Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Full length article

Spatial self-attention network with self-attention distillation for fine-grained image recognition[☆]

Adu Asare Baffour [a], Zhen Qin [a,b,*], Yong Wang [c], Zhiguang Qin [a,b], Kim-Kwang Raymond Choo [d]

[a] School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
[b] Network and Data Security Key Laboratory of Sichuan Province, Chengdu 610054, China
[c] Zhengzhou Aiwen Computer Technology Co. Ltd., Henan 450000, China
[d] Department of Information Systems and Cyber Security, University of Texas at San Antonio, San Antonio, TX 78249-0631, USA

ARTICLE INFO

Keywords:
Fine-grained recognition
Spatial self-attention
Knowledge distillation
Convolutional neural network

ABSTRACT

The underlining task for fine-grained image recognition captures both the inter-class and intra-class discriminate features. Existing methods generally use auxiliary data to guide the network or a complex network comprising multiple sub-networks. They have two significant drawbacks: (1) Using auxiliary data like bounding boxes requires expert knowledge and expensive data annotation. (2) Using multiple sub-networks make network architecture complex and requires complicated training or multiple training steps. We propose an end-to-end Spatial Self-Attention Network (SSANet) comprising a spatial self-attention module (SSA) and a self-attention distillation (Self-AD) technique. The SSA encodes contextual information into local features, improving intra-class representation. Then, the Self-AD distills knowledge from the SSA to a primary feature map, obtaining inter-class representation. By accumulating classification losses from these two modules enables the network to learn both inter-class and intra-class features in one training step. The experiment findings demonstrate that SSANet is effective and achieves competitive performance.

1. Introduction

Fine-grained image recognition (FGIR), an essential task in image representation, facilitates the differentiation of detailed visual features for various sub-categories in a super-category, e.g., types of retail products, models of cars, and species of birds. There are various FGIR applications in image representation, such as image-based recommendation systems [1], image/video search systems [2], and image captioning [3]. Hence, FGIR is a significant research area and a rapidly growing sub-field in image recognition.

Although deep learning networks can automatically extract essential features [4], especially convolutional neural networks (CNN) for image and video representation [2,5], FGIR remains a challenging task that requires learning detailed discriminate visual details. Hence, there has been a focus on learning optimum representation for both the subtle and discriminate details.

Existing FGIR methods can be broadly categorized into strongly supervised and weakly supervised methods [6]. A strongly supervised method uses auxiliary data (e.g., manual annotation and bounding boxes) and image labels to train a network. Weakly supervised approaches use only image labels to capture certain local regions for part localization [6–14]. In recent years, attention-based methods (type of weakly supervised method) are becoming popular due to their capability of end-to-end training without needing auxiliary data. Attention-based methods use CNNs to build a localizing sub-network to capture essential parts of the image. Then, another sub-network is employed for the final classification.

However, some known limitations are associated with these approaches. For example, the number of object parts to be attended is limited and already defined; consequently, limiting the model's effectiveness and flexibility. Also, designing and training CNN sub-networks for each attention part of an object is inefficient, leading to bottlenecks in the CNN network. In addition, while attended parts can be concatenated, they cannot influence the relationship between the multiple localized parts in a global view, which is also crucial to FGIR. These limitations necessitate the design of an efficient model to automatically capture unlimited, primary features (coarse-grained) and attention features (fine-grained) in a way that can be relational to each other.

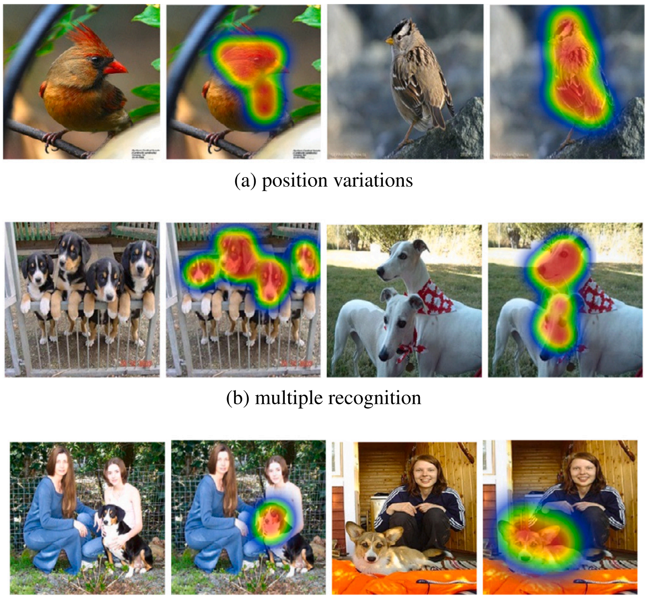
This paper introduces a spatial self-attention module to obtain meaningful features that encapsulate the feature relationships along the

[☆] This paper has been recommended for acceptance by Dr Zicheng Liu.

[*] Corresponding author.

E-mail addresses: baffour@std.uestc.edu.cn (A.A. Baffour), qinzhen@uestc.edu.cn (Z. Qin), wangyong@ipplus360.com (Y. Wang), qinzg@uestc.edu.cn (Z. Qin), raymond.choo@fulbrightmail.org (K.-K.R. Choo).

https://doi.org/10.1016/j.jvcir.2021.103368
Received 3 September 2020; Received in revised form 18 July 2021; Accepted 23 October 2021
Available online 5 November 2021



(a) position variations

(b) multiple recognition

(c) crowded background

Fig. 1. *Best viewed in color.* Visualizing SSANet results for (a) accurate feature localization even at different birds' postures (b) multiple dog recognition in a single image (c) accurate dog recognition in a challenging background.

spatial dimension, representing fine-grained features. A self-attention distillation transfers knowledge from the spatial self-attention module's output to the primary features (coarse-grained). The experiment findings demonstrate that SSANet achieves competitive classification performance on the CUB-Birds [15], Stanford Dogs [16], and Stanford Cars [17] datasets. The main contributions of the paper are summarized as follows:

1. We propose an efficient spatial self-attention network, a trainable end-to-end CNN architecture that does not require annotated data.
2. We introduce a spatial self-attention technique that leverages pixel locations in feature maps for similar correlation while maintaining the original spatial dimension.
3. We propose a simple but effective self-attention distillation technique, which steadily distills attention feature knowledge to the primary features (coarse-grained) output layer.
4. We conduct extensive experiments and report insightful findings on CUB-Birds, Stanford Dogs, and Stanford Cars.

