Weakly Supervised Fine-grained Image Classification via Guassian Mixture Model Oriented Discriminative Learning

Zhihui Wang[1,2], Shijie Wang[1], Shuhui Yang[1], Haojie Li[1,2]*, Jianjun Li[3], Zezhou Li[4]
[1]International School of Information Science & Engineering, Dalian University of Technology, China
[2]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, China
[3]School of Computer Science and Technology, Hangzhou Dianzi University, China
[4] Shanghai Pinlan Data Technology

Abstract

Existing weakly supervised fine-grained image recognition (WFGIR) methods usually pick out the discriminative regions from the high-level feature maps directly. We discover that due to the operation of stacking local receptive filed, Convolutional Neural Network causes the discriminative region diffusion in high-level feature maps, which leads to inaccurate discriminative region localization. In this paper, we propose an end-to-end Discriminative Feature-oriented Gaussian Mixture Model (DF-GMM), to address the problem of discriminative region diffusion and find better fine-grained details. Specifically, DF-GMM consists of 1) a low-rank representation mechanism (LRM), which learns a set of low-rank discriminative bases by Gaussian Mixture Model (GMM) to accurately select discriminative details and filter more irrelevant information in high-level semantic feature maps, 2) a low-rank representation reorganization mechanism (LR^{2}M) which resumes the space information of low-rank discriminative bases to reconstruct the low-rank feature maps. By recovering the low-rank discriminative bases into the same embedding space of high-level feature maps, LR^{2}M alleviates the discriminative region diffusion problem in high-level feature map and discriminative regions can be located more precisely on the new low-rank feature maps. Extensive experiments verify that DF-GMM yields the best performance under the same settings with the most competitive approaches, in CUB-Bird, Stanford-Cars datasets, and FGVC Aircraft.

1. Introduction

Weakly Supervised Fine-grained Image Recognition (WFGIR) focuses on distinguishing subtle visual differences under more detailed categories and granularity with only image-level annotations. WFGIR is still a challenging task due to two reasons. First, the global geometry and appearances of sub-categories can be very similar, and how to identify their subtle variances on the key regions is of vital importance. Second, instead of object or part annotations, WFGIR has only image-level annotations available which brings more difficulty in extracting effective and discriminative features to distinguish the subtle variances between subcategories.

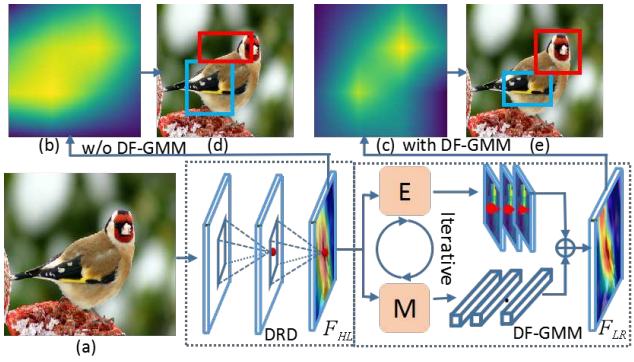

Figure 1. The motivation of Discriminative Feature-oriented Gaussian Mixture Model (DF-GMM). DRD denotes the problem of discriminative region diffusion. F_{HL} denotes the high-level semantic feature maps and F_{LR} indicates the low-rank feature maps. (a) is the original image, (b)(c) are the discriminative response maps to guide network to sample the discriminative regions and (d) (e) are localization results without and with DF-GMM learning, respectively. We can see that after reducing DRD, (c) is more compact and sparse than (b) and the resulted regions in (e) are more accurate and discriminative than those in (d).

Picking out the accurate discriminative regions plays the key role in addressing aforementioned two challenges of WFGIR. From this point, existing fine-grained image recognition approaches can be roughly grouped into

*Corresponding author: hjli@dlut.edu.cn. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants No. 61772108, No. 61932020 and No. 61976038.

three categories. One group localizes the object and local parts/patches by heuristic schemes [12, 13, 24, 32]. The limitation of heuristic schemes is that they cannot guarantee the selected patches are discriminative enough. Therefore, the second group tries to automatically localize the discriminative regions by using learning mechanism in an unsupervised or weakly supervised manner [8] [26] [30]. Instead of picking out discriminative regions independently, more recent works [27] [34] focus on designing end-to-end deep learning process to discover discriminative region group automatically via appropriate loss functions or correlation-guided discriminative learning.

All the previous works try to find discriminative regions/patches from high-level feature maps directly and neglect that the high-level feature map are constructed by fusing both spatial and channel-wise information within local receptive filed in CNN [15]. We argue that this could cause certain spatial propagation of discriminative and less-discriminative response and leads to the problem of discriminative region diffusion (DRD) in WFGIR, which aggravates the difficulty of discriminative region localization. As we can see from Figure 1, the diffused high-level feature map tends to distract the selection of discriminative regions, making the selected regions contain much noisy or background information and therefore degrade the performance of WFGIC.

Inspired by low-rank mechanism [7] [23] in natural language processing, we design a Discriminative Feature-oriented Gaussian Mixture Model (DF-GMM) framework to solve the problem of discriminative region diffusion and improve the WFGIR performance accordingly. The proposed DF-GMM consists of a low-rank representation mechanism (LRM) and a low-rank representation reorganization mechanism (LR^{2}M). The LRM is designed to select regions from the high-level feature maps to construct the low-rank discriminative bases. However, learning low-rank representation with LRM only forces the network to focus on the discriminative details rather than to consider the spatial context of discriminative regions. And the network has difficulty in selecting discriminative patches/regions without spatial information. Based on these consideration, the LR^{2}M is designed to resume the space information of low-rank discriminative bases and construct a new low-rank feature maps by linear weighted combining all low-rank discriminative bases. Comparing with the high-level feature maps, DF-GMM focuses on the discriminative details and distills the useless information on low-rank feature maps, which alleviates DRD problem and achieves better recognition accuracy.

