Dual Attention Networks for Few-Shot Fine-Grained Recognition

Shu-Lin Xu^{1,2}, Faen Zhang^{3}, Xiu-Shen Wei^{1,2,4*}, Jianhua Wang^{3}
[1]School of Computer Science and Engineering, Nanjing University of Science and Technology
[2]State Key Laboratory of Integrated Services Networks, Xidian University
[3]AInnovation Technology Group Co., Ltd
[4]State Key Laboratory for Novel Software Technology, Nanjing University
{xusl, weixs}@njust.edu.cn, {zhangfaen, wangjianhua}@ainnovation.com

Abstract

The task of few-shot fine-grained recognition is to classify images belonging to subordinate categories merely depending on few examples. Due to the fine-grained nature, it is desirable to capture subtle but discriminative part-level patterns from limited training data, which makes it a challenging problem. In this paper, to generate fine-grained tailored representations for few-shot recognition, we propose a Dual Attention Network (DUAL ATT-NET) consisting of two dual branches of both hard- and soft-attentions. Specifically, by producing attention guidance from deep activations of input images, our hard-attention is realized by keeping a few useful deep descriptors and forming them as a bag of multi-instance learning. Since these deep descriptors could correspond to objects' parts, the advantage of modeling as a multi-instance bag is able to exploit inherent correlation of these fine-grained parts. On the other side, a soft attended activation representation can be obtained by applying attention guidance upon original activations, which brings comprehensive attention information as the counterpart of hard-attention. After that, both outputs of dual branches are aggregated as a holistic image embedding w.r.t. input images. By performing meta-learning, we can learn a powerful image embedding in such a metric space to generalize to novel classes. Experiments on three popular fine-grained benchmark datasets show that our DUAL ATT-NET obviously outperforms other existing state-of-the-art methods.

Introduction

Fine-grained visual recognition is a longstanding and fundamental problem in computer vision (Wei et al. 2021), which aims to distinguish each subordinate class that belong to a specific basic class (*e.g.*, birds, dogs, or cars). It is a challenging task due to the small inter-class variance and the large intra-class variance. With the development of deep learning, many works have achieved good results. However, these results always rely on a large number of labeled samples, and even stronger supervision information such as bounding boxes and part annotations. In contrast, the human visual system can quickly learn novel concepts and features from a few examples, and then recognize similar objects in new coming images. To mimic this human ability, in this work, our goal is to learn to classify novel fine-grained categories from a few examples with only category labels, which is a practical but more challenging task.

In general, the generic few-shot learning methods do not achieve satisfactory results in *fine-grained* image recognition because it might not distinguish the subtle differences between samples well. Existing methods (Huang et al. 2020, 2021; Zhu, Liu, and Jiang 2020; Li et al. 2020) for few-shot fine-grained recognition attempt to train a strong end-to-end network to obtain good feature representations, such as obtaining second-order feature interactions (Huang et al. 2020, 2021), ensembling two metrics (Li et al. 2020), and so on. Different from them, our model is designed to learn a powerful feature embedding in a metric space by performing the proposed dual attention approach in a meta learning fashion.

In this paper, we propose DUAL ATT-NET consisting both hard- and soft-attention to capture comprehensive attention information to further generate fine-grained tailored image embedding for few-shot recognition. Specifically, after extracting deep activations from a CNN backbone, the attention guidance can be generated, which is then employed upon the proposed dual attention processes. For the hard-attention, the attention guidance equipped with a threshold will keep a few useful deep descriptors corresponding to fine-grained objects' parts from these activations and also discard noisy descriptors corresponding to noises. We form these remained deep descriptors as a multi-instance bag from the multi-instance learning (MIL) perspective. The motivation of such a procedure is that, MIL can better exploit inherent structure information of instances (*i.e.*, descriptors w.r.t. fine-grained parts) than traditional single-instance learning. It is also noted that, regarding fine-grained objects, their discriminative parts, especially the correlation of these parts, are crucial for distinguishing sub-ordinate categories in fine-grained recognition. In details, we develop a graph-based MIL bag aggregation strategy to model the relation of fine-grained parts. After that, the outputs of hard-attention explicitly enrich part-level cues, as well as more important part correlation. On the other side, regarding the soft-attention, it performs as the counterpart of hard-attention by conducting attention guidance on the ob-

[*] Corresponding author. S.-L. Xu and X.-S. Wei are also with PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, Nanjing University of Science and Technology.

tained deep activations. Then, to abstract global aggregation information, global max- and average-pooling are utilized on the attended representations. Finally, both universal outputs of our dual attention streams are gathered and concatenated as the required image embedding. Upon a framework of meta-learning the embedding, we can learn fine-grained tailored image embedding in a metric space.

In experiments, we perform DUAL ATT-NET on three fine-grained benchmark datasets, *i.e.*, CUB Birds (Wah et al. 2011),Stanford Dogs (Khosla et al. 2011),Stanford Cars (Krause et al. 2013). Empirical results show that our DUAL ATT-NET significantly outperforms competing baseline methods, including state-of-the-art few-shot fine-grained recognition methods (Zhu, Liu, and Jiang 2020; Huang et al. 2021; Li et al. 2020; Huang et al. 2020), and generic few-shot learning methods (Li et al. 2019; Sung et al. 2018; Shell, Swersky, and Zemel 2017; Vinyals et al. 2016).

In summary, our major contributions are three-fold:

- We propose a novel few-shot fine-grained framework consisting dual attention streams for obtaining fine-grained tailored image embedding in a meta learning fashion.
- We develop a dual attention network, including hard- and soft-attention, for not only explicitly modeling the crucial relation of fine-grained objects' parts, but also implicitly capturing discriminative while subtle fine-grained details.
- We conduct comprehensive experiments on three popular fine-grained benchmark datasets, and our proposed model achieves superior recognition accuracy over competing solutions on these datasets.

Related Work

Fine-Grained Visual Recognition

Fine-Grained Visual Recognition (FGVR) is a challenging problem and has been an active research topic emerged in recent years. FGVR aims to distinguish many similar subordinate categories that belong to the same basic category, such as the fine distinction of animal species (Horn et al. 2018), plant species (Hou, Feng, and Wang 2017), cars (Krause et al. 2013), aircraft (Maji et al. 2013) and so on. Due to the challenge of distinguishing between regional localization and fine-grained feature learning, it is difficult to recognize fine-grained categories.

