Selective Sparse Sampling for Fine-grained Image Recognition

Yao Ding[†§*],Yanzhao Zhou[†*],Yi Zhu[†],Qixiang Ye[†§‡] and Jianbin Jiao[†‡]
[†]University of Chinese Academy of Sciences, Beijing, China
[§]Peng Cheng Laboratory, Shenzhen, China
{dingyao16,zhouyanzhao215,zhuyi215}@mails.ucas.ac.cn, {jiaojb,qxye}@ucas.ac.cn

Abstract

*Fine-grained recognition poses the unique challenge of capturing subtle inter-class differences under considerable intra-class variances (*e.g.*, beaks for bird species). Conventional approaches crop local regions and learn detailed representation from those regions, but suffer from the fixed number of parts and missing of surrounding context. In this paper, we propose a simple yet effective framework, called Selective Sparse Sampling, to capture diverse and fine-grained details. The framework is implemented using Convolutional Neural Networks, referred to as Selective Sparse Sampling Networks (S3Ns). With image-level supervision, S3Ns collect peaks,* i.e.*, local maximums, from class response maps to estimate informative receptive fields and learn a set of sparse attention for capturing fine-detailed visual evidence as well as preserving context. The evidence is selectively sampled to extract discriminative and complementary features, which significantly enrich the learned representation and guide the network to discover more subtle cues. Extensive experiments and ablation studies show that the proposed method consistently outperforms the state-of-the-art methods on challenging benchmarks including CUB-200-2011, FGVC-Aircraft, and Stanford Cars*[1].

1. Introduction

Fine-grained recognition refers to identifying subordinate classes under a basic-level category in images, *e.g.*, bird species [24], flower breeds[19], car models [10], and aircraft types [15]. Compared with general image classification, fine-grain recognition is more challenging due to the subtle differences between inter-class images.

The study of cognitive neuroscience [8, 16] discovers that when understanding a scene, the human visual system will experience three stages. These stages include visual saccadic activating salient regions, visual foraging selecting regions of interest, and visual fixation gazing on local regions to make the final decision.

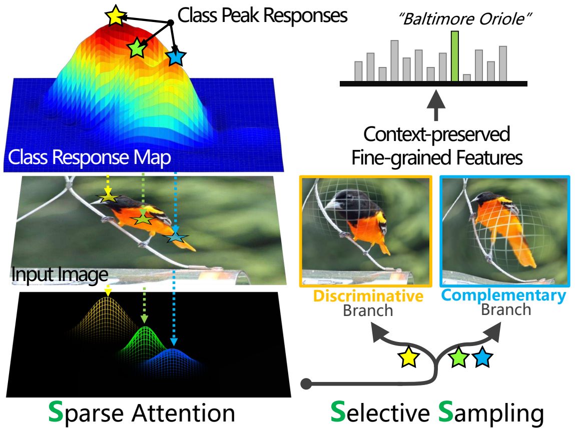

Figure 1. Selective Sparse Sampling Networks (S3Ns) learn sparse attention from class peak responses, which typically corresponds to informative object parts. The sparse attention selectively samples image to two branches in a probabilistic manner to produce a rich representation of both discriminative and complementary visual evidence. The white grids covered on images indicate sparse attention sampling for visualization purpose. Best viewed in color.

Inspired by this, a significant line of previous works [31, 7, 13, 6, 30, 20] tackles the fine-grained image recognition problem by two principal components, *i.e.*, discriminative parts localization and ROI[2] feature extraction. However, the main drawbacks are threefold: 1) Accurate estimation for part bounding boxes under the image-level supervision remains an open problem and often resorts to complicated and time-consuming pipelines, *e.g.*, weakly supervised detection model [33, 29], recurrent mining [3], or reinforcement learning [12]. 2) The number of localized parts is often a predefined hyper-parameter, which is fixed (*e.g.*, four parts [29]) and does not adapt to the image content. 3) Last but not least, the "hard" crop operation neglects the

[‡] Corresponding authors.
[*] Equal contribution.
[1] Source code is released at https://github.com/Yao-DD/S3N.git

[2] The region of interest which is typically defined by a bounding box.

surrounding context of each local region which limits the expressive power of the resulting features, especially when localization errors occur.