The main contributions of this paper are listed as follows:

- To the best of our knowledge, we are the first to discover the problem of discriminative region diffusion in WFGIR.
- We propose an end-to-end discriminative feature-oriented Gaussian Mixture Model (DF-GMM) to learns low-rank feature maps to alleviate discriminative region diffusion problem and improve the WFGIR performance accordingly. This work also provides a generic framework to use other low-rank algorithms for WFGIR.
- We evaluate the proposed method on three challenging datasets (CUB-Bird, Stanford Cars, and FGVC Aircraft), and the results demonstrate that our DF-GMM achieves state-of-the-art.

2. Related Work

In the following, we will briefly review two lines of related work: feature representation and discriminative region localization.

Feature representation: End-to-end encoding approaches [9, 21, 16, 2, 5] encode the CNN features into high-order information. More recent advances reduce the high feature dimensionality [9] [16] and extract higher order information with kernel modules[2] [5]. Kernel-Pooling [2] defines Tayler series kernel and shows its explicit feature map can be compactly approximated. Kernel-Activation [5] designs the convolutional filter to select parts by the convolutional activations in a single spatial position. Due to the invariance to translation and posture of the object, these methods achieve better recognition accuracy.

Discriminative region localization: Recent WFGIR works mainly focus on designing end-to-end learning frameworks [6, 30, 33, 35]. S3Ns [6] produces sparse attention to localize object and discriminative parts by collecting local maximums of class response maps. TASN [35] learns subtle feature representations from hundreds of part proposals and uses an attention-based sampler to highlight attention regions. DCL [4] automatically detects the discriminative regions by region confusion mechanism. More recent [27] [34]works try to find discriminative region groups to improve discriminative ability for WFGIR. MA-CNN [34] proposes a part learning approach to implicitly select discriminative region group by a channel group loss, where part generation and feature learning can reinforce each other. CDL [27] establishes correlation between regions to discover the more discriminative region groups for WFGIR.

However, all the previous works try to find discriminative details from high-level feature maps directly and the problem of discriminative region diffusion is neglected. To address this, we propose an end-to-end Discriminative Feature-oriented Gaussian Mixture Model (DF-GMM) to reconstruct the low-rank feature maps. To our best knowledge, this is the first work to discover the problem of discriminative region diffusion for WFGIR and the first work

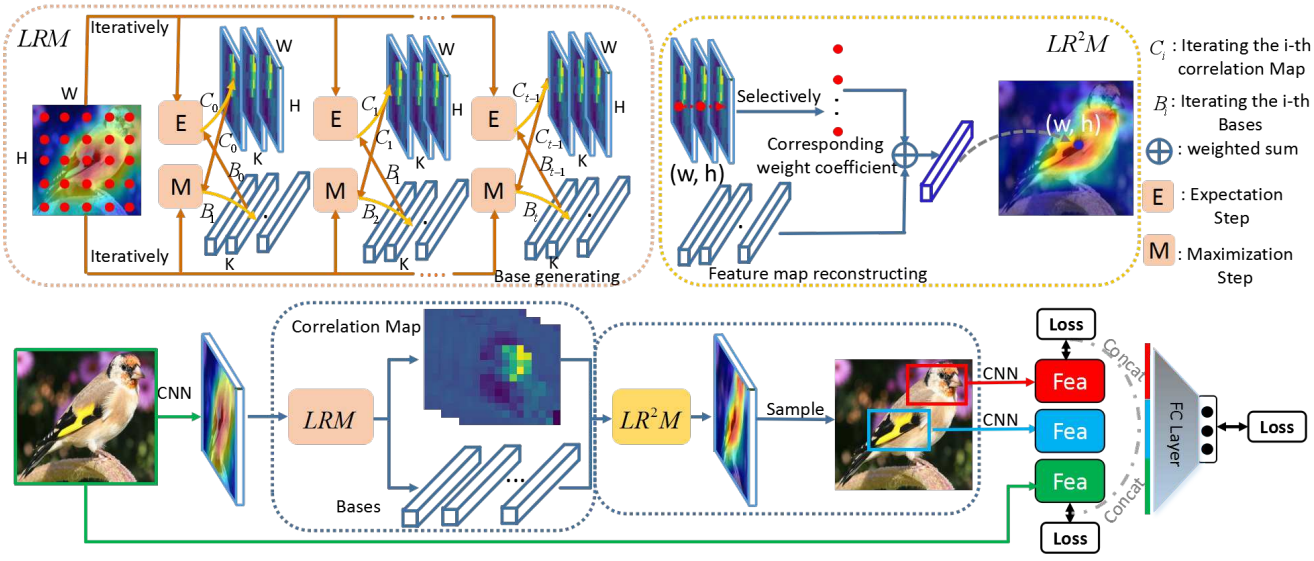

Figure 2. The framework of the proposed Discriminative Feature-oriented Gaussian Mixture Model (DF-GMM). DF-GMM first produces discriminative bases and linear weight correlation coefficient map by the low-rank representation mechanism (LRM). Then the low-rank representation reorganization mechanism (LR^{2}M) constructs the new low-rank feature maps by linear weighted combining all low-rank discriminative bases. At the sampling phase, the discriminative object patches are located by collecting local maximums from new low-rank feature maps. Next, we crop and resize the patches to 224×224 from the original image. Finally, the features of all branches are aggregated to produce the final recognition vectors. Note that the CNN parameters for all branches are shared.

to recognize the fine-grained image through exploring low-rank mechanism.