Thanks to the development of powerful deep learning techniques with large annotated datasets, a number of efficient fine-grained recognition methods have achieved high performance. Among them, some work, *e.g.*, (Wei et al. 2018; Wang et al. 2020; Liu et al. 2020), captured the discriminative semantic parts of fine-grained objects by localization-classification subnetworks, and then constructed a mid-level representation corresponding to these parts for the final classification to bring accuracy improvements. Some methods attempted to model subtle differences between fine-grained categories by stronger end-to-end feature coding, such as performing high-order feature interactions (Lin, RoyChowdhury, and Maji 2017; Kong and Fowlkes 2017) or designing novel loss function (Dubey et al. 2018). Moreover, some of them leverage external information to further assist fine-grained recognition, such as web data (Sun, Chen, and Yang 2019), multi-model data (Reed et al. 2016), or human-computer interactions (Deng et al. 2015).

On the basis of FGVR, we are studying FGVR in a more challenging few-shot learning setting, where the model is required to use only a few labeled images to identify novel fine-grained categories. To deal with this problem, we focus on capturing different fine-grained semantic parts to improve the model's representation ability. In our work, we select useful deep descriptors corresponding to fine-grained object parts based on attention guidance, and then form a multi-instance learning bag to model the relationship between these fine-grained parts.

Generic Few-Shot Learning

In recent years, for eliminating the dependence of deep learning models on large amounts of data, Few-Shot Learning (FSL) has attracted more and more attention. The successful progress of FSL enables the learning system to quickly learn novel patterns from a few examples with supervised information, even by incorporating prior knowledge.

The literature on FSL has been diverse, we mainly focus on the meta-learning framework. There are two main streams in the FSL methods, including metric-based approaches (Vinyals et al. 2016; Finn, Abbeel, and Levine 2017; Shell, Swersky, and Zemel 2017; Finn, Xu, and Levine 2018; Oreshkin, Rodriguez, and Lacoste 2018; Sung et al. 2018; Yoon, Seo, and Moon 2019; Li et al. 2019; Guo and Cheung 2020; Yang et al. 2020) and optimization-based approaches (Luketina et al. 2016; Lee and Choi 2018; Antoniou, Edwards, and Storkey 2019; Jamal and Qi 2019; Lifchitz et al. 2019). The metric-based methods usually embed support samples and query samples into the same feature space first, and then calculate the similarity of the embedded features to make predictions. The optimization-based approaches firstly train a network with auxiliary data, and then fine-tune the classifier or the whole network with support data from unseen novel classes. In addition, many other methods have been proposed to deal with the FSL problem, such as work based on graph theory (Kim et al. 2019), reinforcement learning (Chu et al. 2019), differentiable SVM (Lee et al. 2019), etc.

The most related work of ours is Prototypical Networks (Shell, Swersky, and Zemel 2017), which first uses the mean value of the sample feature of each class to construct the prototype representation, and then calculates the similarity between the query sample representation and the prototype representation for inference. However, it simply concatenates/merges all descriptors of a sample into a single vector. This may cause misalignment of parts for fine-grained classification. Hereby, we propose DUAL ATT-NET to better mine the information of deep descriptors. Our proposed DUAL ATT-NET can select the key deep descriptors corresponding to the fine-grained object parts and then use a multi-instance learning based approach to generate a key part representation. Moreover, we also use global pooling for all deep descriptors to generate a global representation that do not need to consider the positional relationship to supplement the samples information. Aggregating the above two representations, we can map the samples to a better embedding space, and then calculate the similarity between samples.

Multi-Instance Learning

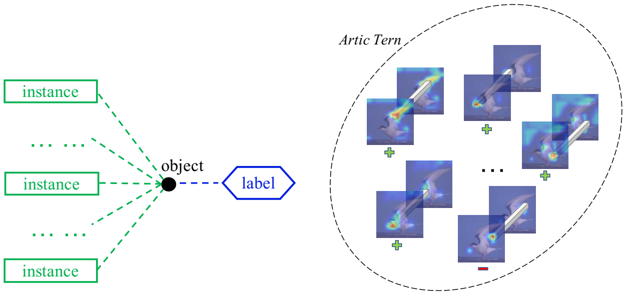
Multi-Instance Learning (MIL) is a type of weakly supervised task in machine learning, in which a labeled bag is associated with multiple unlabeled instances or descriptions. With the pioneering proposal of MIL in (Dietterich, Lathrop, and Lozano-Pérez 1997), many MIL algorithms have been developed since it helps to solve a range of real applications. In recent years, with the development of deep learning, many MIL methods on neural networks have been proposed (Ilse, Tomczak, and Welling 2018; Wang et al. 2019; Tu et al. 2019), on the basis of classic machine learning (Andrews, Tsochantaridis, and Hofmann 2003; Zhou, Sun, and Li 2009; Zhou et al. 2012). In particular, the MIL methods for image recognition (Tang et al. 2018; Carbonneau et al. 2018; Angles et al. 2021) has also achieved some good results. In initial MIL research, the standard MI assumption was made. It is normally referred that each instance has an unknown class label which identifies it as either positive or negative. A bag is considered to be positive if and only if it contains at least one positive instance. But this assumption is not available in all MIL problems. In some cases, a generalized assumption is required: collective assumption (Xu 2003) is often used, where class label of a bag as a property that is related to all the instances within that bag. Under the latter assumption, there are two approach to get the class label of a bag. One is the instance-level approach, where there is an instance-level transfer function to get the score of each instance, and then the class label of the bag is obtained by a MIL pooling, *e.g.*, max pooling and mean pooling. The other is the bag-level approach, which transforms the multi-instance data into a single-bag representation then train a bag-level classifier on the transformed data. In most cases, the second approach is more flexible and more competitive.

In few-shot fine-grained scenarios, key parts of an image can often play a decisive role in image classification, so we try to treat a patch of an image as an instance, the original image as a bag, and classify the class label of a bag through MIL. In our work, we are inspired by work (Zhou, Sun, and Li 2009) and use the collective assumption with the bag-level approach to solve few-shot fine-grained recognition.

Recap of Multi-Instance Learning

Multi-instance learning (Dietterich, Lathrop, and Lozano-Pérez 1997) was originally proposed to investigate the problem of drug activity prediction, which later has been widely applied on diverse applications involving complicated data objects, such as images (Wang, Ruan, and Si 2014; Wei and Zhou 2016; Yuan et al. 2021). Contrasting to traditional single-instance learning, the multi-instance representation enables the learning process to exploit some inherent structure information in input data.