In this paper, we address the problem of fine-grained recognition by proposing the Selective Sparse Sampling framework, Fig 1. Our approach imitates the human visual system to predict a dynamic set of sparse attention conditioned on the image content. Each attention focuses on an informative region to estimate the appropriate scale and capture fine-detailed visual evidence without losing the context information.

The proposed framework implemented using convolutional neural networks (CNNs) is referred to as S3Ns. S3Ns are trained with image-level supervision, *i.e*., object categories. We first collect class peak responses, *i.e*., local maximums from class response maps [34, 35], as the estimation of informative receptive fields that contain visual cues for the objects of interest. We then estimate the scale for each identified class peak response to form a set of sparse attention. The resulting sparse attended parts are used to selectively sample the image by an inhomogeneous transform to highlight corresponding regions and guide the network to learn both the discriminative and complementary features.

Compared with the conventional approaches, our method is simple yet effective. By exploiting representation learned by CNNs, *i.e*., class peak responses, our approach requires no additional supervision yet can accurately localize informative regions (See more in Sec. 4.1). Moreover, the number of sparse attended parts is dynamic and depends on the image content. Thus the proposed framework is more flexible and can be applied to different domains, *e.g*., bird, aircraft, and cars, without tuning hyperparameters for each specific tasks. Furthermore, S3N highlights the informative regions in a "soft" manner, which facilitates capturing fine-grained features along with preserving the context information, achieving significant performance gain over the baselines, Fig. 2.

The main contributions of this paper include:

- The development of a novel Selective Sparse Sampling framework, which tackles the challenging fine-grained image recognition problem by learning a set of sparse attention to selectively sample informative regions and extract discriminative and complementary features while preserving the context information.
- The implementation of our approach with popular CNNs such as ResNet50 that demonstrates substantial improvement over the baselines concerning model accuracy and the ability of mining visual evidence.
- The comprehensive experiment analysis as well as the new state-of-the-art performance on common fine-grained recognition benchmarks including CUB-200-2011 Birds, FGVC-Aircraft, and Stanford Cars.

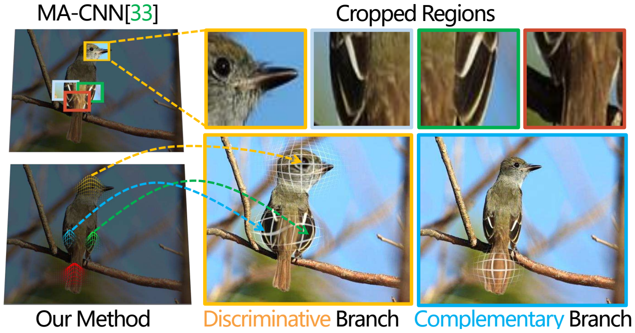

Figure 2. Cropping a fixed number of object parts often involves hyper-parameter and loses context. Our method samples a dynamic number of sparse attention to encode discriminative and complementary fine-grained visual evidence while retaining the surrounding context information.

2. Related Work

In this section, we briefly review previous work from the perspective of feature learning and discriminative region localization.

Fine-grained Feature Learning: Learning representative features is crucial for fine-grained image recognition. The deep features [11, 23, 5] have achieved unprecedented performance for general image recognition tasks, but they are less satisfying for fine-grained image recognition.

In [14], Lin *et al*. argue that the effectiveness of features for fine-grained classification is due to their invariance to position and pose of the object. They propose a bi-linear framework, which is a kind of orderless texture representations and captures localized feature interactions in a translationally invariant manner. Gao *et al*. [4] update the bilinear model into a compact structure, which can reduce the feature dimensionality two orders of magnitude. Kong *et al*. [9] use a bilinear classifier replacing the bilinear feature, which improves the computational efficiency as well as decreases the number of parameters to be learned.

The feature learning approaches mainly focus on exploring invariant features for object representation, but often ignore the spatial distributions of discriminative regions, which limits their performance when facing objects of significant deformation. Our S3N enhances the local features of the sampled sparse attention, which naturally achieves the encoding of spatial information.