We organize the remaining sections as follows. Section 2 briefly revisits the related literature. The proposed SSANet is described in Section 3. Comprehensive experiments, performance evaluations, and discussions are conducted in Section 4. Section 5 outlines the conclusion.

2. Related work

A crucial component often used in FGIR models is object region localization. Hence, we categorize existing methods according to how they learn to localize an object region. We group them into strongly supervised methods and weakly supervised methods. Since our work is related to the latter, we will focus our discussion on attention-based methods in weakly supervised methods.

Strongly supervised methods. A simple technique to locate an object region is by guiding a network to learn with object annotations or part/keypoint annotation [18–20]. For example, Zhang et al. [18] used double detectors (i.e., whole and part object detector) and imposed learned geometric constraints between them to make predictions from pose-normalized representation. Lin et al. [19] presented a network containing three sub-networks, which integrates part localization, alignment, and classification into one network using a valve linkage function. Wei et al. [20] employed a Mask-CNN to locate and generate object/part masks using part annotation.

Weakly supervised methods. The methods in this group require only image labels to avoid the labor required to obtain part annotations for part/object feature detection. For example, Fu et al. [11] presented a multi-scale network that uses R-CNN to learn to discriminate parts at different scales repeatedly. Zheng et al. [21] proposed locating and learning multiple object parts and enforced the correlation between them using three sub-networks (i.e., convolution sub-network, channel grouping sub-network, and part classification sub-network). Similar to [21], Sun et al. [22] proposed to regulate several object parts in different input images. Specifically, a one-squeeze multiple-excitation module learns multiple attention features, and a constraint module is employed to place similar class features together while separating different class features. Weakly supervised methods generally seek to capture multiple semantic parts before sharing the appearances across fine-grained categories while preserving the subtle differences between these part representations.

Attention-based methods. Implementing attention techniques in neural networks helps the network focus on the essential parts of a problem, maximizing accuracy and efficiency. Apart from FGIR, many other attention mechanisms have been proposed to capture essential information in various tasks such as pedestrian re-identification [23], machine translation, and human activity recognition. Similarly, FGIR attention mechanisms seek to capture essential details in images to complicate joint training of multiple localization or classification sub-networks.

He and Peng [24], for example, combined vision and language to achieve two-level attention. The vision-level attention takes as input a localized image and outputs attention prediction from the image view only. Then, the language-level attention outputs classification by computing the shared compatibility with the vision-level attention. Yang et al. [25] proposed to train a Teacher agent that focuses on a Navigator agent to detect the most significant informative regions. Other recent attention mechanisms are proposed in the literature [13,14,26,27]. For instance, Mnih et al. [26] presented an RNN that guides the proposed attention region using reinforcement learning. Linsley et al. [27] reinforced the significance of human attention for FGIR by guiding a network to focus attention on the same image parts, which humans consider essential for recognition via ClickMe.[1]

Generally, the methods for FGIR further concatenate several part/object features into a single image representation and feed them into a sub-network for final classification. Also, the attention-based networks either rely on RNN sub-networks or a heavily parameterized fully-connected sub-network for attention localization. Such approaches are highly prone to overfitting and can be challenging to implement due to architecture complexity.

Contrary to these approaches, we propose to learn the optimum whole image (not part/object level) feature representation at a single forward pass. Furthermore, the proposed SSANet uses only image labels for training and does not require human-annotated parts or bounding boxes.

3. Spatial self-attention network with self-attention distillation

The proposed spatial self-attention network (SSANet) consists of three main components (see Fig. 2). We first implement a CNN as a feature extractor (FE) that extracts primary features from the input images through several convolution and pooling layers (see Section 3.1). Then,

[1] A simple game that allows dataset downloads in large-scale. http://clickme.ai.

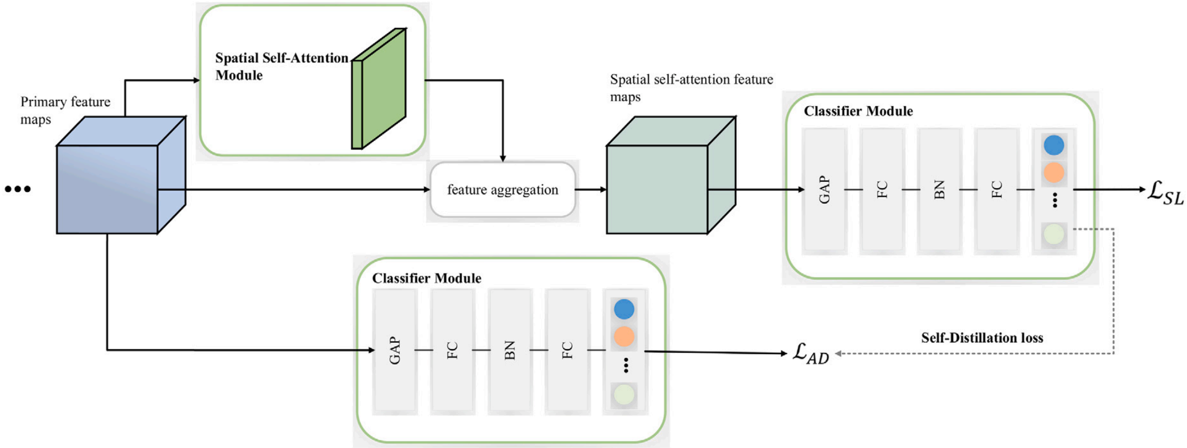


Fig. 2. Our proposed SSANet. It produces two separate loss outputs from the spatial self-attention maps and primary feature maps via separate classifier modules. GAP means global average pooling. FC means a fully connected layer. BN means batch normalization.