Specifically, multi-instance learning (MIL) receives a set of bags X_i associated with their labels y_i, *i.e.*, $\{(X_1, y_1), \ldots, (X_i, y_i), \ldots, (X_{N_B}, y_{N_B})\}$, where the bag $X_i = \{\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{ij}, \ldots, \boldsymbol{x}_{i,n_i}\}$ consists of multiple instances. In concretely, $\boldsymbol{x}_{ij} = [x_{ij1}, \ldots, x_{ijl}, \ldots, x_{ijd}]^\top \in \mathcal{X}$ represents an instance, where x_{ijl} is the value of $\boldsymbol{x}_{ij}$ at the l-th attribute. N_B is the number of training bags, n_i is the number of instances in the corresponding bag X_i, and d is the number of attributes. The goal of MIL is to generate a classifier to classify unseen bags as one of the C categories. In this paper, as illustrated in Fig. 1, we hereby regard several key deep descriptors gathered by our hard-attention as a MIL bag, which is expected to capture the correlation of these descriptors and use this to further model the inherent structure of fine-grained objects' parts.



(a) Multi-instance learning (b) Fine-grained representations from the MIL perspective

Figure 1: Motivation of handling fine-grained representations from the multi-instance learning (MIL) perspective. Regarding an input image (*i.e.*, an MIL bag) of a fine-grained class, it contains several descriptors corresponding key objects' parts or perhaps noises. These descriptors are treated as multiple positive/negative instances in MIL. (Best viewed in color.)

Methodology

Learning Strategy, Framework and Notations

Our work is built upon the framework of meta-learning which aims to meta-learn a proper and powerful feature embedding function $f_{\text{embed}}(\cdot)$ to perform few-shot fine-grained recognition. In concretely, given an auxiliary training set $\mathcal{B}$, it contains N labeled training images $\mathcal{B} = \{(I_1, y_1), (I_2, y_2), \ldots, (I_N, y_N)\}$, where I_i is an example image and $y_i \in \{1, 2, \ldots, C_{\mathcal{B}}\}$ is its corresponding label. Once the embedding is learned, it is then applied on another testing set $\mathcal{N}$ for performance evaluation, where $\mathcal{N}$ contains images of novel categories that do not appear in $\mathcal{B}$.

To meta-learn the embedding, we randomly sample a set of "meta-training sets" from $\mathcal{B}$. Each meta-training set (corresponding to a training episode) contains $C_{\mathcal{S}} < C_{\mathcal{B}}$ randomly chosen categories and a few images associated with them. A meta-training set is composed of a "support set" $\mathcal{S}$ and a "query set" $\mathcal{Q}$ to mimic the scenario in testing. Specifically, $\mathcal{S}$ contains N_s (*e.g.*, 1 or 5) supported images per category. The query set $\mathcal{Q}$ is coupled with $\mathcal{S}$ (with the same categories), but has no overlapped images. Each category of $\mathcal{Q}$ contains N_q query images.

As the framework of Dual Att-Net illustrated in Fig. 2, given an input image I_i and a CNN model $f_{\text{cnn}}(\cdot)$, we can obtain the activations of a convolution layer as an order-3 tensor $\boldsymbol{T}_i$ with $H \times W \times D$ elements. $\boldsymbol{T}_i$ is typically considered as having $H \times W$ cells and each cell contains one

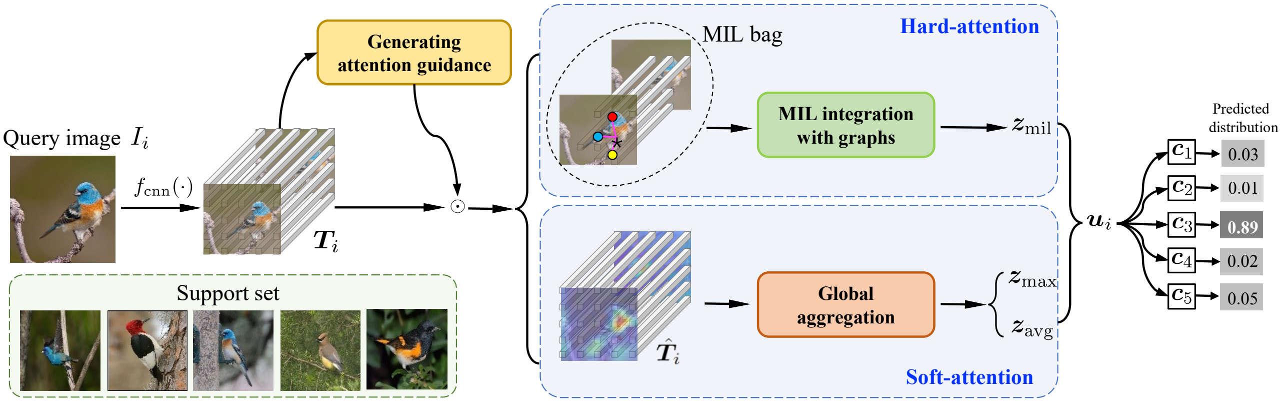


Figure 2: Overall framework of the proposed DUAL ATT-NET, which consists of several crucial components, *i.e.*, 1) generating attention guidance derived from the extracted deep activation tensor $\boldsymbol{T}_i$; 2) selecting useful deep descriptors corresponding to fine-grained objects' parts as *hard-attention* and forming a multi-instance learning bag for modeling the relationship of fine-grained parts; 3) performing *soft-attention* with attention guidance and obtaining global aggregated features. Finally, once the holistic image embedding $\boldsymbol{u}_i$ is obtained, the whole network is end-to-end trainable for few-shot recognition in the meta-learning fashion. Note that, "MIL" is the abbreviation of "multi-instance learning".

D-dimensional deep descriptor (Wei et al. 2017), *i.e.*, $\boldsymbol{x}_{h,w}$ in the (h,w)-th cell. Then, attention guidance can be generated from $\boldsymbol{T}_i$. Since these deep descriptors can represent their corresponding local image patches in raw pixels, we apply attention guidance to perform both hard- and soft-attention upon $\boldsymbol{T}_i$ for capturing fine-grained patterns.