Discriminative Region Localization: These approaches typically include two stages: 1) localizing object parts and cropping a fixed number of local regions. 2) extracting features from those restricted parts and aggregating all the features for final recognition.

Numerous earlier studies focus on localizing significant regions for fine-grained recognition with bounding-box and part annotations [24, 31, 13, 6, 30, 20]. Although effective, such supervised annotations are costly to obtain.

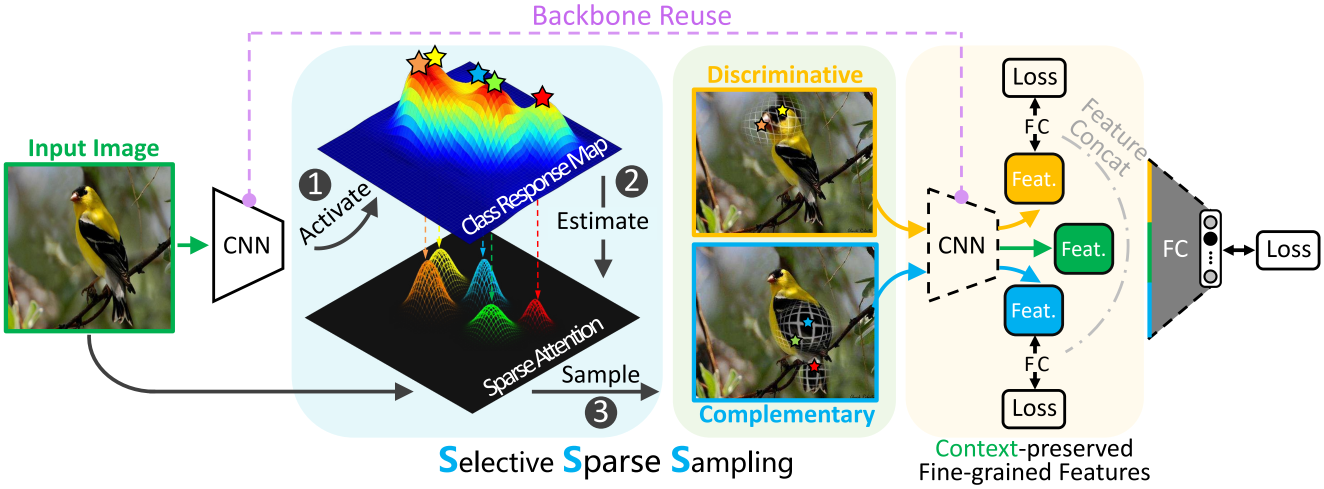

Figure 3. An overview of the proposed Selective Sparse Sampling framework for fine-grained image recognition. The framework first produces sparse attention to localize informative object parts by collecting local maximums of class response maps. Then the following two parallel sampling branches select a dynamic number of sparse attention conditioned on the image content to construct rich representations for both discriminative and complementary visual evidence, respectively. Finally, the features of each branch are aggregated to produce the final output. Best viewed in color.

Consequently, an increasing number of researches explore weakly-supervised methods to estimate discriminative regions. Object detection methods provide the reference to generate proposal regions. For example, Xiao *et al.* [28] and Zhang *et al.* [32] use Selective Search while Yang *et al.* [29] introduce feature pyramid network into fine-grained recognition. Then, some strategies, such as a part discriminator or a particular loss, are designed to filter out informative patches. Zheng *et al.* [33] group channels in convolutional networks to generate different part patterns. Beyond that, attention mechanisms have been applied. Fu *et al.* [3] recursively learn discriminative region at multiple scales (*i.e.* 3). Li *et al.* [17, 22, 12] use a recurrent visual attention model to select a sequence of attention regions. Despite the effectiveness of part-based discriminative approaches, they use a "hard" part crop strategy and overlook the surrounding context around cropped regions, which limit the predictive power of the corresponding features. Instead, we use a soft manner to amplify local regions while retaining context.