Table 1
The detailed information of the FGIR datasets used for experiments.

Datasets	Categories	Training	Testing
CUB-Birds [15]	200	5,994	5,794
Stanford Dogs [16]	120	12,000	8,580
Stanford Cars [17]	196	8,144	8,041

the spatial self-attention (SSA) module detects spatial self-attention regions (see Section 3.2). Further, a classifier module takes the primary features and spatial self-attention features and refines the learned features for better representation (see Section 3.4).

Note that we will refer to SSA output layer as the prediction output layer in the classifier module implemented exclusively for the SSA module. Likewise, the FE output layer refers to the prediction output layer in the classifier module implemented exclusively for FE. Finally, a self-attention distillation (Self-AD) distills knowledge from the SSA output layer to the FE output layer (see Section 3.3).

3.1. Feature extractor

Recently, for image processing tasks, a conventional approach towards extracting primary features is using a pre-trained CNN in order to benefit from meaningful weight initialization. Such pre-trained CNNs can extract high-level features from images. For a fair comparison with other state-of-the-art methods, we employ VGG-16 [5] pre-trained on the ImageNet dataset.

We remove the last three fully connected layers to extract primary feature maps from input images. For a given input image X of a dataset, we take the feature map output from the final convolutional layer. This process is illustrated as

$$F = VGG(X).$$

Specifically, the CNN feature extractor maps an input image X to a K dimensional feature map $F \in \mathbb{R}^{H \times W \times K}$, where H, W, and K represent the spatial height, spatial width, and the number of channels/kernel contained in the last layer, respectively.

3.2. Spatial self-attention

The spatial self-attention module leverage a self-attention mechanism proposed by [28], which accumulates local attention through a softmax function. We extend this idea to attend to the spatial pixel positions of the original features and apply feature aggregation to obtain spatial self-attention feature maps.

As illustrated in Fig. 3, given primary feature maps $F \in \mathbb{R}^{H \times W \times K}$ obtained from the FE, we first generate three new feature maps A, B, and C using 1×1 convolution. $\{A, B, C\} \in \mathbb{R}^{H \times W \times K}$ have the same spatial dimension as F. We reshape A, B, C to $\mathbb{R}^{N \times K}$, where $N = H \times W$ is the number of pixels. Then, we compute element-wise multiplication between A and B transpose. We apply spatial-wise softmax to compute the spatial self-attention map $S \in \mathbb{R}^{N \times N}$ as

$$S_{j,i} = \frac{\exp(A_i \otimes B_j)}{\sum_{i=1}^{N} \exp(A_i \otimes B_j)},$$

where $\otimes$ represents element-wise multiplication. The S_{ji} shows how the network measures the ith spatial location's influence on the jth spatial position. Hence, more related feature representations between A and B results in a more significant and enriched correlation between them and vice versa. To amplify the attended positions, we perform element-wise multiplication between $S \in \mathbb{R}^{N \times N}$ and $C \in \mathbb{R}^{N \times K}$ and reshape the results to $\mathbb{R}^{H \times W \times K}$.

Finally, we implement a feature aggregation mechanism to infer the influence of the spatial self-attention regions across all positions on the original feature map F through Eqs. (1) and (2) as

$$H_j^{plus} = \alpha \sum_{i=1}^{N} \left(s_{ji} C_i \right) \oplus F_j \tag{1}$$

$$H_j^{mul} = \alpha \sum_{i=1}^{N} \left(s_{ji} C_i \right) \otimes F_j \tag{2}$$

We initialize a learnable scale parameter $\alpha = 0$. We introduce α as a learnable parameter to enable the network to initially rely on the local spatial neighborhood and steadily assign more weight to the attended regions. It can be deduced from Eqs. (1) and (2) that the features obtained by H_j^{plus} and H_j^{mul} represent an overall aggregation of the contextual view according to the spatial self-attention maps. Experiment results of the Eqs. (1) and (2) are discussed in Section 4.1.

3.3. Self-attention distillation

First, we formulate a standard supervised loss $\mathcal{L}_{SL}$ from the cross-entropy (CE) as

$$\mathcal{L}_{SL} = \mathrm{CE}(y, \hat{y}_{SSA}). \tag{3}$$

We denote y and $\hat{y}_{SSA}$ as the ground-truth label and SSA predicted label, respectively. However, since the proposed SSANet can be integrated into existing CNNs, we use both the SSA output layer and FE output layer during training, as shown in Fig. 2. In particular, we transfer knowledge from the SSA output layer to the FE's output layer.

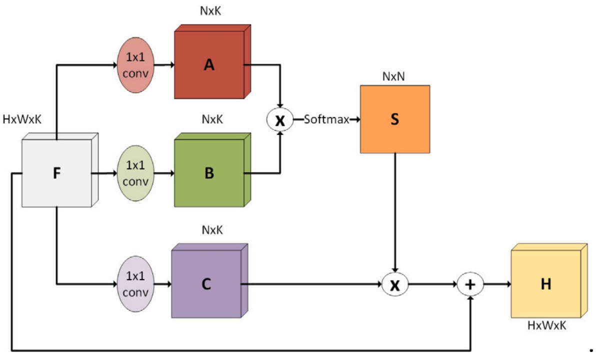


Fig. 3. The specifics of the SSA module. We denote ⊗ as element-wise multiplication and ⊕ as element-wise addition.

Table 2

FGIR results and comparison on the CUB-Birds (Birds), Stanford Cars (Cars), and Stanford Dogs (Dogs) datasets. The "Aux." means training with a bounding box or part annotation. We did not include numbers for some state-of-the-art baselines because the authors did not report on the underlining dataset in their paper.