Specifically, the hard-attention of DUAL ATT-NET attempts to remain key deep descriptors w.r.t. discriminative parts of fine-grained objects (*e.g.*, heads or tails) and mask out other useless descriptors. Then, these selected deep descriptors of an image form a multi-instance learning (MIL) bag X_i, and one deep descriptor correspond an instance of a bag. These selected key descriptors can be analogous to key instances in MIL (Zhang and Zhou 2017). After that, in order to model the relationship between these key descriptors (aka key fine-grained parts), we propose to integrate these instances of a MIL bag with graph based methods and return $\boldsymbol{z}_{\text{mil}}$. On the other side, regarding soft-attention, the learnt attention guidance is dot-product with $\boldsymbol{T}_i$ to get an attended tensor representation $\hat{\boldsymbol{T}}_i$. Both global max- and average-pooling are then applied on $\hat{\boldsymbol{T}}_i$ for global aggregation, *i.e.*, $\boldsymbol{z}_{\text{max}}$ and $\boldsymbol{z}_{\text{avg}}$. Finally, the outputs of dual attention are concatenated into a holistic image representation $\boldsymbol{u}_i$ which will be employed in subsequent processes. Briefly, the aforementioned procedure can be highly abstracted as a learnable fine-grained tailored feature embedding:

$$\boldsymbol{u}_i \leftarrow f_{\text{embed}}(I_i;\theta_{\text{embed}})\,, \tag{1}$$

where θ_{embed} is the parameter of the embedding function.

For network training, based on Eq. (1) and following (Shell, Swersky, and Zemel 2017), for a given query sample $\boldsymbol{u}' \in \mathcal{Q}$ with label $y' = k$, we calculate distance between $\boldsymbol{u}'$ and the categorical prototypes from $\mathcal{S}$. The class predicted distribution can be obtained via softmax by

$$p_{\theta_{\text{embed}}}(y'=k|\boldsymbol{u}') = \frac{\exp(-d(\boldsymbol{u}',\boldsymbol{c}_k))}{\sum_{k'}\exp(-d(\boldsymbol{u}',\boldsymbol{c}_{k'}))}\,, \tag{2}$$

where $d(\cdot)$ is the distance function, and $\boldsymbol{c}_k = \frac{1}{|\Omega_k|}\sum_{i\in\Omega_k}\boldsymbol{u}_i$ represents the categorical prototype of class k, particularly $\Omega_k = \{i|y_i = k\}$. Overall, the model parameter θ_{embed} is trained via minimizing the negative log-likelihood $\mathcal{J}(\boldsymbol{u}',y') = -\log(p_{\theta_{\text{embed}}}(k|\boldsymbol{u}'))$ over $\mathcal{Q}$.

In the following, we elaborate the crucial components of DUAL ATT-NET, *i.e.*, attention guidance generation, as well as both hard- and soft-attention mechanisms.

Attention Guidance Generation

As aforementioned, based on a CNN model, the activation tensor $\boldsymbol{T}_i$ is obtained via

$$\boldsymbol{T}_i = f_{\text{cnn}}(I_i) \in \mathbb{R}^{H\times W\times D}\,. \tag{3}$$

Then, we perform both global max- and average-pooling on $\boldsymbol{T}_i$ to get the aggregated signal as two $H \times W \times 1$ tensors. These two tensors are concatenated as $H \times W \times 2$, and conducted by a convolution operation (*e.g.*, $2\times 7\times 7\times 1$ with the stride of 1 and padding of 3). Thus, a matrix $\boldsymbol{G}_i \in \mathbb{R}^{H\times W}$ is returned, and the sigmoid function is used upon $\boldsymbol{G}_i$ to normalize its values into an interval of $[0,1]$. We denote the normalized matrix with $\hat{\boldsymbol{G}}_i$ as attention guidance for evaluating the importance of deep descriptors (Wei et al. 2017) in these $H \times W$ cells.

Based on $\hat{\boldsymbol{G}}_i$, it is designed to produce two dual streams of outputs, *i.e.*, hard-attention and soft-attention. The hard-attention is to directly mask out less important deep descriptors and remain key descriptors by referring to a threshold δ. The soft-attention is to conduct dot-product upon $\boldsymbol{T}_i$ to generate an attended tensor representation $\hat{\boldsymbol{T}}_i$. The dual attention framework is as illustrated in Fig. 2.

Hard-Attention in DUAL ATT-NET

Regarding hard-attention, we employ δ as a threshold on $\hat{\boldsymbol{G}}_i$ to filter key descriptors from $\boldsymbol{T}_i$, and form these remained descriptors as a MIL bag:

$$X_i = \{\boldsymbol{x}_{h,w} | \hat{\boldsymbol{G}}_i(h, w) > \delta\} \,. \tag{4}$$

Then, we remark the index of descriptors in X_i from 1 to n_i as $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_i}\}$ for facilitating the subsequent statement.

The motivation of forming these selected descriptors into a MIL bag is that, MIL can better exploit inherent structure information than traditional single-instance learning. While, for fine-grained objects, their discriminative parts, especially the correlation of these parts, are crucial for distinguishing sub-ordinate categories in fine-grained recognition. Therefore, after key descriptor selection by our hard-attention, the instance in X_i, which is indeed a descriptor, has potential to correspond to a specific fine-grained part.[1]

Here, we propose to use a graph-based method to both model the relationship of instances (*i.e.*, fine-grained parts) and integrate these instances into a universal feature vector. More specifically, we first transform MIL bags into graphs, where each bag generates a single graph $\mathcal{G}$ and the nodes $\mathcal{V}$ of graphs are the instances in the MIL bag X_i. We denote $\boldsymbol{A} \in \mathbb{R}^{n_i \times n_i}$ as the correlation matrix, which guides the information propagation between these graph nodes and further updates the node representations. For the elements of $\boldsymbol{A} \in \mathbb{R}^{n_i \times n_i}$, it is set to 1 or 0 according to the Euclidean distance between two nodes in $\mathcal{G}$:

$$A_{u,v} = \begin{cases} 1 & \text{if} \quad dist(\boldsymbol{x}_u, \boldsymbol{x}_v) \in \text{top}k\{\mathcal{D}\} \\ 0 & \text{otherwise} \end{cases} \,, \tag{5}$$

where a set of $\text{top}k\{\mathcal{D}\}$ is the top-k smallest distance among all the node pairs. Then, we obtain the graphs as $\mathcal{G} =< \mathcal{V}, \boldsymbol{A} >$ to represent X_i.