Recasens *et al.* [21] first propose to perform non-uniformed sampling on saliency maps. Our approach is different from this work in the following three aspects. Firstly, we propose to sample sparse attention, which is class-aware and enables richer representation than the class-agnostic saliency employed in [21]. Secondly, our sparse attention often corresponds to fine-detailed object parts such as throat, nape, and crown, providing more subtle visual evidence than the salient region of [21]. Thirdly, S3N explicitly divides the visual evidence into two parallel sampling branches, *i.e.*, the discriminative and complementary branch, while the model in [21] samples it together.

3. Methodology

The proposed Selective Sparse Sampling framework first learns a set of sparse attention, which specifies the location and scale for candidate regions particularly informative to the task. The framework then uses the learned sparse attention to selectively sample the input image into discriminative and complementary branches for extracting context-preserved fine-detailed features. The framework is implemented using a convolutional neural network (CNN) backbone, *e.g.*, ResNet50, and can be trained end-to-end with standard classification settings, *i.e.*, image-level supervision, and cross-entropy loss, Fig. 3.

3.1. Revisiting Class Peak Response

Given an input image X, our approach predicts a set of sparse attention by leveraging the class peak responses, *i.e.*, local maximums, emerged from class response maps of the classification networks trained with image-level category supervision [34, 35].

We first feed the image X into the CNN backbone and extract feature maps from the top convolutional layer. The resulting feature maps are indicated as $S \in \mathbb{R}^{D\times H\times W}$, where D is the number of feature channels and $H \times W$ is the spatial size of the feature map. The feature maps S are then fed to a Global Average Pooling (GAP) layer followed by a Fully Connected (FC) layer to obtain class scores $s \in \mathbb{R}^C$, where C is the number of fine-grained object categories. With the weight matrix of the FC layer $W^{fc} \in \mathbb{R}^{D\times C}$, we can compute the class response map M_c as

$$M_c = \sum_{d=1}^{D} W_{d,c}^{fc} \times S_d. \tag{1}$$

Class peak responses for the c-th category are defined to be the local maximums within a window size of r from the corresponding class response map M_c. And the peak locations are denoted as $P_c = \{(x_0, y_0), (x_1, y_1), ..., (x_{N_c}, y_{N_c})\}$, where N_c is the number of valid peaks for the c-th class. Class peak responses typically correspond to strong visual cues residing inside regions of interest [35].

3.2. Learning Sparse Attention

We leverage the learned peaks to localize receptive fields particularly informative to the task and estimate a set of sparse attention for extracting fine-grained visual evidence.

To keep training and testing phases consistent, we use the predicted class scores s to select candidate peaks for both learning and inference periods. Based on the experimental observation, peaks in the top-1 class response map are not always enough to cover discriminative parts. However, peaks in top-k are quite a few but may be noisy. To balance the recall and precision of visual evidence, we selectively collect peaks from the top-1 or top-5 predicted classes.

Let $Prob = softmax(s) \in \mathbb{R}^C$ be the predicted probabilities of all $\mathbb{C}$ classes and $\hat{Prob} \in \mathbb{R}^5$ be the subset of $Prob$ for top-5 class scores, listed in the descending order. We compute the entropy as

$$H = -\sum_{i=1}^{5} p_i \log p_i, \; p_i \in \hat{Prob}, \tag{2}$$

and construct a response map R based on the following strategy,

$$R = \begin{cases} \hat{M}_1, & if\ H \leq \delta \\ \sum_{k=1}^{5} \hat{M}_k, & if\ H > \delta \end{cases}, \tag{3}$$

where $\hat{M} \in \mathbb{R}^{5 \times H \times W}$ is the class response maps correspond to $\hat{Prob}$ and δ is a threshold[3].

We then map R into $[0, 1]$ by Min-Max Normalize, *i.e.*, $R = \frac{R - min(R)}{max(R) - min(R)}$. Finally, we find all local maximums within a window size of r[4] in R and denote the locations of them as $T = \{(x_1, y_1), (x_2, y_2), ..., (x_{N_t}, y_{N_t})\}$, where N_t is the number of detected peaks.