Method	Aux	Birds	Cars	Dogs
B-CNN [8]	No	84.1	–	–
SPDA-CNN [29]	Yes	85.1	–	–
FCAN [9]	No	84.3	91.3	84.5
DB [30]	No	–	–	87.7
FCAN [9]	Yes	–	89.1	–
PDFR [10]	No	82.6	–	71.9
RA-CNN [11]	No	85.3	92.5	87.3
DVAN [6]	No	79.0	87.1	81.5
PC-DenseNet-161 [31]	No	–	–	83.6
LTPA [32]	Yes	73.1	–	–
WS-DAN [33]	No	–	–	90.0
DFL-CNN [34]	No	86.5	93.8	–
WLHR [7]	No	83.7	–	–
HDWE [35]	Yes	84.3	–	76.9
ACBNT [12]	No	87.6	–	–
EfficientNet-B0 [36]	No	–	–	61.2
HPB [13]	No	87.1	93.7	–
SEF [37]	No	–	–	88.8
WSCP [38]	Yes	86.2	–	–
NPA [39]	Yes	–	92.8	–
PC [14]	No	–	83.6	61.9
HIHCA [40]	No	85.3	91.7	–
DCL-VGG [41]	No	86.9	94.1	–
Bilinear CNN [42]	No	84.1	91.3	–
OPAM [43]	No	85.8	92.2	–
API-Net [44]	No	–	–	90.3
Our SSANET$_{\text{no-attention}}$	No	75.0	82.9	77.9
Our SSANET$_{\text{no-alpha}}$	No	82.3	91.5	87.0
Our SSANET$_{\text{alpha}}$	No	82.7	92.0	88.1
Our SSANET$_{\text{multiply}}$	No	86.5	93.0	90.4
Our SSANET$_{\text{plus}}$	**No**	**88.4**	**94.2**	**92.2**

To this end, we design a self-attention distillation (Self-AD) loss $\mathcal{L}_{AD}$, containing outputs from the SSA and FE output layers as

$$\mathcal{L}_{AD} = \mathrm{KDL}(\hat{y}_{SSA}, \hat{y}_{FE}) + \gamma \mathrm{CE}(y, \hat{y}_{SSA}), \tag{4}$$

where $\hat{y}_{FE}$ represents the FE output logits, and γ is the weight factor of the two loss terms. We empirically set $\gamma = 0.5$. The KDL is the Kullback–Leibler divergence introducing relative entropy between the two modules. Ultimately, the total loss $\mathcal{L}$ accumulates the output of Eqs. (3) and (4) as

$$\mathcal{L} = \mathcal{L}_{SL} + \mathcal{L}_{AD}, \tag{5}$$

we may achieve more advancement by applying different loss ratios.

3.4. Classifier

To fully benefit from global contextual information and intra-class semantic information, we aggregate the features learned from the FE and SSA modules. We apply global average pooling for each set of feature maps of FE and SSA module, a fully connected layer of 1024 units, 1-D batch normalization, and a softmax output layer.

4. Experiments

This section presents experiment procedures, results, and discussions of the proposed SSANet.

4.1. Setup

We assess the performance in two views: spatial self-attention feature accuracy and the most effective feature aggregation mechanism between Eqs. (1) and (2).

Models. Specifically, we show the effectiveness of our method by initially training the network with no SSA module (i.e., including the FE and the classifier module), herein referred to as SSANet$_{\text{no-attention}}$. We then experiment with two variations of the SSA module: (1) with a learnable parameter α, herein called SSANet$_{\text{alpha}}$, (2) with no learnable parameter α, herein called SSANet$_{\text{no-alpha}}$. We experimented with the two feature aggregation mechanisms expressed in Eqs. (1) and (2) in a 'fully fleshed' model (i.e., containing all modules). Eq. (1) model herein referred to as SSANet$_{\text{plus}}$, and Eq. (2) model referred to as SSANet$_{\text{multiply}}$. The experiment results are shown in Table 2.

Dataset. We carry out investigations on three FGIR datasets: CUB-Birds [15], Stanford Dogs [16], and Stanford Cars [17]. The information about the number of classes and train/test splits is outlined in Table 1.

Implementation Details. For data augmentation, we adhere to the standard practices adopted in the literature [9,11,13,22,34]; thus, resized images to 448 × 448 for both train and test images; and horizontally flipping train images. VGG-16 is used in the FE, containing 16 convolution layers, 4 max-pool layers, 3 fully-connected layers, and a softmax layer. We discard the last four layers to create a feature extractor, thus, we do not update the ImageNet pre-trained weights. The results from the last convolutional layer after the max-pool layer (7 × 7) become the SSA module's input.

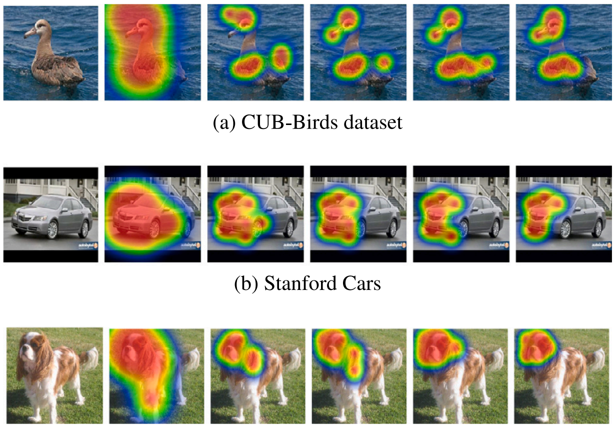
During training, we feed-forward training samples through the network to output the prediction. Thus, the training process is a single-step process with no separate module training. We train the network using the stochastic gradient descent optimizer with a momentum of 0.9, weight decay of $5{\times}10^{-4}$, and 10^{-3} learning rate. We reduce the learning rate by a factor of 0.3 when learning stagnates. All the model and training hyperparameters were empirically selected based on preliminary experiments.

4.2. Results of CUB-birds

Table 2 presents the performance comparison with other baseline methods [6–13,29,32,34,35,38,40–43]. For this task, even with the availability of human-defined annotations, we train the network with only the class labels. However, some of the baseline methods used the bounding box data to train the network. We make some conclusions from Table 2.