3. Proposed Method

As shown in Figure 2, the network of DF-GMM learns a set of discriminative bases from high-level semantic feature maps by Gaussian Mixture Model (GMM) in Low-rank Representation Mechanism (LRM), and then utilizes them to reconstruct low-rank discriminative feature maps by Low-rank Representation Reorganization Mechanism (LR^{2}M), which can be considered as the low-rank matrix recovery for alleviating discriminative region diffusion in high-level feature maps.

3.1. Low-rank Representation Mechanism

Our proposed Low-rank Representation Mechanism (LRM) is designed to learn regions from the high-level feature maps to construct the low-rank discriminative bases through Gaussian Mixture Model (GMM). The GMM consists of 1) feature-guided base initialization module, which makes low-rank bases more unique for each image in WFGIC, 2) expectation step (E-step) module, which computes the expected value of the linear weight correlation coefficients, 3) maximization step (M-step) module, which updates the low-rank bases by using the linear weight correlation coefficients weighted summation of high-level feature maps. M-step makes the low-rank bases lie in a low dimensional manifold.

Specifically, given an image X, we feed X into the CNN backbone and extract the high-level feature maps from the top convolutional layer. The high-level feature maps are indicated as $M_I \in \mathbb{R}^{C\times H\times W}$, where C, H and W denote the channel, height and width of feature maps. Then, M_I is fed into a Gaussian Mixture Model (GMM) function to get the low-rank discriminative bases μ and the linear weight correlation coefficients Z:

$$(\mu, Z) = GMM(M_I), \tag{1}$$

where $\mu \in \mathbb{R}^{C\times K}$ denotes the low-rank discriminative bases, K is the number of bases. $Z \in \mathbb{R}^{N\times K}$ indicates the linear weight correlation coefficients, and N equals to $W \times H$. Here Z is applied to select the discriminative regions to construct the low-rank discriminative bases.

Base Initialization: For fine-grained image recognition, there are thousand of images in the datasets. As each image has different discriminative region feature distributions from others, it is not suitable to use unified bases computed upon one image. We propose the initialization of low-rank bases is guided by high-level feature maps M_I. Concretely, M_I is fed to a Global Average Pooling (GAP) layer followed by a copy operation to obtain the feature matrix $V \in \mathbb{R}^{K\times C}$. With the weight matrix in GMM $W^m \in \mathbb{R}^{K\times C}$, we can compute the initialization of low-

rank bases μ by element-wise multiplication as follows:

$$\mu_{ij} = R_{ij} \odot W_{ij}^{m}, \tag{2}$$

where μ_{ij} denotes the j^{th} element in i^{th} base, R_{ij} is the j^{th} element in i^{th} vector and W_{ij}^{m} denotes the i^{th} row and the j^{th} column weight coefficient. Note that W^m is initialized by Kaiming's initialization [10].

Gaussian Mixture Model: Let M_I be reshaped into $M_I \in \mathbb{R}^{C\times N}$, where N equals to $W \times H$. Note that the discriminative bases μ can be regarded as the mean parameters in GMM and the linear weight correlation coefficients Z as latent variables. Then our task-related GMM can be defined as a linear superposition of Gaussian according to the distribution of data M_I:

$$p(M_I^n) = \sum_{k=1}^{K} Z_{nk}\mathcal{N}(M_I^n|\mu_k, \sigma_k^2), \tag{3}$$

where the covariance σ_k^2 is parameter for the k-th Gaussian basis, $M_I^n \in \mathbb{R}^{C\times 1}$ denotes the n^{th} vectors in high-level semantic feature maps M_I. The likelihood of the complete data $\{M_I, Z\}$ is formulated as:

$$ln\ p(M_I, Z|\mu, \sigma) = \sum_{n=1}^{N} ln\left[\sum_{k=1}^{K} Z_{nk}\mathcal{N}(M_I^n|\mu_k, \sigma_k^2)\right], \tag{4}$$

where $\sigma_k^2 \times Z_{nk} = 1$, Z_{nk} can be viewed as the responsibility that the k-th basis takes for the observation M_I^n. Concretely, we choose inner dot $\mathcal{K}$ as the general kernel function in GMM. Using $\mathcal{K}$, Eq. (4) is simplified to

$$ln\ p(M_I^n|\mu_k) = \sum_{n=1}^{N} ln\mathcal{K}(M_I^n, \mu_k), \tag{5}$$

where $ln\ p(M_I^n|\mu_k)$ indicates the posterior probability of M_I^n given μ^k.

For GMM, it contains two steps: an expectation step (E-step) and a maximization step (M-step).

E-Step: It aims to estimate the posterior distributions of the latent variables Z, i.e. $Z_{nk} = P(M_I^n|\mu_m, \theta^{old})$, by using the current estimated parameters $\theta^{old} : \{\mu^{(old)}, \sigma^2\}$. Specifically, the new expected value of Z_{nk} is given by:

$$Z_{nk}^{new} = \frac{\mathcal{N}(M_I^n|\mu_k^{(old)}, \sigma^2)}{\sum_{k=1}^{K}\mathcal{N}(M_I^n|\mu_k^{old}, \sigma^2)} \tag{6}$$

According to Eq. (5), Eq. (6) can be reformulated into a more general from:

$$Z_{nk} = \gamma \cdot \frac{ln\mathcal{K}(M_I^n, \mu_k)}{\sum_{k=1}^{K} ln\mathcal{K}(M_I^n, \mu_k)} \tag{7}$$

where γ is a learning rate parameter and is gradually learned to regulate the distribution of correlation weight coefficient matrix. In practice, there is a learning rate parameter γ for each Gaussian component.

