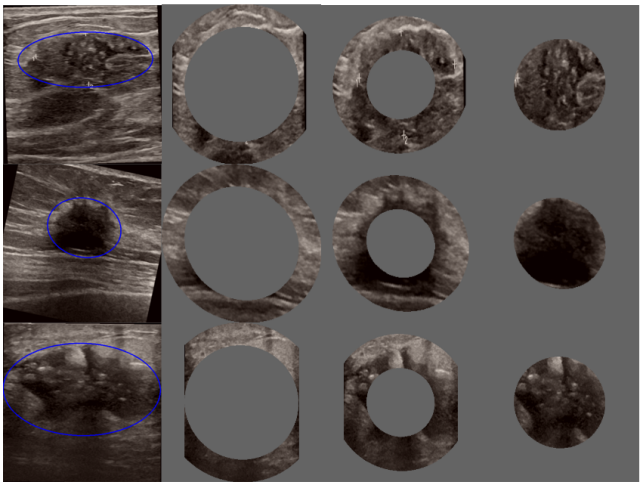

Fig. 3. Visualization of masks. From left to right shows full image, outer mask, marginal mask, central mask respectively. Ellipse in full image shows the boundary prediction of the tumor. All masks are reshaped into squares for better visualization.

lesions separately. Ultrasonic equipment includes Aplio 500, GE Logic E9, Philips IU22, and Siemens S3000. Our dataset is severely unbalanced and noisy, which makes our dataset difficult to do the classification task. We only use the BI-RADS classification label during training. The distribution of our data is shown in the first two rows in Table I.

The input images are resized to a fixed size of 512 $\times$ 512 and randomly cropped into 448 $\times$ 448 for scale normalization. We adopt random rotation and horizontal flip for data augmentation. All the above transformations are standard in the literature. We use ResNet-50 as the backbone of all models for simplicity and fair comparison with the literature. All models are trained for 240 epochs to ensure complete convergence. We use stochastic gradient descent (SGD) optimizer with momentum and an initial learning rate of $10^{-3}$. We use 5-fold cross-validation to evaluate our method.

### B. Results

For the multi-class BI-RADS classification task, we evaluate the precision and recall/sensitivity of every class. We report the probability of a single case in each class being malignant ($p_{mal}$) based on pathological results in Table I. Our result is roughly compliant with the BI-RADS Atlas Fifth Edition except for class BI-RADS 2 because we only contain a limited number of images in BI-RADS 2.

We also visualize the masks captured in Subsection II-C to show the interpretability of our model (refer to Fig. 3). Our model can roughly fit the tumor, and the masks successfully capture the corresponding features such as internal calcification in the central mask.

Finally, We compare our results with recent FGVC methods. All methods use ResNet50 as the backbone network. As is shown in Table II, our method outperforms recent FGVC methods by a large margin under a realistic medical dataset setting using classification accuracy as the metric. The bold values in Table II are the best result.

### C. Ablation Studies

We conduct an ablation study to evaluate our model components. Our study considers two variants: one without object detection and one without soft-labeling. The results is shown in Table III. Soft-labeling improves accuracy by 0.95%. Object detection improves accuracy by 1.23%, which we believe can be further improved by a fine-grained boundary prediction method in future works. Our proposed method with the weakly-supervised object detection to provide medical knowledge and soft-labeling obtains the best result.

Our method helps the neural network focus on discriminative parts of the image instead of random patches proposed by the network alone. Soft-labeling significantly improves performance than one-hot labeling by appropriately representing the characteristics of the input image.

### D. Discussion

Recent weakly-supervised FGVC works either enhance the detailed feature representation ability of the backbone network or attempt to discriminate similar classes by informative areas proposed directly by the network. However, they fail to solve the problems of FGVC: large intra-class variance and little inter-class variance. We believe that these problems are caused by lacking concentration on discriminative areas, which can be more beneficial to classification. By transforming 'informative' areas into 'discriminative' areas, the proposed work improves how networks learn relevant features for classification without enhancing the backbone network. Also, by making full use of prior knowledge of BI-RADS categories, a soft-labeling method is proposed to reduce severe misclassification and better represent the characteristics of different types of tumors. Both methods are proven to be effective by experiments.

However, failure cases can be found on images with more than two tumors due to failure to fit the ellipse. Also, our method relies on the accuracy of the RPN, which reduces stability. Future works will focus on the stability of our method and other ways to fit the boundary.

## IV. Conclusion

There are two key points in FGVC: 1) locating discriminative areas instead of informative areas, 2) fully mining and utilizing the prior knowledge of the dataset. In this paper, we propose an interpretable paradigm to capture discriminative features from images by roughly locating the object and producing semantic masks under a weakly supervised setting. We further utilize prior information of the dataset by soft-labeling instead of one-hot labeling. Our method outperforms some previous FGVC methods in BI-RADS classification on our dataset. Meanwhile, our method reports the probability of one tumor in each BI-RADS class being malignant tumor compliant with the ACR BI-RADS® Atlas. The study reported herein opens doors to a new way of understanding FGVC tasks.

## References


[1] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5157–5166, 2019.

[2] Jianlong Fu, Heliang Zheng, and Tao Mei. 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*, pages 4438–4446, 2017.

[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*, pages 770–778, 2016.

[4] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. Fine-grained recognition without part annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5546–5555, 2015.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[6] P. Li, J. Xie, Q. Wang, and Z. Gao. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[7] P. Li, J. Xie, Q. Wang, and W. Zuo. Is second-order information helpful for large-scale visual recognition? In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[8] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[10] Edward A Sickles, Carl J D'Orsi, Lawrence W Bassett, Catherine M Appleton, Wendie A Berg, Elizabeth S Burnside, et al. Acr bi-rads® atlas, breast imaging reporting and data system. *Reston, VA: American College of Radiology*, pages 39–48, 2013.

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

[12] 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*, pages 1–9, 2015.

[13] 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*, pages 3156–3164, 2017.

[14] Yaming Wang, Vlad I Morariu, and Larry S Davis. Learning a discriminative filter bank within a cnn for fine-grained recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4148–4157, 2018.

[15] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 420–435, 2018.

[16] Mohan Zhou, Yalong Bai, Wei Zhang, Tiejun Zhao, and Tao Mei. Look-into-object: Self-supervised structure modeling for object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11774–11783, 2020.