Firstly, we observe superior results of SSANet$_{\text{plus}}$ contrasting with baseline attention methods [29,32,35,38] that utilized bounding boxes. Specifically, SSANet$_{\text{plus}}$ achieves higher accuracy of +3.3%, +15.3%, +4.1%, and +2.2% than SPDA-CNN [29], LTPA [32], HDWE [35], and WSCP [38], respectively.

Second, we compare weakly supervised attention baselines with no bounding boxes [6–13,34,40–43]. B-CNN [8] used bilinear vectors from double CNN streams to obtain 84.1% accuracy. FCAN [9] adopted



(a) CUB-Birds dataset

(b) Stanford Cars

(c) Stanford Dogs

Fig. 4. *Best viewed in color.* Visualizing SSANet results of the various SSANet models. For each dataset (a, b, c), the images are arranged in this order (from left to right): input image, SSANet$_{\text{no-attention}}$, SSANet$_{\text{no-alpha}}$, SSANet$_{\text{alpha}}$, SSANet$_{\text{multiply}}$, and SSANet$_{\text{plus}}$.

reinforcement learning for attention region localization and achieved 84.3% accuracy. PDFR [10] selected CNN filters to learn part detectors and local CNN descriptions to obtain 82.6% accuracy. DVAN [6] proposed multiple attention canvas and LSTM recurrent units to locate attention regions and achieved only 79% accuracy. DFL-CNN [34] leverages mid-level convolutional filters to achieve 86.5% accuracy. WLHR [7] exploits the hierarchical relationship between coarse and fine classes. ACBNT [12] proposed root-to-leaf convolutional operation pathways in a tree structure and obtained accuracy up to 87.6%. HPB [13] proposed pooling and integrating cross-layer features in a bilinear fashion and achieved 87.1% accuracy.

We observe that the use of spatial self-attention improves the feature representation and helps improve performance. The performance varies depending on the discriminative power of the models. The SSANET$_{\text{no-attention}}$ model achieves the lowest performance among the baseline methods. However, the performance improves by enhancing the feature representation via the SSA module, and we achieve superior results to the baseline methods via the SSANET$_{\text{multiply}}$ and SSANET$_{\text{plus}}$ models. We demonstrate that the proposed network automatically learns to locate discriminative spatial regions essential for classification, mainly located on birds' heads, beaks, and wings, as shown in Fig. 4(a).

4.3. Results of Stanford Cars

Table 2 shows the comparison results of our methods with other baseline methods [6,9,11,13,14,34,39–43]. This dataset also contains bounding box data. We do not use the bounding box data but only used the class labels for training. Meanwhile, some baseline methods [9,39] used the bounding box data during training.

The baseline methods FCAN [9], DVAN [6], RA-CNN [11], DFL-CNN [34], and HPB [13] (briefly discussed in Section 4.2) achieved an accuracy of 91.3%, 87.1%, 92.5%, and 93.7%, respectively. NPA [39] used bounding boxes for training to obtain performance accuracy up to 92.8%. PC [14] used pairwise confusion regularization in activations to achieve 83.6% accuracy. HIHCA [40] used a polynomial kernel predictor to capture higher-order statistics of convolutional activation and achieved 91.7%. DCL-VGG [41] proposed to "destruct" and "reconstruct" the input image to learn discriminative features and achieved 94.1%. Bilinear CNN [42] multiplied the output of two feature extractors to obtain image descriptors and achieved 91.3% performance accuracy. OPAM [43] used object-level attention and part-level attention and achieved 92.2% performance accuracy.

We make the following conclusions. Contrary to the bounding box baselines [9,39], our SSANet$_{\text{multiply}}$ and SSANet$_{\text{plus}}$ models automatically find the discriminative regions (see Fig. 4(b)) without requiring any bounding boxes and achieve superior accuracy. We also obtain the best performance of up to 94.2% among baseline methods [6,11, 13,14,34,40–43] that do not use bounding box data. Note that we do not employ a separate RNN sub-network for localization, additional regularization, or cross-layer features to achieve performance accuracy. We train the SSANet in a single training step without bounding box data, making our method advantageous and less complex to implement.

4.4. Results of Stanford Dogs

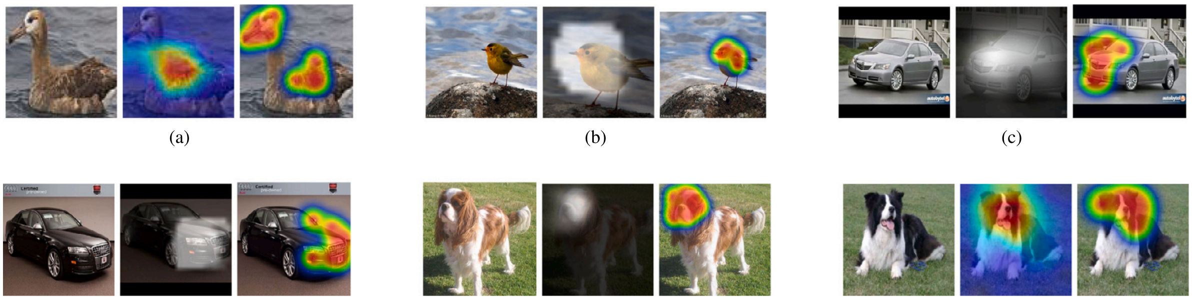
The performance comparison with baseline methods [6,9–11,14,30, 31,33,35–37] on Stanford Dogs is summarized in Table 2. This dataset also has bounding boxes. However, we similarly conduct experiments as CUB-Birds and Stanford Cars. Thus, we train the network with only class labels.

Nonetheless, HDWE [35] used the Clickture-Dog [47] dataset as an auxiliary click data source and achieved even less accuracy (−15.3%) to SSANET$_{\text{plus}}$. Comparing with the two most effective baseline methods: WS-DAN [33] and API-Net [44], we obtain higher performance accuracy of +2.2% and +1.9%, respectively. This performance is achieved through accurate spatial self-attention localization, as illustrated in Fig. 4(c). It is evident from Fig. 4(c) that the attended regions are often found around the dog's head.