$\mathcal{K}$ indicates the matrix multiplication between M_I^n and μ_k, while $\sum_{k=1}^{K} ln\mathcal{K}(M_I^n, \mu_k) = 1$. Now, Eq. (7) can be simplified to

$$Z^{(new)} = \gamma \cdot M_I \odot (\mu^{(old)})^T. \tag{8}$$

Then Z is passed through a softmax layer to normalize the weight correlation coefficient Z_{nk} in the n^{th} row and the k^{th} column of correlation weight coefficient matrix Z:

$$Z_{nk}^{(new)} = \frac{e^{Z_{nk}^{(new)}}}{\sum_{n=1}^{N}\sum_{k=1}^{K} e^{Z_{nk}^{(new)}}}. \tag{9}$$

M-Step: The parameters of GMM are re-estimated by likelihood maximization as follows:

$$\mu_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N} Z_{nk}^{(new)} M_I^n, \tag{10}$$

$$\sigma^2 = \frac{1}{N_k}\sum_{n=1}^{N} Z_{nk}^{(new)}(M_I^n - \mu_k^{old})(M_I^n - \mu_k^{old})^T, \tag{11}$$

where

$$N_k = \sum_{n=1}^{N} Z_{nk}^{(new)}. \tag{12}$$

M-step updates the low-rank discriminative bases μ by maximizing the complete data $ln\ p(M_I, Z, \theta)$, where θ is the set of all parameters of GMM. We re-estimated the low-rank bases μ through using the weighted summation of M_I with the latent variables $Z^{(new)}$. Therefore, Eq. (10) can be rewritten as:

$$\mu_k^{(new)} = \frac{Z_{nk}^{(new)} \cdot M_I^n}{\sum_{n=1}^{N} Z_{nk}^{(new)}}. \tag{13}$$

The Low-rank Representation Mechanism (LRM) executes the expectation step and maximization step alternately until the low-rank bases are the most discriminative.

3.2. Low-rank Representation Reorganization

Learning low-rank representation with LRM only forces the network to focus on the discriminative details rather than to consider the spatial context of discriminative regions. The network has difficulty in selecting discriminative patches/regions without spatial information. To deal with this limitation, we propose a Low-rank Reorganization Representation Mechanism (LR^{2}M) to resume the spatial information from the low-rank discriminative bases.

After the Gaussian Mixture Model is convergent, we reshape $Z \in \mathbb{R}^{N\times K}$ into $Z \in \mathbb{R}^{W\times H\times K}$ to make linear weight coefficients correspond with the space localization

Table 1. The stride, patch scale size, scale step and aspect ratios of the three different layers. M_D^1 and M_D^2 are feature maps after down-sampling M_D from the output of base decomposition. Note that the stride is the original image scaling ratio. Patch width & height = scale $\times$ scale step $\times$ aspect ratio.

Feature Map	Stride	Scale	Scale Step	Aspect ratio
M_D	32	32	$2^{\frac{1}{3}}, 2^{\frac{2}{3}}$	$\frac{2}{3}, 1, \frac{3}{2}$
M_D^1	64	64	$2^{\frac{1}{3}}, 2^{\frac{2}{3}}$	$\frac{2}{3}, 1, \frac{3}{2}$
M_D^2	128	128	$1, 2^{\frac{1}{3}}, 2^{\frac{2}{3}}$	$\frac{2}{3}, 1, \frac{3}{2}$

of original feature maps M_I. Given the low-rank discriminative bases μ and the linear weight coefficients Z, the vectors M_D^{wh} located at (w, h) in re-estimation feature maps M_D can be calculated as follows:

$$M_D^{wh} = \sum_{k=1}^{K} Z_{whk} \cdot \mu_k, \tag{14}$$

where Z_{whk} denotes the linear weight coefficient located at (w, h) and k^{th} channel value in Z. After all M_D^{wh} are computed, M_D is constructed from discriminative bases.

M_D has the low-rank property compared with the original input M_I. As Z keeps the mapping correlation between M_I and μ, M_D can resume the discriminative details with corresponding spacial information. Meanwhile, each feature vector in channel direction integrates all low-rank discriminative bases with different linear combinations, which can emphasize the discriminative regions while distill the false positive highlighting in original feature maps M_I.

3.3. Discriminative Information Sampling

We use the low-rank feature maps with three different scales to generate default patches, inspired by Feature Pyramid Network [20]. Table 1 shows the design details, containing the scale size, scale step and aspect ratio of default patches.

Let's take feature map M_D as an example. We feed the low-rank features M_D into a score layer. Concretely, we add a $1 \times 1 \times N$ convolution layer and a sigmoid function σ to learn discriminative response maps $R \in \mathbb{R}^{N \times H \times W}$, which indicates the impact of discriminative regions on the final classification, as follows:

$$R = \sigma(W_R * M_D + b_R), \tag{15}$$

where $W_R \in \mathbb{R}^{C \times 1 \times 1 \times H}$ represents the convolution kernels, H is the number of the default patches at a given location in the feature maps, and b_R denotes the bias. Meanwhile, we assign the discriminative response value to each default patch p_{ijk}:

$$p_{ijk} = [t_x, t_y, t_w, t_h, R_{ijk}], \tag{16}$$

where s_{ijk} denotes the value of the i^{th} row, the j^{th} column and the k^{th} channel, and (t_x, t_y, t_w, t_h) denotes each patch's coordinates. Finally, the network picks the top-M patches with a response value, where M is a hyper-parameter.