For modeling the instance correlation, we realize it as the graph convolutional operation (Kipf and Welling 2017) by

$$\boldsymbol{H}^{L+1} = h(\boldsymbol{A}\boldsymbol{H}^L\boldsymbol{W}^L) \,, \tag{6}$$

where $\boldsymbol{H}^L \in \mathbb{R}^{n_i \times d}$ represents feature descriptions of the nodes $\mathcal{V}$, L denotes the L-th operation layer, $\boldsymbol{W}^L \in \mathbb{R}^{d \times d'}$ is a transformation matrix to be learned, and $h(\cdot)$ denotes a non-linear activation function, *e.g.*, ELU (Clevert, Unterthiner, and Hochreiter 2016). Since the information can be propagated through the graph nodes, we can learn and model the inherent structure of the nodes (*i.e.*, instances in X_i) by stacking multiple graph convolution layers. Finally, by learning such a graph-based function $f_{\text{graph}}(\cdot)$ upon $\mathcal{G}$, we can obtain

$$\boldsymbol{H} = f_{\text{graph}}(\boldsymbol{A}, \mathcal{V}) \,, \tag{7}$$

where $\boldsymbol{H} \in \mathbb{R}^{n_i \times D'}$ is the output w.r.t. the original input $\{\boldsymbol{x}\}$ in X_i.

For integrating these instances, in order to transform variant numbers of nodes of $\mathcal{G}$ to a fixed-length feature vector, it is desirable to learn an assignment matrix $\boldsymbol{M} \in \mathbb{R}^{n_i \times t}$ to assign the nodes of $\boldsymbol{H}$ towards t semantical clusters. The so-called semantical cluster hereby can correspond to the semantics of fine-grained parts, cf. Fig. 4. Besides, all graphs share the same value of t to ensure that the integrated outputs have the same dimensions. In fact, $\boldsymbol{M}$ represents the probability of a node belongs to a cluster, which can be learnt by a differentiable pooling algorithm (Ying et al. 2018). Based on the assignment matrix $\boldsymbol{M}$, $\boldsymbol{H}$ is transformed by

$$\boldsymbol{H}^* = \boldsymbol{M}^\top \boldsymbol{H} \,. \tag{8}$$

We concatenate nodes of $\boldsymbol{H}^*$ into a single vector $\boldsymbol{z}_{\text{mil}} \in \mathbb{R}^{(D' \times t) \times 1}$ as the final representation of hard-attention w.r.t. a MIL bag X_i.

Soft-Attention in DUAL ATT-NET

Beyond the hard-attention of DUAL ATT-NET, soft-attention is also utilized as the counterpart of hard-attention, which is desirable to obtain more comprehensive attention information from $\boldsymbol{T}_i$. By formulation, the soft-attention is simple to implement by a dot-product with the attention guidance:

$$\hat{\boldsymbol{T}}_i = \boldsymbol{T}_i \odot \hat{\boldsymbol{G}}_i \,, \tag{9}$$

where $\hat{\boldsymbol{T}}_i$ is the obtained tensor representation with soft attentions. To abstract global aggregation information, global max- and average-pooling are conducted on $\hat{\boldsymbol{T}}_i$, and thus return $\boldsymbol{z}_{\max}$ and $\boldsymbol{z}_{\text{avg}}$.

Finally, both universal outputs of dual attention streams, *i.e.*, $\boldsymbol{z}_{\text{mil}}$, $\boldsymbol{z}_{\max}$ and $\boldsymbol{z}_{\text{avg}}$, are gathered. They are hereby concatenated into a holistic image representation $\boldsymbol{u}_i = [\boldsymbol{z}_{\text{mil}}; \boldsymbol{z}_{\max}; \boldsymbol{z}_{\text{avg}}]$. Such a representation $\boldsymbol{u}_i$ not only enriches inherent structural correlations of fine-grained parts, but also contains fine-grained tailored subtle and discriminative attention clues. We then depend on $\boldsymbol{u}_i$ to perform few-shot fine-grained recognition.

Experiments

Datasets, Setups and Implementation Details

We conduct the experiments on three popular used benchmark datasets for few-shot fine-grained recognition, *i.e.*, CUB Birds (200 categories of birds, 11,788 images), Stanford Dogs (120 categories of Dogs, 20,580 images), Stanford Cars (196 categories of cars, 16,185 images). For each dataset, we follow (Wei et al. 2019; Huang et al. 2019, 2020) to randomly split its original image categories into two disjoint subsets: One as the auxiliary training set $\mathcal{B}$, and the other as the FSFG testing set $\mathcal{N}$, which is shown in Table 1.

Category	CUB	Cars	Dogs
C_{total}	200	196	120
$C_{\mathcal{B}}$	150	147	90
$C_{\mathcal{N}}$	50	49	30

Table 1: Category split for three datasets. C_{total} represents the total number of categories in the dataset, $C_{\mathcal{B}}$ represents the number of categories in $\mathcal{B}$, and $C_{\mathcal{N}}$ represents the number of categories in $\mathcal{N}$.

[1] In experiments, it can be validated that these selected descriptors indeed correspond to fine-grained parts, cf. Fig. 4.

Methods	Type	Published in	CUB		Cars		Dogs	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MatchingNet	FS	NeurIPS’16	57.59±0.74	70.57±0.62	48.03±0.60	64.22±0.59	45.05±0.66	60.60±0.62
ProtoNet	FS	NeurIPS’17	53.88±0.72	70.85±0.63	45.27±0.61	64.24±0.61	42.58±0.63	59.49±0.65
RelationNet	FS	CVPR’18	59.82±0.77	71.83±0.61	56.02±0.74	66.93±0.63	44.75±0.70	58.36±0.66
DN4	FS	CVPR’19	53.15±0.84	81.90±0.60	61.51±0.85	**89.60±0.44**	45.73±0.76	66.33±0.66
LRPABN	FSFG	IEEE TMM’19	67.97±0.44	78.26±0.22	63.11±0.46	74.66±0.22	54.82±0.46	67.12±0.23
BSNet	FSFG	IEEE TIP’20	65.20±0.92	84.18±0.64	61.41±0.92	86.68±0.54	51.06±0.94	71.90±0.68
MattML	FSFG	IJCAI’20	66.29±0.56	80.34±0.30	66.11±0.54	82.80±0.28	54.84±0.53	71.34±0.38
TOAN	FSFG	IEEE TCSVT’21	65.34±0.75	80.43±0.60	65.90±0.72	84.24±0.48	49.30±0.77	67.16±0.49
Ours	FSFG	This paper	**72.89±0.50**	**86.60±0.31**	**70.21±0.50**	85.55±0.31	**59.81±0.50**	**77.19±0.35**

Table 2: Comparison results (mean±std) on three fine-grained datasets. The highest average accuracy of each column is marked in bold. Note that, “FS” is the abbreviation of few-shot, and “FSFG” is the abbreviation of few-shot fine-grained.