4.5. Ablation study

To understand the contributions of the different network components, we study the results of ablation experiments on the three datasets.

Effectiveness of spatial self-attention module. The SSANet$_{\text{no-attention}}$ model extracts primary features and performs classification by the classifier module. The performance accuracies of



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 5. *Best viewed in color.* Attention localization comparison with other methods. (a), (d), and (f) are attention localization from DVAN [6]. Moreover, (c) and (e) are attention localization from [45]. Furthermore, (b) and (g) are attention localization from LTPA [32,46], respectively. All figures are taken from corresponding published papers.

Table 3
Assessing the effectiveness of Self-AD on CUB-Birds.

Loss	SSANet$_{alpha}$	SSANet$_{multiply}$	SSANet$_{plus}$
$\mathcal{L}_{CE}$	81.4	85.6	87.1
$\mathcal{L}_{AD}$	**82.7**	**86.5**	**88.4**

SSANet$_{\text{no-attention}}$ are 77.9%, 82.9%, and 75.0% for Stanford Dogs, Stanford Cars, and CUB-Birds datasets, respectively. After adopting the simplest variant of the SSA module, SSANet$_{\text{no-alpha}}$, we observe a significant improvement in performance up to +9.1% (87.0%) for Stanford Dogs, +8.6% (91.5%) for Stanford Cars, and +7.3% (82.3%) for CUB-Birds. This result demonstrates the significance and effectiveness of the SSA module for the SSANet.

Effectiveness of alpha. The learnable parameter α assigns more weights to the spatially attended regions. To show the effectiveness of α, we contrast the classification results of SSANet$_{\text{alpha}}$ with SSANet$_{\text{no-alpha}}$. We observe that SSANet$_{\text{alpha}}$ achieves slightly higher accuracy than SSANet$_{\text{no-alpha}}$ for all three datasets. Specifically, SSANet$_{\text{alpha}}$ gains accuracy of +1.1%, +0.5%, and +0.4% for Stanford Dogs, Stanford Cars, and CUB-Birds, respectively. This performance shows that α does not significantly affect the network as far as the SSA module is concerned. Hence, the SSA module can rely on the local spatial attended pixels for localization and visual discrimination.

Effectiveness of feature aggregation. For feature aggregation of primary features (coarse-grained) and spatial self-attention features (fine-grained), we experiment with two feature aggregations, Eqs. (1) and (2). The basic idea is to amplify the spatially attended regions by an element-wise combination.

Firstly, we train the network using Eq. (2); thus, the SSANet$_{\text{multiply}}$ model obtains performance accuracy of 90.4%, 93.0%, and 86.5% for Stanford Dogs, Stanford Cars, and CUB-Birds, respectively. Then we train using Eq. (1), thus, SSANet$_{\text{plus}}$, and record performance accuracy of 92.2% for Stanford Dogs, 94.2% for Stanford Cars, and 88.4% for CUB-Birds. We observe that SSANet$_{\text{plus}}$ attains greater accuracy than SSANet$_{\text{multiply}}$ for all three datasets. This result can be due to two reasons:

1. An element-wise multiplication of self-attention features carrying more weights (say 5.45) with less or zero weighted pixel in the primary features (say $F_j = 0.01$) can cause the resulting spatial self-attention features to be of less or no significance.
2. On the other hand, element-wise multiplication of self-attention features carrying less or no weights (say 0.01) with a more weighted pixel in the primary features (say $F_j = 5.45$) can cause the resulting SSA features to be misleading. Thus, making the primary features less or not significant.

Hence, Eq. (2) suffers imposition by higher/lower pixel values. Instead, Eq. (1) accounts for the higher/lower values through summation, which avoids value imposition.

Effectiveness of self-attention distillation. We further conduct experiments on Stanford birds to assess the effectiveness of our self-attention distillation. Rather than distilling outputs from the SSA output layer, the standard cross-entropy (CE) loss is exclusively applied to the SSA output layer and FE output layer at the training phase. The results are shown in Table 3. We observe that training the network with CE loss for both the SSA output layer and FE output layer results in a fall in performance compared to our self-attention distillation loss, $\mathcal{L}_{AD}$. This result implies that the attention knowledge transfer from the SSA module output layer to the FE output layer by our self-attention distillation is helpful to improving performance.

Visualization Visualizing artificial intelligent models helps to make better explanations and build user confidence. To visualize and explain the network's intuition, we use Grad-CAM [48]. Grad-CAM is a technique to display feature localization of CNN models using target gradients.

We replace the CNN part of the Grad-CAM model with our trained SSANet on each distinct dataset. We forward-pass through the updated Grad-CAM model with 448 × 448 images and their corresponding class labels as input. For a given image propagated through the CNN part of Grad-CAM, we obtain the prediction score for the category with the gradient of the target class set to 1 while all other classes' gradients are set to 0. This cue is then propagated back to the interested CNN feature maps, which are integrated to calculate the localization (heatmap) to show where the SSANet looks to make the prediction decision.

For Stanford birds visualization, we observe two phenomena:

1. The proposed SSANet learns that the discriminate part features of birds are located at the head, beak, and wings, which is consistent with previous research [6,9,11,32,39] as illustrated in Figs. 5(a) and 5(b).
2. As shown in Fig. 1(a), the proposed SSANet can localize discriminate parts of the bird at different angles.

Stanford dogs visualization reveals the following observations:

1. The proposed SSANet can accurately recognize multiple dogs in a picture at a single shot (see Fig. 1(b)).
2. The proposed SSANet learns that discriminate dog features are found at the head region. This finding is as well agreeable to previous research [6,10,11,46], as shown in Figs. 5(e) and 5(f).

In Stanford car visualization, the proposed SSANet learns to make classification mostly by localizing the car bonnet (Fig. 4(b)). This finding is consistent with previous research [6,45], as shown in Figs. 5(c) and 5(d).