Note that in the above peak selection strategy, we use entropy to determine the confidence of network predictions. When the confidence is high, we use peaks from the top-1 class response map, and when low, we aggregate all top-5 class response maps for peak finding to improve the recall of informative region candidates.

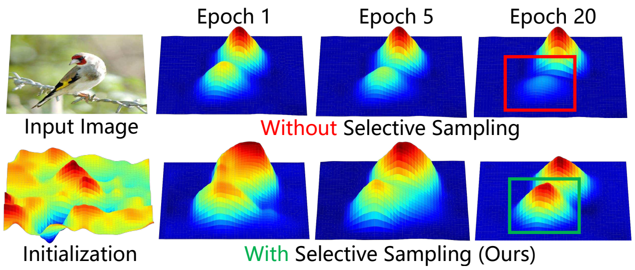

Figure 4. The discovered visual evidence at different training epoch. Model learning can be dominated by strong features and fail to preserve subtle features. Our method selectively samples and balances discriminative and complementary features to encourage the model to learn a more diverse image representation. Best viewed in color.

For each peak $(x, y) \in T$ detected by the above procedure, we generate random number $\zeta_{(x,y)}$ from the uniform distribution between 0 and 1. We then partition peaks into two sets, T_d and T_c, according to their response values,

$$\begin{aligned} T_d &= \{(x, y) |\, (x, y) \in T\ if R_{x,y} \geqslant \zeta\ \} \\ T_c &= \{(x, y) |\, (x, y) \in T\ if R_{x,y} < \zeta\ \}\,. \end{aligned} \tag{4}$$

Peaks of high response value which localize discriminative evidence (*e.g.*, unique patterns for the fine-grained categories) are more likely to be partitioned into T_d, while peaks of low response value which localize complementary evidence (*e.g.*, supporting patterns) are more likely to be partitioned into T_c.

Finally, we utilize Gaussian kernel to compute a set of sparse attention $A \in \mathbb{R}^{N_t \times H \times W}$ attending to each peak as

$$A_{i,x,y} = \begin{cases} R_{x_i,y_i} e^{-\frac{(x-x_i)^2 + (y-y_i)^2}{R_{x_i,y_i}\ \beta_1^2}}, & if\ (x_i, y_i) \in T_d \\ \frac{1}{R_{x_i,y_i}} e^{-\frac{(x-x_i)^2 + (y-y_i)^2}{R_{x_i,y_i}\ \beta_2^2}}, & if\ (x_i, y_i) \in T_c, \end{cases} \tag{5}$$

where β_1 and β_2 are learnable parameters and R_{x_i,y_i} is the peak value of the i-th peak in T. Note that the amplitude and radius of each sparse attention control sampling scale of the corresponding visual evidence (Sec. 3.3), and are dynamically affected by the corresponding peak response value; thus conditioned on the input image content.

3.3. Selective Sampling

With the sparse attention defined in Eq. 5, we perform image re-sampling to highlight fine-grained details from informative local regions while preserving surrounding context information. We construct two sampling maps Q^d and Q^c for the discriminative branch and complementary branch of feature extraction as

$$\begin{aligned} Q^d &= \sum A_i,\ if\ (x_i, y_i) \in T_d \\ Q^c &= \sum A_i,\ if\ (x_i, y_i) \in T_c. \end{aligned} \tag{6}$$

[3] We set $\delta = 0.2$ in all experiments. The model accuracy is insensitive to δ when δ falls into a certain range, *i.e.* [0.1, 0.3].

[4] We set $r = 3$ in all experiments, empirically.

Denote an input image X as a mesh grid with vertices V, where $\mathbf{V} = [\mathbf{v_0}, \mathbf{v_1}, ...\mathbf{v_{end}}]$ and $\mathbf{v_i} = (v_x^i, v_y^i) \in \mathbb{R}^2$. The vertices can be connected into horizontal and vertical grid lines.