To mimic testing scenarios, all meta-training sets and testing sets contain $C_{\mathcal{S}} = 5$. Furthermore we follow (Zhu, Liu, and Jiang 2020) to set $N_s = 1$ ($N_s = 5$) for 1-shot recognition (5-shot recognition) and N_q is set to 15 in all settings. The results are reported with mean accuracy (MA) with 0.95 confidence intervals (CIs) over sampled 2000 testing sets.

For a fair comparison with state-of-the-art methods, our CNN model $f_{\text{cnn}}(\cdot)$ is CNN-4 (Shell, Swersky, and Zemel 2017; Zhu, Liu, and Jiang 2020), which is composed of four convolutional blocks and each block comprises a 64-filter 3×3 convolution, a batch normalization layer and a ReLU nonlinearity. The first three blocks contain a 2×2 max-pooling layer. The input of this network is 84×84, and obtain $\boldsymbol{T}_i$ with $10 \times 10 \times 64$ elements. For hyperparameters, we set $\delta = 0.6$ in Eq. (4), and $k = \sum_{u=1}^{n_i} \sum_{v=1}^{n_i} A_{u,v} = \lceil \frac{n_i^2}{t} \rceil$ in Eq. (5). During meta-training, all of models are trained from scratch in an end-to-end manner. We use the Adam optimizer with initial learning rate of 0.001. The total number of episode is 200,000 and the learning rate is of reduced as 1/2 after each 50,000 episodes. We apply data augmentation, which includes random crops, random horizontal flips, and color jitter at the meta-training stage, as well as center crops at the testing stage, in all implemented experiments.

Main Results

Comparison methods In our experiments, we compare our proposed model to the following eight competitive baselines. As our method belongs to the metric-learning branch, we mainly compare our model with seven state-of-the-art metric-learning based models, including Matching Nets (Vinyals et al. 2016), Prototypical Nets (Shell, Swersky, and Zemel 2017), Relation Net (Sung et al. 2018), DN4 (Li et al. 2019), LRPABN (Huang et al. 2020), BSNet (Li et al. 2020), TOAN (Huang et al. 2021). An optimization-based based model called MattML (Zhu, Liu, and Jiang 2020) is also picked for reference. Among them, the first four are generic few-shot methods, and the others are few-shot fine-grained methods.

Comparison results Table 2 presents the average accuracy rates of FSFG in the test stage on the three fine-grained datasets. For each dataset, we report both 1-shot and 5-shot recognition results. As shown in that table, our proposed model is significantly better than other baseline methods almost all the cases of these three datasets. Especially for 1-shot setting on CUB, Cars, and Dogs, we achieve 4.92%, 4.1% and 4.97% improvements than state-of-the-art methods. In addition, in the 5-shot setting on CUB and Dogs, the accuracy of our Dual Att-Net are 2.42% and 5.29% higher than the best performing methods.

By comparing with other baseline methods, we generally observe that the FSFG methods perform better than the FS methods on the three fine-grained datasets, and our proposed method can better recognize novel fine-grained categories. Moreover, it can also be seen from the results that, compared with 5-shot, our accuracy has a more stable improvement under 1-shot on the three datasets. By comparing with DN4, for each descriptor of a query image, DN4 selects the first k-nearest local descriptors in the whole support set as the prototype to calculate the cosine distance. Therefore, when the intra-class variance of the dataset is small, such as a rigid car, DN4 can achieve a good result under the 5-shot setting.

Ablation Studies

To further inspect our Dual Att-Net for FSFG, we conduct ablation experiments on three aspects for 1-shot 5-way recognition. First, we change the number of semantical clusters $t \in \{1, 2, 3, 4, 5\}$ to show its influence on the hard-attention performance. Second, we investigate the influence of the bag representations on FSFG performance via changing the MIL methods. Finally, we set the correlation matrix $\boldsymbol{A}$ in Eq. (5) as the identity matrix to evidence that we need consider the relationship between instances/fine-grained parts.

What is the optimal number of semantical clusters? As aforementioned, we assign all nodes of $\boldsymbol{H}$ towards t semantical cluster to transform variant numbers of nodes of $\mathcal{G}$ to a fixed-length feature vector. Because the number of t affects the length of $\boldsymbol{z}_{\text{mil}}$ and the effect of clustering, we just modify the value of t while keeping the other settings the same on the three fine-grained datasets to explore its impact. Fig. 3 shows the results of our model on three datasets with different number of clusters. We can see from the figure that our model is robust with different numbers of clusters (except for 1), which shows the number of clusters is not sensitive. In

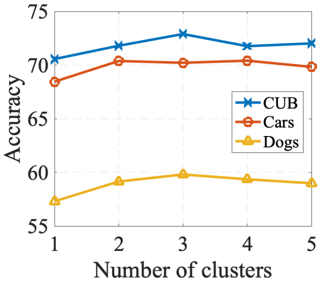


Figure 3: Comparison results in 1-shot learning with different numbers of clusters in our method on the three datasets.

Configurations			CUB	Cars	Dogs
Our MIL aggregation	Naive MIL	Global feature			
✓			58.40	54.30	50.76
	✓		57.72	54.09	47.57
	✓	✓	71.96	69.59	58.95
✓		✓	**72.89**	**70.41**	**59.81**

Table 3: Comparisons of bag representations. Here, "Our MIL aggregation" is our proposed MIL aggregation method, and "Naive MIL" is using simple global average pooling for all instances in the bag. "Global feature" denotes the concatenation of $\boldsymbol{z}_{\max}$ and $\boldsymbol{z}_{\mathrm{avg}}$ in our soft-attention.

particular, the accuracy of our DUAL ATT-NET is the highest when the number of clusters is 3 on the three datasets.

What kinds of bag representations? In order to prove that the proposed MIL-based aggregation approach in our model is effective, we compare it with other MIL pooling methods in Table 3. In addition to only using MIL, we also compare the results of the concatenation between these two MIL aggregation and the global representation. From the results in Table 3, we can see that our MIL-based aggregation is better than that using naive MIL pooling regardless of whether the global representation is concatenated or not.