3.4. Loss Function

The full multi-task loss $\mathcal{L}$ can be represented as the following:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \cdot \mathcal{L}_{gud} + \lambda_2 \cdot \mathcal{L}_{rela} + \lambda_3 \cdot \mathcal{L}_{rank}, \tag{17}$$

where $\mathcal{L}_{cls}$ represents the fine-grained classification loss. $\mathcal{L}_{gud}$, $\mathcal{L}_{rela}$ and $\mathcal{L}_{rank}$ represent the guided loss, correlation loss and rank loss, respectively. The balance among these losses is controlled by hyper-parameter λ_1, λ_2, λ_3.

We denote the selected discriminative patches as $P = \{P_1, P_2, ..., P_N\}$ and the corresponding discriminative response values as $R = \{R_1, R_2, ..., R_N\}$. Then the guided loss and the correlation loss as well as the rank loss are defined as follows:

$$\mathcal{L}_{gud}(X, P) = \sum_{i}^{N} (max\{0, log\mathcal{C}(X) - log\mathcal{C}(P_i)\}), \tag{18}$$

$$\mathcal{L}_{rela}(P_c, P) = \sum_{i}^{N} (max\{0, log\mathcal{C}(P_i) - log\mathcal{C}(P_c)\}), \tag{19}$$

$$\mathcal{L}_{rank}(R, P) = \sum_{log\mathcal{C}(P_i) < log\mathcal{C}(P_j)} (max\{0, (R_i - R_j)\}), \tag{20}$$

where X is the original image and the function $\mathcal{C}$ is the confidence function which reflects the probability of classification into the correct category, P_c is the concatenation of all selected patch features.

The guided loss is designed to guide the network to select the more discriminative regions. The correlation loss can guarantee that the prediction probability of combined features is greater than that of single patch features. The rank loss strives for consistency of the discriminative scores and the final classification probability values of the selected patches, encouraging them in the same order.

3.5. Back-propagation in GMM

As the proposed DF-GMM is an end-to-end framework, the loss $\mathcal{L}$ in sec.3.4 can directly influence the parameter in GMM. Concretely, we calculate the derivatives of weight matrix W^m in low-rank bases μ:

$$\frac{\partial \mathcal{L}}{\partial W^m} = \frac{\partial \mathcal{L}}{\partial M_D} \cdot \frac{\partial M_D}{\partial M_I^n} \cdot \frac{\partial M_I^n}{\partial W^m}, \tag{21}$$

where the weight matrix can be modified through back-propagation to improve the internal discriminative ability of base elements.

Table 2. The ablative recognition results and speed of different variants of our method. We test the models on CUB-200-2011.

Method	Accuracy	Speed
BL [19]	84.5%	n/a
BL + Sample	86.2%	50 fps
BL + Sample + DF-GMM	88.8%	41 fps

We use Q to represent GMM module, which is a self-supervised clustering algorithm. According to Eq. (10) and Eq.(11), we have:

$$\frac{\partial Q}{\partial \mu_k} = \sum_{n=1}^{N} \frac{1}{\sigma_k^2}(M_I^n - \mu_k), \tag{22}$$

$$\frac{\partial Q}{\partial \sigma_k^2} = -\sum_{n=1}^{N} \frac{1}{2\sigma_k^2} + \sum_{n=1}^{N} \frac{1}{2\sigma_k^4}(M_I^n - \mu_k)^2, \tag{23}$$

It is obvious that covariance σ^2 and mean μ both can be adjusted indirectly by the learning process of network with feature M_I^n.

4. Experiments

4.1. Datasets

We comprehensively evaluate our algorithm on Caltech-UCSD Birds [1] (CUB-200-2011), Stanford Cars [18] (Cars) and FGVC Aircraft (Airs) [22] datasets, which are widely used benchmark for fine-grained image recognition. The CUB-200-2011 dataset contains 11,788 images spanning 200 sub-species. The ratio of train data and test data is roughly 1:1. The Cars dataset has 16,185 images from 196 classes officially split into 8,144 training and 8,041 test images. The Airs dataset contains 10,000 images over 100 classes, and the train and test sets split ratio is around $2:1$.

4.2. Implementation Details

In all our experiments, all images are resized to 448×448, and we crop and resize the patches to 224×224 from the original image. We use fully-convolutional network ResNet-50 as feature extractor and apply Batch Normalization as regularizer. We also use Momentum SGD with initial learning rate 0.001 and multiplied by 0.1 after 60 epochs. We use weight decay $1e^{-4}$. To reduce patch redundancy, we adopt the non-maximum suppression (NMS) on default patches based on their discriminative scores, and the NMS threshold is set to 0.25. According to the results of multiple experiments, the loss balance parameter can be set into $\lambda_1 = \lambda_2 = \lambda_3 = 1$. Note that the architecture in principle contains multiple CNN modules and for clarity, these CNN modules share the same parameters.

4.3. Ablation Experiments

We conduct ablation studies to understand the influence of different components in our proposed method. We design different runs on CUB-200-2011 dataset using ResNet-50 as the backbone network and report the results in Table 2.