The sampling procedure targets at exploring a new mesh geometry $\mathbf{V}' = \left[\mathbf{v}'_\mathbf{0}, \mathbf{v}'_\mathbf{1}, ...\mathbf{v}'_\mathbf{end}\right]$, where regions of higher significance enjoy uniform scaling and those of lower significance are allowed to be suppressed to a large extent. This problem can be converted to find a mapping between the re-sampling image and the input image. Such a mapping can be written as two functions, $f(\mathbf{v})$ and $g(\mathbf{v})$, so that $X_{new}(\mathbf{v}) = X(f(\mathbf{v}), g(\mathbf{v}))$, where X_{new} denotes the re-sampled image.

The goal of designing f and g is to map pixels proportionally to the normalized weight assigned to them by the sampling map. An exact approximation to this problem is that f and g can satisfy the condition: $\int_0^{f(\mathbf{v})} \int_0^{g(\mathbf{v})} Q(\mathbf{v}')dv_x'dv_y' = v_x v_y$. Following the method in [21], the solution can be described as

$$f(\mathbf{v}) = \frac{\sum_{\mathbf{v}'} Q(\mathbf{v}')k(\mathbf{v}', \mathbf{v})v_x'}{\sum_{\mathbf{v}'} Q(\mathbf{v}')k(\mathbf{v}', \mathbf{v})}, \tag{7}$$

$$g(\mathbf{v}) = \frac{\sum_{\mathbf{v}'} Q(\mathbf{v}')k(\mathbf{v}', \mathbf{v})v_y'}{\sum_{\mathbf{v}'} Q(\mathbf{v}')k(\mathbf{v}', \mathbf{v})}, \tag{8}$$

where $k(\mathbf{v}', \mathbf{v})$ is a Gaussian distance kernel to act as a regularizer and avoid extreme cases, such as all the pixels converge to the same value. By substituting Q in Eq. 7 and in Eq. 8 with Q^d and Q^c that are computed in Eq. 6, we can get two re-sampled images. One corresponding to Q^d, named as the discriminative branch image, highlights significant regions for extraction of detailed evidence. The other one corresponding to Q^c, named as the complementary branch image, enlarges regions that are not that significant for mining more visual cues. As shown in Fig. 4 that the proposed selective sampling can prevent powerful features from dominating the gradient learning and encourage the network to learn a more diverse image representation. The re-sampling process is implemented by convolution and embedded into the end-to-end training, where β_1 and β_2 can be updated by the classification loss from the re-sampled images.

3.4. Fine-grained Feature Learning

With the sparse attention and selective sampling procedure defined above, the feature learning procedure is implemented in an end-to-end manner. During the process, an image X is first fed to S3N and generates two re-sampled images, the same size as the input image. They amplify a dynamic number of informative regions corresponding to discriminative and complementary features. The two re-sampled images are then taken as inputs by S3N for extracting fine-grained features. The same backbone is reused for feature extraction of all inputs; thus, no significant model parameter is introduced by our proposed method.

Benefit from assembling global and local informative features of an image, we define the feature representations for each image: $F_J = \{F_O, F_D, F_C\}$, where F_O, F_D, F_C denotes the feature extracted from the original image, the discriminative branch image, and the complementary branch image, respectively. These features are concatenated and fed to a fully-connection fusion layer with a softmax function for the final classification.

During learning, the whole model is optimized by classification losses defined as

$$L(X) = \sum_{i \in I} L_{cls}(\mathbf{Y}^i, \mathbf{Y}^*) + L_{cls}(\mathbf{Y}^j, \mathbf{Y}^*) \tag{9}$$

where L_{cls} denotes the cross-entropy loss. I is $\{O, D, C\}$. $\mathbf{Y}^i$ is the predicted label vector from original and re-sampling images based on features F_O, F_D, and F_C. $\mathbf{Y}^j$ is the predicted label vector using joint features F_J and $\mathbf{Y}^*$ is the ground-truth label vector.

3.5. Discussion

The proposed S3N leverages the class peak responses learned by image classification networks to estimate informative regions for the task, *i.e.*, fine-grained image recognition, which guide a selective sampling procedure to highlight fine-detailed visual evidence without losing surrounding context information. The re-sampled images are then fed into the shared network backbone to update the learned class peak responses. By multiple epochs of peak prediction and image re-sampling, S3N implements a special kind of iterative learning.