Necessary to consider relationship between fine-grained parts? As aforementioned in Eq. (6), we use a graph-based method to model the relationship of instances/fine-grained parts. In the process of constructing the graph, we construct a correlation matrix based on the Euclidean distance between nodes, and treat the closer nodes as inter-correlated components. In order to explore whether it is necessary to consider relationship between fine-grained parts, we conduct experiments on setting the correlation matrix $\boldsymbol{A}$ to an identity matrix $\boldsymbol{I}$. Thus, the instances/fine-grained parts in the MIL bag will be treated as obeying an independently identically distribution. Comparison results on the three datasets are reported in Table 4. It is apparent to observe that, by indeed considering the relationship between fine-grained parts, the accuracy of few-shot fine-grained recognition will significantly boost.

Visualization Results Fig. 4 illustrates some examples of both hard- and soft-attention on these three datasets of birds, dogs and cars. In concretely, in the first row, we show the input images. The locations of these selected deep descriptors by our hard-attention are presented as red points over input images (note that we merely draw the first 6 locations for a clearer presentation). It is apparently to see that the selected descriptors indeed correspond fine-grained objects' part, *e.g.*, head, tail, ear, wheel and so on. Visualization of our soft-attention is shown by Grad-CAM (Selvaraju et al. 2017) in the third row. As seen, more detailed fine-grained patterns can attended by the soft-attention.

Configurations	CUB	Cars	Dogs
$\boldsymbol{A}=\boldsymbol{I}$	72.22	69.48	59.19
$\boldsymbol{A}$ in our DUAL ATT-NET	**72.98**	**70.41**	**59.81**

Table 4: Comparisons of whether considering the part correlation of fine-grained objects, where $\boldsymbol{I}$ is the identity matrix.

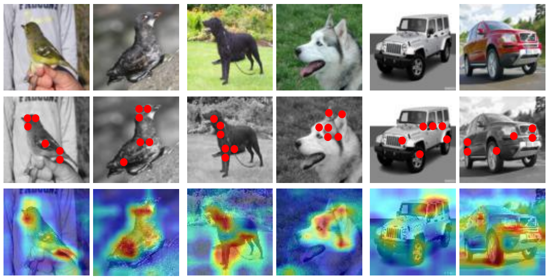


Figure 4: Visualization of our hard- and soft-attention on these three fine-grained datasets, *i.e.*, CUB, Cars, and Dogs.

Conclusion

In this paper, we proposed a Dual Attention Network for dealing with the few-shot fine-grained visual recognition task. Our DUAL ATT-NET consisted of two dual attention branches including hard- and soft-attention. The hard-attention explicitly kept useful deep descriptors corresponding to fine-grained objects' parts. Then, these descriptors were formed into a multi-instance bag for better modeling the relation of fine-grained parts. While, the soft-attention can bring comprehensive attention information as the counterpart of hard-attention. By aggregating both outputs of dual attentions, a holistic representation was obtained w.r.t. input images. By performing meta-learning in an end-to-end manner, we can learn a good image embedding in such a metric space to generalize to novel fine-grained classes. Experiments on three fine-grained benchmark datasets validated the effectiveness of our DUAL ATT-NET. In the future, an advanced method of modeling part correlation deserves further exploration.

Acknowledgements

This work was supported by National Key R&D Program of China (2021YFA1001100), Natural Science Foundation of Jiangsu Province of China under Grant (BK20210340), the Fundamental Research Funds for the Central Universities (No. 30920041111), and Beijing Academy of Artificial Intelligence (BAAI). We also gratefully acknowledge the support of MindSpore, CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research.

References

Andrews, S.; Tsochantaridis, I.; and Hofmann, T. 2003. Support vector machines for multiple-instance learning. In *NIPS*, 577–584.

Angles, B.; Jin, Y.; Kornblith, S.; Tagliasacchi, A.; and Yi, K. M. 2021. MIST: Multiple instance spatial transformer. In *CVPR*, 2412–2422.

Antoniou, A.; Edwards, H.; and Storkey, A. 2019. How to train your MAML. *arXiv preprint: 1810.09502*.

Carbonneau, M.-A.; Cheplygina, V.; Granger, E.; and Gagnon, G. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77: 329–353.

Chu, W.-H.; Li, Y.-J.; Chang, J.-C.; and Wang, Y.-C. F. 2019. Spot and learn: A maximum-entropy patch sampler for few-shot image classification. In *CVPR*, 6251–6260.

Clevert, D.-A.; Unterthiner, T.; and Hochreiter, S. 2016. Fast and accurate deep network learning by exponential linear units (ELUs). In *ICLR*, 1–14.

Deng, J.; Krause, J.; Stark, M.; and Li, F.-F. 2015. Leveraging the wisdom of the crowd for fine-grained recognition. *IEEE TPAMI*, 38(4): 666–676.

Dietterich, T. G.; Lathrop, R. H.; and Lozano-Pérez, T. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2): 31–71.

Dubey, A.; Gupta, O.; Raskar, R.; and Naik, N. 2018. Maximum-entropy fine grained classification. In *NeurIPS*, 635–645.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 1–10.

Finn, C.; Xu, K.; and Levine, S. 2018. Probabilistic model-agnostic meta-learning. In *NeurIPS*, 9537–9548.

Guo, Y.; and Cheung, N.-M. 2020. Attentive weights generation for few shot learning via information maximization. In *CVPR*, 13499–13508.

Horn, G. V.; Aodha, O. M.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The iNaturalist species classification and detection dataset. In *CVPR*, 8769–8778.

Hou, S.; Feng, Y.; and Wang, Z. 2017. VegFru: A domain-specific dataset for fine-grained visual categorization. In *ICCV*, 541–549.

Huang, H.; Zhang, J.; Yu, L.; Zhang, J.; Wu, Q.; and Xu, C. 2021. TOAN: Target-oriented alignment network for fine-grained image categorization with few labeled samples. *IEEE T-CSVT*. Doi:10.1109/TCSVT.2021.3065693.

Huang, H.; Zhang, J.; Zhang, J.; Wu, Q.; and Xu, J. 2019. Compare more nuanced: Pair wise alignment bilinear network for few-shot fine-grained learning. In *ICME*, 91–96.

Huang, H.; Zhang, J.; Zhang, J.; Xu, J.; and Wu, Q. 2020. Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. *IEEE TMM*, 23: 1666–1680.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *ICML*, 2127–2136.

Jamal, M. A.; and Qi, G.-J. 2019. Task agnostic meta-learning for few-shot learning. In *CVPR*, 11719–11727.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Fei-Fei, L. 2011. Novel dataset for fine-grained image categorization. In *CVPR Workshop on Fine-Grained Visual Categorization*, 806–813.