5. Conclusion

We proposed a novel Spatial Self-Attention Network (SSANet) to facilitate FGIR, which automatically learns attention features by the SSA module and Self-AD. The SSA module effectively learns discriminate parts localization and significantly improves performance. We observed that the addition operation proves favorable in the coarse-grained and fine-grained feature aggregation mechanism, while the multiplication operation may suffer imposition by higher/lower pixel values. Our proposed Self-AD distills valuable knowledge from the SSA output layer to the FE output layer. The ablation studies indicate that our Self-AD further boosts performance. Finally, our SSANet can also be applied to alternative networks like Inception, ResNet, and DenseNet by removing their fully connected layers and plugging in the SSANet modules.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62076054, No. 62072074, No. 62027827), the Frontier Science and Technology Innovation Projects of National Key R&D Program (No. 2019QY1405), the Sichuan Science-Technology Support Plan Program (No. 2020YFSY0010, No. 2019YJ0636), the Sichuan Science and Technology Innovation Platform and Talent Plan (No. 2020JDJQ0020), and Kim-Kwang Raymond Choo was supported only by the Cloud Technology Endowed.

References

[1] W. Zhou, P. Mok, Y. Zhou, Y. Zhou, J. Shen, Q. Qu, K. Chau, Fashion recommendations through cross-media information retrieval, J. Vis. Commun. Image Represent. 61 (2019) 112–120.

[2] L. Jing, X. Yang, Y. Tian, Video you only look once: Overall temporal convolutions for action recognition, J. Vis. Commun. Image Represent. 52 (2018) 58–65.

[3] N. Xu, A.-A. Liu, J. Liu, W. Nie, Y. Su, Scene graph captioner: Image captioning based on structural visual representation, J. Vis. Commun. Image Represent. 58 (2019) 477–485.

[4] Z. Qin, Y. Zhang, S. Meng, Z. Qin, K.-K.R. Choo, Imaging and fusing time series for wearable sensor-based human activity recognition, Inf. Fusion 53 (2020) 80–87.

[5] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[6] B. Zhao, X. Wu, J. Feng, Q. Peng, S. Yan, Diversified visual attention networks for fine-grained object classification, IEEE Trans. Multimed. 19 (6) (2017) 1245–1256.

[7] Q. Jiao, Z. Liu, L. Ye, Y. Wang, Weakly labeled fine-grained classification with hierarchy relationship of fine and coarse labels, J. Vis. Commun. Image Represent. 63 (2019) 102584.

[8] T. Lin, A. RoyChowdhury, S. Maji, Bilinear CNN models for fine-grained visual recognition, in: 2015 IEEE International Conference on Computer Vision, ICCV, 2015, pp. 1449–1457, http://dx.doi.org/10.1109/ICCV.2015.170.

[9] X. Liu, T. Xia, J. Wang, Y. Lin, Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition, CoRR, abs/1603.06765, 2016.

[10] X. Zhang, H. Xiong, W. Zhou, W. Lin, Q. Tian, Picking deep filter responses for fine-grained image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 1134–1142, http://dx.doi.org/10.1109/CVPR.2016.128.

[11] J. Fu, H. Zheng, T. Mei, Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 4476–4484, http://dx.doi.org/10.1109/CVPR.2017.476.

[12] R. Ji, L. Wen, L. Zhang, D. Du, Y. Wu, C. Zhao, X. Liu, F. Huang, Attention convolutional binary neural tree for fine-grained visual categorization, CoRR, abs/1909.11378, 2019.

[13] C. Yu, X. Zhao, Q. Zheng, P. Zhang, X. You, Hierarchical bilinear pooling for fine-grained visual recognition, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision, ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI, in: Lecture Notes in Computer Science, vol. 11220, Springer, 2019, pp. 595–610, http://dx.doi.org/10.1007/978-3-030-01270-0_35.

[14] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, N. Naik, Training with confusion for fine-grained visual classification, CoRR, abs/1705.08016, 2017.

[15] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, P. Perona, Caltech-UCSD Birds 200, Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[16] A. Khosla, N. Jayadevaprakash, B. Yao, L. Fei-Fei, Novel Dataset for Fine-Grained Image Categorization, in: First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, 2011.

[17] J. Krause, M. Stark, J. Deng, L. Fei-Fei, 3D object representations for fine-grained categorization, in: 2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013, IEEE Computer Society, 2013, pp. 554–561, http://dx.doi.org/10.1109/ICCVW.2013.77.

[18] N. Zhang, J. Donahue, R. Girshick, T. Darrell, Part-based R-CNNs for fine-grained category detection, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision, ECCV 2014, Springer International Publishing, Cham, 2014, pp. 834–849.

[19] D. Lin, X. Shen, C. Lu, J. Jia, Deep LAC: Deep localization, alignment and classification for fine-grained recognition, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 1666–1674, http://dx.doi.org/10.1109/CVPR.2015.7298775.

[20] X.-S. Wei, C.-W. Xie, J. Wu, C. Shen, Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization, Pattern Recognit. 76 (2018) 704–714.

[21] H. Zheng, J. Fu, T. Mei, J. Luo, Learning multi-attention convolutional neural network for fine-grained image recognition, in: 2017 IEEE International Conference on Computer Vision, ICCV, 2017, pp. 5219–5227, http://dx.doi.org/10.1109/ICCV.2017.557.

[22] M. Sun, Y. Yuan, F. Zhou, E. Ding, Multi-attention multi-class constraint for fine-grained image recognition, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision, ECCV 2018, Springer International Publishing, Cham, 2018, pp. 834–850.

[23] Z. Qin, W. He, F. Deng, M. Li, Y. Liu, SRPRID: Pedestrian Re-Identification based on super-resolution images, IEEE Access 7 (2019) 152891–152899.

[24] X. He, Y. Peng, Fine-grained image classification via combining vision and language, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, 2017, pp. 7332–7340, http://dx.doi.org/10.1109/CVPR.2017.775.

[25] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, L. Wang, Learning to navigate for fine-grained classification, in: Computer Vision, ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV, 2018, pp. 438–454, http://dx.doi.org/10.1007/978-3-030-01264-9_26.