The S3N incorporates the sparse attention mechanisms with image content re-sampling in an integrated framework, which provides a new methodology to fuse local and global features. In the first step, the global image features are used to activate class peak responses. In the second step, the activated peaks reinforce the image content and global features. Therefore, local and global features reinforce mutually.

4. Experiments

Datasets: We evaluate the proposed approach on three fine-grained datasets. The CUB-200-2011 [24] dataset consists of 11,788 images from 200 bird species, which are split into 5,994 training and 5794 testing images. The FGVC-Aircraft [15] dataset contains 10,000 images of 100 aircraft variants, among which 6,667 images for training and 3,333 images for testing. The Stanford Cars [10] dataset includes 16,185 images of 196 classes of cars, where 8144 images are for training, and 8044 images are for testing. Our proposed method does not utilize any extra annotations (*e.g.*, part annotations [30], object bounding boxes [31], and prior

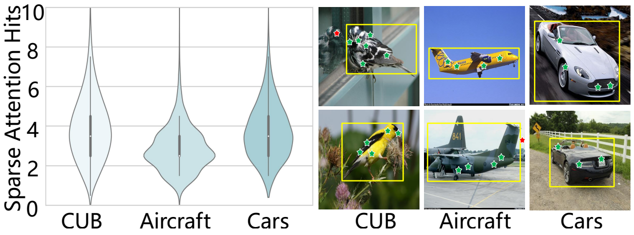

Figure 5. Distribution of the per-image hit number of the sampled sparse attention on three datasets. Our method collects a different number of sparse attention for different images, rather than sampling a fixed number of parts for all the images. Best viewed zooming on screen.

Dataset	CUB	Aircraft	Cars
Mean hit rate (%)	94.63	97.22	98.76

Table 1. Mean hit rate (%) of the sparse attention sampled by S3N on three datasets.

knowledge of class relationships [25, 2]) except image-level labels throughout the experiments.

Implementation Details: In all our experiments, we preprocess images to size 448 × 448. We use Resnet-50 as feature extractor. We train S3Ns for 60 epochs with a batch size of 16 using the Momentum SGD optimizer. We set the weight decay as 1e-4 and momentum as 0.9. For parameters that are initialized from pre-trained models on Imagenet, we use an initial learning rate of 0.001; for other parameters, we use an initial learning rate of 0.01.

4.1. Sparse Attention Analysis

To analyze the quality of the sparse attention produced by S3N, we perform a series of statistical analysis about the ability of the sparse attention to localize informative object parts. We first calculate the hit number of the sparse attention. A hit is counted if the pixel of the sparse attention falls in the ground truth bounding boxes of the image. Otherwise, a miss is counted. We visualize the distribution of the hit number of each dataset to see whether S3N networks can collect informative object parts as visual evidence, Fig. 5. It can be seen that our method can adaptively select informative, sparse attention for each image, rather than a fixed number of object parts as hyper-parameters. We measure the per-image hit rate by $\frac{Hits}{Hits+Misses}$. The overall results for each dataset are the mean value of per-image sparse attention hit rate, shown in Tab. 1, demonstrating that our collected sparse attention often hit valid object parts.

We further select the maximum and the minimum sparse attention of each image and count the numbers of max/min sparse attention for each object part category. It can be seen in Fig. 6 that the **D**iscriminative branch of S3N usually sample informative object parts like “throat”, “nape”, and “crown” as strong evidence to give a major decision of fine-grained bird categories. As for the **C**omplementary branch, object parts like “wings” and “legs” are sampled as weak evidence to provide supportive information.