First, the features are extracted from the original image through ResNet-50 [11] without any object or partial annotation for fine-grained recognition, and we set it as the baseline (BL) of our model. Then the default patches are selected as local features to improve recognition accuracy. However, massive redundant default patches result in the low recognition speed. When we introduce the score mechanism (Sample) to only preserve the highly discriminative patches and reduce the number of patches to single-digit, the top-1 recognition accuracy on CUB-200-2011 dataset improves 1.7% and achieves a real-time recognition speed of 50 fps. Finally, we take account into the problem of discriminative region diffusion through DF-GMM, and achieve the state-of-the-art result of 88.8%. Ablation experiments have verified that the proposed DF-GMM indeed learns the low-rank discriminative bases to precisely localizes the discriminative regions by solving the problem of discriminative region diffusion, thus effectively improves the recognition accuracy.

Table 3. Comparison of different methods on CUB-200-2011.

Method	Box	Part	Accuracy
PN-DCN [1]	BBox	Parts	85.4%
M-CNN [29]	n/a	Parts	84.2%
PG [17]	BBoxs	n/a	82.8%
SCDA [28]	n/a	n/a	80.1%
AutoBD [31]	n/a	n/a	81.6%
OPAM [24]	n/a	n/a	85.8%
Bilinear [21]	n/a	n/a	84.0%
Kernel-Pooling [5]	n/a	n/a	86.2%
NTS-Net [30]	n/a	n/a	87.5%
PA-CNN [36]	n/a	n/a	87.8%
DCL [4]	n/a	n/a	87.8%
TASN [35]	n/a	n/a	87.9%
CDL [27]	n/a	n/a	88.4%
S3Ns [6]	n/a	n/a	88.5%
StackDRL [14]	n/a	n/a	86.6%
KERL [3]	n/a	n/a	87.0%
Our DF-GMM	n/a	n/a	**88.8%**

4.4. Performance Comparison

Accuracy comparison. Our comparisons focus on the weakly supervised methods because the proposed model only utilizes image-level annotations. Table 3, Table 4 and Table 5 show the performance of different methods on CUB-200-2011 dataset, Stanford Cars-196 dataset and FGVC-Aircraft dataset, respectively. In each table from top to bottom, the methods are separated into six groups, which are (1) supervised multi-stage methods, (2) weakly supervised multi-stage frameworks, (3) weakly supervised

Table 4. Comparison of different methods on Stanford Cars-196.

Method	Annotation	Accuracy
PG [17]	BBoxs	92.8%
SCDA [28]	n/a	85.1%
AutoBD [31]	n/a	88.9%
OPAM [24]	n/a	92.2%
Bilinear [21]	n/a	91.3%
Kernel-Pooling [5]	n/a	92.4%
PA-CNN [36]	n/a	93.3%
NTS-Net [30]	n/a	93.9%
TASN [35]	n/a	93.8%
CDL [27]	n/a	94.2%
DCL [4]	n/a	94.5%
S3Ns [6]	n/a	94.7%
DT-RAM [19]	n/a	93.1%
Our DF-GMM	n/a	**94.8%**

Table 5. Comparison of different methods on FGVC-Aircraft.

Method	Annotation	Accuracy
BoT [25]	BBoxs	88.4%
SCDA [28]	n/a	79.5%
Kernel-Pooling [5]	n/a	85.7%
LB-CNN [16]	n/a	87.3%
Kernel-Activation [2]	n/a	88.3%
PA-CNN [36]	n/a	91.0%
NTS-Net [30]	n/a	91.4%
DFL-CNN [26]	n/a	92.0%
S3Ns [6]	n/a	92.8%
DCL [4]	n/a	93.0%
-	-	-
Our DF-GMM	n/a	**93.8%**

end-to-end feature encoding, (4) end-to-end localization-classification sub-networks, (5) other methods (e.g. reinforcement learning [14], knowledge representation [3]) and (6) our DF-GMM.

Earlier multi-stage methods rely on the object and even part annotations to achieve comparable results. However, using the object or part annotations limits the performance due to the fact that human annotations only give the coordinates of important parts rather than the accurate discriminative region location. Weakly supervised multi-stage frameworks gradually exceed the strong supervised methods though picking out discriminative regions. The end-to-end feature encoding methods have good performance via encoding the CNN feature vectors into high-order information, while they result in high computational cost. Although the localization-classification sub-networks works well on various datasets, they neglect the problem of discriminative region diffusion and have difficulty in picking out the accurate discriminative regions. Other methods also achieve comparable performance due to using the extra information (e.g. the semantic embedding).

Table 6. Comparison with the efficiency and effectiveness of other method on CUB-200-2011. K means the number of selected discriminative regions for each image.

Method	Annotation	Accuracy	Speed
M-CNN(K=2) [29]	Parts	84.20%	12.90
WSDL(K=1) [13]	n/a	83.45%	10.07
Bilinear(K=0) [21]	n/a	84.00%	30.00
Our DF-GMM(K=2)	n/a	88.10%	**43.00**
Our DF-GMM(K=4)	n/a	**88.80%**	41.00

Table 7. Effect of Global Max Pooling vs. Global Average Pooling on base initialization, the recognition accuracy on CUB-200-2011.

Initialization Method	Accuracy
Random initialization	87.1%
Global Max Pooling	87.9%
Global Average Pooling	88.8%

Table 8. The recognition accuracy on CUB-200-2011 of model trained with different number of GMM iterations.

k	1	2	3	4	5
Accuracy	86.9%	87.5%	88.8%	88.4%	88.1%

As shown in Table 3, Table 4 and Table 5, our approach outperforms these strong supervised methods in the first group, which indicates that the proposed method can find the discriminative patches without any fine-grained annotations. Compared with recent weakly supervised end-to-end methods, which find discriminative patches from high-level feature maps directly. We run DF-GMM to learn low-rank feature maps to alleviate discriminative region diffusion problem and achieves the new state-of-the-arts.