[26] V. Mnih, N. Heess, A. Graves, K. Kavukcuoglu, Recurrent models of visual attention, in: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, 2014, pp. 2204–2212, URL: http://papers.nips.cc/paper/5542-recurrent-models-of-visual-attention.

[27] D. Linsley, D. Scheibler, S. Eberhardt, T. Serre, Global-and-local attention networks for visual recognition, CoRR, abs/1805.08819, 2018.

[28] H. Zhang, I.J. Goodfellow, D.N. Metaxas, A. Odena, Self-attention generative adversarial networks, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, in: Proceedings of Machine Learning Research, 97, PMLR, 2019, pp. 7354–7363, URL: http://proceedings.mlr.press/v97/zhang19d.html.

[29] H. Zhang, T. Xu, M. Elhoseiny, X. Huang, S. Zhang, A.M. Elgammal, D.N. Metaxas, SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 1143–1152, http://dx.doi.org/10.1109/CVPR.2016.129.

[30] G. Sun, H. Cholakkal, S. Khan, F.S. Khan, L. Shao, Fine-grained recognition: Accounting for subtle differences between similar classes, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 12047–12054, URL: https://aaai.org/ojs/index.php/AAAI/article/view/6882.

[31] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, N. Naik, Pairwise confusion for fine-grained visual classification, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision, ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII, in: Lecture Notes in Computer Science, vol. 11216, Springer, 2018, pp. 71–88, http://dx.doi.org/10.1007/978-3-030-01258-8_5.

[32] S. Jetley, N. Lord, N. Lee, P. Torr, Learn to pay attention, Arxiv, abs/1804.02391, 2018.

[33] A. Imran, V. Athitsos, Domain adaptive transfer learning on visual attention aware data augmentation for fine-grained visual categorization, in: G. Bebis, Z. Yin, E. Kim, J. Bender, K. Subr, B.C. Kwon, J. Zhao, D. Kalkofen, G. Baciu (Eds.), Advances in Visual Computing - 15th International Symposium, ISVC 2020, San Diego, CA, USA, October 5-7, 2020, Proceedings, Part II, in: Lecture Notes in Computer Science, vol. 12510, Springer, 2020, pp. 53–65, http://dx.doi.org/10.1007/978-3-030-64559-5_5.

[34] Y. Wang, V.I. Morariu, L.S. Davis, Learning a discriminative filter bank within a CNN for fine-grained recognition, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, IEEE Computer Society, 2018, pp. 4148–4157, http://dx.doi.org/10.1109/CVPR.2018.00436, URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Learning_a_Discriminative_CVPR_2018_paper.html.

[35] J. Yu, M. Tan, H. Zhang, D. Tao, Y. Rui, Hierarchical deep click feature prediction for fine-grained image recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 1.

[36] D. Haase, M. Amthor, Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved MobileNets, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, IEEE, 2020, pp. 14588–14597, http://dx.doi.org/10.1109/CVPR42600.2020.01461.

[37] W. Luo, H. Zhang, J. Li, X. Wei, Learning semantically enhanced feature for fine-grained image classification, IEEE Signal Process. Lett. 27 (2020) 1545–1549, http://dx.doi.org/10.1109/LSP.2020.3020227.

[38] W. Ge, X. Lin, Y. Yu, Weakly supervised complementary parts models for fine-grained image classification from the bottom up, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 3034–3043, http://dx.doi.org/10.1109/CVPR.2019.00315, URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Ge_Weakly_Supervised_Complementary_Parts_Models_for_Fine-Grained_Image_Classification_From_CVPR_2019_paper.html.

[39] J. Krause, H. Jin, J. Yang, F. Li, Fine-grained recognition without part annotations, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, pp. 5546–5555, http://dx.doi.org/10.1109/CVPR.2015.7299194.

[40] S. Cai, W. Zuo, L. Zhang, Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 511–520, http://dx.doi.org/10.1109/ICCV.2017.63.

[41] Y. Chen, Y. Bai, W. Zhang, T. 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, Computer Vision Foundation / IEEE, 2019, pp. 5157–5166, http://dx.doi.org/10.1109/CVPR.2019.00530, URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Chen_Destruction_and_Construction_Learning_for_Fine-Grained_Image_Recognition_CVPR_2019_paper.html.

[42] T. Lin, A. RoyChowdhury, S. Maji, Bilinear convolutional neural networks for fine-grained visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1309–1322, http://dx.doi.org/10.1109/TPAMI.2017.2723400.

[43] Y. Peng, X. He, J. Zhao, Object-part attention model for fine-grained image classification, IEEE Trans. Image Process. 27 (3) (2018) 1487–1500, http://dx.doi.org/10.1109/TIP.2017.2774041.

[44] P. Zhuang, Y. Wang, Y. Qiao, Learning attentive pairwise interaction for fine-grained classification, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 13130–13137, URL: https://aaai.org/ojs/index.php/AAAI/article/view/7016.

[45] W. Shen, R. Liu, Generating attention from classifier activations for fine-grained recognition, CoRR, abs/1811.10770, 2018.

[46] X. Sun, H. Xv, J. Dong, H. Zhou, C. Chen, Q. Li, Few-shot learning for domain-specific fine-grained image classification, IEEE Trans. Ind. Electron. (2020) 1.

[47] X. Hua, L. Yang, J. Wang, J. Wang, M. Ye, K. Wang, Y. Rui, J. Li, Clickage: Towards bridging semantic and intent gaps via mining click logs of search engines, in: A. Jaimes, N. Sebe, N. Boujemaa, D. Gatica-Perez, D.A. Shamma, M. Worring, R. Zimmermann (Eds.), ACM Multimedia Conference, MM ’13, Barcelona, Spain, October 21-25, 2013, ACM, 2013, pp. 243–252, http://dx.doi.org/10.1145/2502081.2502283.

[48] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. (2019) http://dx.doi.org/10.1007/s11263-019-01228-7.