Kim, J.; Kim, T.; Kim, S.; and Yoo, C. D. 2019. Edge-labeling graph neural network for few-shot learning. In *CVPR*, 11–20.

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*, 1–10.

Kong, S.; and Fowlkes, C. 2017. Low-rank bilinear pooling for fine-grained classification. In *CVPR*, 365–374.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3D object representations for fine-grained categorization. In *ICCV Workshop on 3D Representation and Recognition*.

Lee, K.; Maji, S.; Ravichandran, A.; and Soatto, S. 2019. Meta-learning with differentiable convex optimization. In *CVPR*, 10657–10665.

Lee, Y.; and Choi, S. 2018. Gradient-based meta-learning with learned layerwise metric and subspace. In *ICML*, 2927–2936.

Li, W.; Wang, L.; Xu, J.; Huo, J.; Gao, Y.; and Luo, J. 2019. Revisiting local descriptor based image-to-class measure for few-shot learning. In *CVPR*, 7260–7268.

Li, X.; Wu, J.; Sun, Z.; Ma, Z.; Cao, J.; and Xue, J.-H. 2020. BSNet: Bi-similarity network for few-shot fine-grained image classification. *IEEE TIP*, 30: 1318–1331.

Lifchitz, Y.; Avrithis, Y.; Picard, S.; and Bursuc, A. 2019. Dense classification and implanting for few-shot learning. In *CVPR*, 9258–9267.

Lin, T.-Y.; RoyChowdhury, A.; and Maji, S. 2017. Bilinear convolutional neural networks for fine-grained visual recognition. *IEEE TPAMI*, 40(6): 1309–1322.

Liu, C.; Xie, H.; Zha, Z.-J.; Ma, L.; Yu, L.; and Zhang, Y. 2020. Filtration and distillation: Enhancing region attention for fine-grained visual categorization. In *AAAI*, 11555–11562.

Luketina, J.; Berglund, M.; Greff, K.; and Raiko, T. 2016. Scalable gradient-based tuning of continuous regularization hyperparameters. In *ICML*, 2952–2960.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Oreshkin, B. N.; Rodriguez, P.; and Lacoste, A. 2018. TADAM: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS*, 719–729.

Reed, S.; Akata, Z.; Lee, H.; and Schiele, B. 2016. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 49–58.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 618–626.

Shell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *NeurIPS*, 4077–4087.

Sun, X.; Chen, L.; and Yang, J. 2019. Learning from web data using adversarial discriminative neural networks for fine-grained classification. In *AAAI*, 273–280.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *CVPR*, 1199–1208.

Tang, P.; Wang, X.; Bai, S.; Shen, W.; Bai, X.; Liu, W.; and Yuille, A. 2018. PCL: Proposal cluster learning for weakly supervised object detection. *IEEE TPAMI*, 42(1): 176–191.

Tu, M.; Huang, J.; He, X.; and Zhou, B. 2019. Multiple instance learning with graph neural networks. *arXiv preprint arXiv:1906.04881*.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2016. Matching networks for one shot learning. In *NeurIPS*, 3630–3638.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD birds-200-2011 dataset. *Techique Report CNS-TR-2011-001*.

Wang, Q.; Ruan, L.; and Si, L. 2014. Adaptive knowledge transfer for multiple instance learning in image classification. In *AAAI*, 1334–1340.

Wang, X.; Yan, Y.; Tang, P.; Liu, W.; and Guo, X. 2019. Bag similarity network for deep multi-instance learning. *Information Sciences*, 504: 578–588.

Wang, Z.; Wang, S.; Li, H.; Dou, Z.; and Li, J. 2020. Graph-propagation based correlation learning for weakly supervised fine-grained image classification. In *AAAI*, 12289–12296.

Wei, X.-S.; Luo, J.-H.; Wu, J.; and Zhou, Z.-H. 2017. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE TIP*, 26(6): 2868–2881.

Wei, X.-S.; Song, Y.-Z.; Aodha, O. M.; Wu, J.; Peng, Y.; Tang, J.; Yang, J.; and Belongie, S. 2021. Fine-grained image analysis with deep learning: A Survey. *IEEE TPAMI, in press*. Doi:10.1109/TPAMI.2021.3126648.

Wei, X.-S.; Wang, P.; Liu, L.; Shen, C.; and Wu, J. 2019. Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples. *IEEE TIP*, 28(12): 6116–6125.

Wei, X.-S.; Xie, C.-W.; Wu, J.; and Shen, C. 2018. Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognition*, 76: 704–714.

Wei, X.-S.; and Zhou, Z.-H. 2016. An empirical study on image bag generators for multi-instance learning. *Machine Learning*, 105(2): 155–198.

Xu, X. 2003. *Statistical learning in multiple instance problems*. Ph.D. thesis, The University of Waikato.

Yang, L.; Li, L.; Zhang, Z.; Zhou, X.; Zhou, E.; and Liu, Y. 2020. DPGN: Distribution propagation graph network for few-shot learning. In *CVPR*, 13390–13399.

Ying, R.; You, J.; Morris, C.; Ren, X.; and Hamilton, W. L. 2018. Hierarchical graph representation learning with differentiable pooling. In *NeurIPS*, 4805–4815.

Yoon, S. W.; Seo, J.; and Moon, J. 2019. TapNet: Neural network augmented with task-adaptive projection for few-shot learning. In *ICML*, 7115–7123.

Yuan, T.; Wan, F.; Fu, M.; Liu, J.; Xu, S.; Ji, X.; and Ye, Q. 2021. Multiple instance active learning for object detection. In *CVPR*, 317–326.

Zhang, Y.-L.; and Zhou, Z.-H. 2017. Multi-instance learning with key instance shift. In *IJCAI*, 3441–3447.

Zhou, Z.-H.; Sun, Y.-Y.; and Li, Y.-F. 2009. Multi-instance learning by treating instances as non-iid samples. In *ICML*, 1249–1256.

Zhou, Z.-H.; Zhang, M.-L.; Huang, S.-J.; and Li, Y.-F. 2012. Multi-instance multi-label learning. *Artificial Intelligence*, 176(1): 2291–2320.

Zhu, Y.; Liu, C.; and Jiang, S. 2020. Multi-attention meta learning for few-shot fine-grained image recognition. In *IJCAI*, 1090–1096.