Speed Comparison. Table 6 shows the speed comparison with other methods. All the experiments are under the setting with batch size 8 using a graphics card of Titan X. While selecting 2 discriminative patches according to the discriminative score maps, we outperform other methods both in speed and accuracy. When we increase the discriminative patches from 2 to 4, the proposed model achieves the state-of-the-art recognition precision, and still stay real-time at 41 fps.

4.5. Visualization Analysis

Insights about the influence of our proposed approach can be obtained by visualizing the effects of feature maps M_I and M_D, i.e. the feature maps without and with DF-GMM respectively. As shown in Figure 3, the feature map response can be shrank to pay attention to the accurate discriminative regions with DF-GMM, which improves the accuracy of localizing discriminative regions. We also visualize the latent variables in GMM, as shown in Figure 4. The linear weight coefficients can be displayed at the area of object that indicates the network focuses on the discriminative

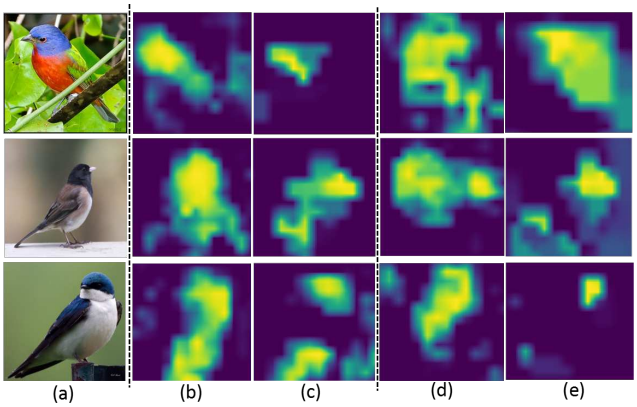

Figure 3. Visualization of intermediate results in DF-GMM. (a) is the original images, (b)(d)indicate the original feature maps M_I and (c)(d) denote the reconstructing feature maps of the special channel, respectively. (b)(c) are the same channel feature map. (d)(e) are also the same channel feature map.

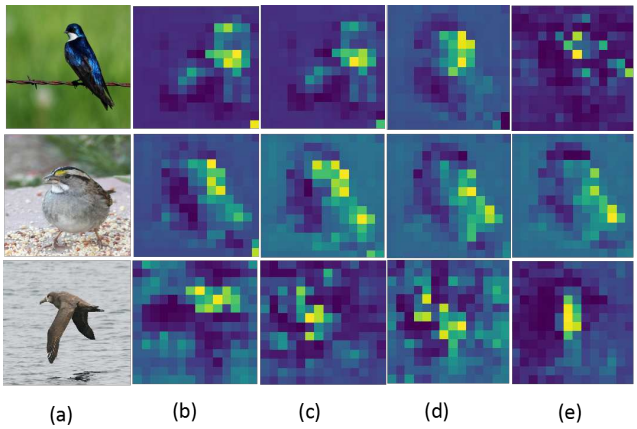

Figure 4. Visualization of the latent variables at the last iteration in GMM. (a) is the original images. (b)(c)(d)(e) indicate the latent variables corresponding certain-th base.

regions. We draw the discriminative regions and display the discriminative response map predicted by our model without and with DF-GMM in Figure 5, respectively. It can be seen that the discriminative response maps without DF-GMM focus on the wide area which results in the problem of hard localization, as shown in Figure 5(b). However, Our DF-GMM could pay attention to a small area in discriminative response maps, where the discriminative patches can be located more easily and accurately. For more intuitive presentation, we display the localization results in original images, as shown in Figure 5(d)(e).

4.6. Discussions

The deeper, the better? We show the recognition results with different iterative number of GMM, as shown in Table 8. It is obvious that the performance of DF-GMM drops when the iterative number increases to 4. The possible reason of the performance drop is that after using more E-step and M-step, the propagation between bases μ and latent variables Z will be overwhelmed.

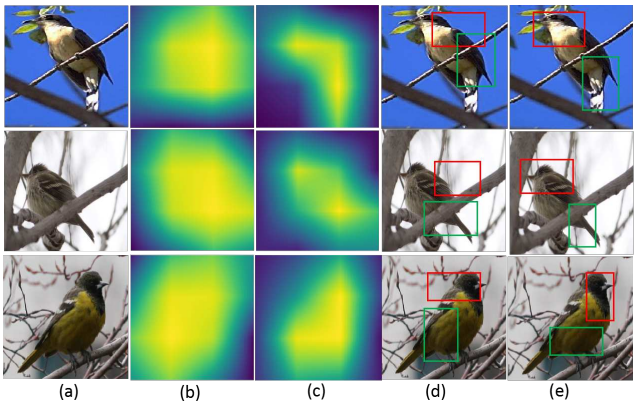

Figure 5. Visualization of discriminative response maps and localization results with and without DF-GMM.(a) is the original images. (b)(c) are the discriminative response maps through sampling stage without and with the DF-GMM, respectively. (d)(e) are the localization results without and with DF-GMM, respectively.

GMP vs. GAP: As it can be seen in Table 5, switching the pooling method from GAP to GMP leads to a significant performance drop. Therefore, although the low-rank bases are initialized to same state, GAP makes discriminative bases focus on all discriminative information by encouraging the GMM to have high response over the whole discriminative regions and the gradients affect every spatial location of discriminative regions during training procedure. On the other side, GMP makes filters pay attention to the most discriminative region to have a single response at a certain location of the feature map and the gradients will only be back-propagated to that location.

5. Conclusion

In this paper, we first discover the discriminative region diffusion problem of high-level feature maps in WFGIR methods. We argue that DRD problem aggravates the difficulty of discriminative region localization for existing methods. We propose an end-to-end Discriminative Feature-oriented Gaussian Mixture Model method to learn low-rank feature maps to address DRD problem. Extensive experiments show that the recognition accuracy can be improved significantly by localizing patches on the new low-rank feature maps, which proves the DRD problem does play a key role in WFGIR. The last but the most important, our algorithm is end-to-end trainable and achieves state-of-the-art in CUB-Bird, FGVC Aircraft and Stanford Cars datasets.

References

[1] Steve Branson, Grant Van Horn, Serge J. Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. *CoRR*, abs/1406.2952, 2014.

[2] Sijia Cai, Wangmeng Zuo, and Lei Zhang. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 511–520, 2017.

[3] Tianshui Chen, Liang Lin, Riquan Chen, Yang Wu, and Xiaonan Luo. Knowledge-embedded representation learning for fine-grained image recognition. In *IJCAI*, pages 627–634, 2018.

[4] Yue Chen, Yalong Bai, Wei Zhang, and Tao Mei. Destruction and construction learning for fine-grained image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5157–5166, 2019.

[5] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin, and Serge J. Belongie. Kernel pooling for convolutional neural networks. In *CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3049–3058, 2017.

[6] Yao Ding, Yanzhao Zhou, Yi Zhu, Qixiang Ye, and Jianbin Jiao. Selective sparse sampling for fine-grained image recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[7] Daniel Fried, Tamara Polajnar, and Stephen Clark. Low-rank tensors for verbs in compositional distributional semantics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 731–736, 2015.

[8] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4476–4484, 2017.

[9] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 317–326, 2016.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1026–1034, 2015.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[12] Xiangteng He and Yuxin Peng. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. In *AAAI February 4-9, 2017, San Francisco, California, USA.*, pages 4075–4081, 2017.

[13] Xiangteng He, Yuxin Peng, and Junjie Zhao. Fine-grained discriminative localization via saliency-guided faster R-CNN. In *ACM MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 627–635, 2017.

[14] Xiangteng He, Yuxin Peng, and Junjie Zhao. Stackdrl: Stacked deep reinforcement learning for fine-grained visual categorization. In *IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 741–747, 2018.

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7132–7141, 2018.

[16] Shu Kong and Charless C. Fowlkes. Low-rank bilinear pooling for fine-grained classification. In *CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 7025–7034, 2017.

[17] Jonathan Krause, Hailin Jin, Jianchao Yang, and Fei-Fei Li. Fine-grained recognition without part annotations. In *CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 5546–5555, 2015.

[18] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pages 554–561, 2013.

[19] Zhichao Li, Yi Yang, Xiao Liu, Feng Zhou, Shilei Wen, and Wei Xu. Dynamic computational time for visual attention. In *ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, pages 1199–1209, 2017.

[20] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 936–944, 2017.

[21] Tsung-Yu Lin, Aruni Roy Chowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1449–1457, 2015.

[22] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013.

[23] Jiaqi Mu, Suma Bhat, and Pramod Viswanath. Representing sentences as low-rank subspaces. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 629–634, 2017.

[24] Yuxin Peng, Xiangteng He, and Junjie Zhao. Object-part attention model for fine-grained image classification. *TIP*, 27(3):1487–1500, 2018.

[25] Yaming Wang, Jonghyun Choi, Vlad I. Morariu, and Larry S. Davis. Mining discriminative triplets of patches for fine-grained classification. In *CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1163–1172, 2016.

[26] Yaming Wang, Vlad I. Morariu, and Larry S. Davis. Learning a discriminative filter bank within a CNN for fine-grained recognition. In *CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4148–4157, 2018.

[27] Zhihui Wang, Shijie Wang, Pengbo Zhang, Haojie Li, Wei Zhong, and Jianjun Li. Weakly supervised fine-grained image classification via correlation-guided discriminative learning. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 1851–1860, 2019.

[28] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. *TIP*, 26(6):2868–2881, 2017.

[29] Xiu-Shen Wei, Chen-Wei Xie, and Jianxin Wu. Mask-cnn: Localizing parts and selecting descriptors for fine-grained image recognition. *CoRR*, abs/1605.06878, 2016.

[30] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *ECCV, Germany, September 8-14, 2018, Proceedings, Part XIV*, pages 438–454, 2018.

[31] Hantao Yao, Shiliang Zhang, Chenggang Yan, Yongdong Zhang, Jintao Li, and Qi Tian. Autobd: Automated bi-level description for scalable fine-grained visual categorization. *TIP*, 27(1):10–23, 2018.

[32] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. Picking deep filter responses for fine-grained image recognition. In *CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1134–1142, 2016.

[33] Yu Zhang, Xiu-Shen Wei, Jianxin Wu, Jianfei Cai, Jiangbo Lu, Viet Anh Nguyen, and Minh N. Do. Weakly supervised fine-grained categorization with part-based image representation. *TIP*, 25(4):1713–1725, 2016.

[34] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5219–5227, 2017.

[35] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5012–5021, 2019.

[36] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, Jiebo Luo, and Tao Mei. Learning rich part hierarchies with progressive attention networks for fine-grained image recognition. *IEEE Trans. Image Processing*, 29:476–488, 2020.