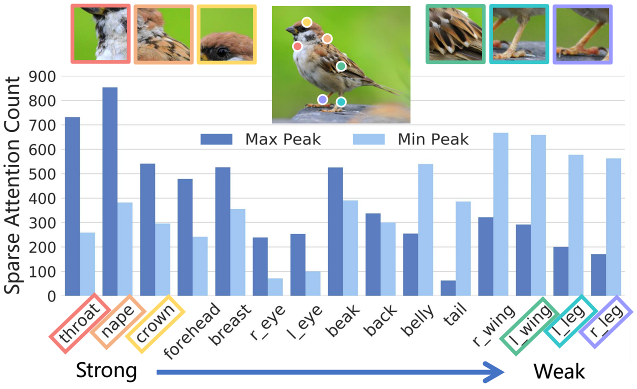

Figure 6. Statistics of maximum and minimum sparse attention in terms of object part classes on the CUB dataset. The discrimination of the part categories is ranked from strong to weak according to the difference between the counts of maximum and the minimum attention. Best viewed in color.

4.2. Ablation Study

In this section, we conduct a series of ablation studies to understand the designation of our proposed S3N on the CUB-200-2011 dataset.

Impact of different sampling branch: To investigate the contribution of each branch in the proposed method, we omit different sparse attention sampling branches of S3Ns. In Tab. 2, we report the total top-1 classification accuracy for each architecture as well as per-branch classification accuracy. Moreover, we also report the top-1 localization accuracy defined by [34]. The localization is estimated based on the aggregated top-1 predicted class response maps from the O, D, and C branches; class response maps from the D, C branches are mapped to the same space of the O branch by their corresponding inverse transform.

From the results reported in Tab. 2, we can draw the following conclusions: 1). The top-1 localization accuracy (2nd column) drops from 65.2% to 56.6% and 59.2% when omitting the discriminative branch and the complementary branch respectively, which demonstrates the effectiveness of both the branches to localize informative objects regions. 2). The classification accuracy at the complementary branch (5th column) is lower than other branches (3rd and 4th column) in all the architectures, showing that the visual evidence learned by this branch is less discriminative. 3). The network with only discriminative branch (S3N O+D) improves the total classification accuracy (last column) of baseline by 1.6% (86.0% vs. 87.6%) shows that our method can collect more strong visual evidence for the fine-grained image recognition task, while the network with complementary branch (S3N O+C) improves 1.3% (86.0% vs. 87.3%) shows that our method can capture weak vi-

Setting	Loc.	O-branch	D-branch	C-branch	Total
S3N O	57.7	86.0	-	-	86.0
S3N O+D	59.2	87.0	86.5	-	87.6
S3N O+C	56.6	86.8	-	85.3	87.3
S3N D+C	62.6	-	87.1	85.6	87.5
S3N O+D+C	**65.2**	87.9	86.7	85.1	**88.5**

Table 2. Ablation study on the CUB dataset based on different branches of S3N. The first column shows top-1 accuracy (%) of object localization and the subsequent columns show top-1 accuracy of image classification for the **O**riginal, **D**iscriminative, **C**omplementary branches, and the total aggregated accuracy for the whole S3N.

	Acc (%)	Comments
Saliency based attention [21]	85.9	Class-agnostic
Class response based attention	87.8	Class-aware
Sparse attention (Ours)	**88.5**	Part-aware

Table 3. Top-1 classification accuracy (%) on the CUB dataset for sparse attention and dense attention, namely, saliency based attention and class response based attention.

Partition	All	Random	Probability (Ours)
Acc (%)	87.7	87.9	**88.5**

Table 4. Comparison to different selective sampling strategies for discriminative and complementary branch.

sual evidence which is relatively easier to be ignored. 4). Note that the absence of the discriminative branch causes more loss (1.2%, from 88.5% to 87.3%) of total classification accuracy than that of the complementary branch (0.9%, from 88.5% to 87.6%), demonstrating that the discriminative branch can help collect critical visual evidence which is necessary for fine-grained image recognition.

Sparse attention vs. dense attention: In Tab. 3, we compare our sparse attention with two types of dense attention, including the class-agnostic saliency-based attention and the class response based attention. Our sparse attention can explicitly localize class-aware object parts and informative parts. The sparse attention sampling can enhance the informative object regions and discard noisy responses, making it easier to capture subtle visual evidence which is often ignored in dense attention maps.

Sparse attention partition: In our method, the sparse attention would be divided into two sets in a probabilistic manner. Sparse attention with higher scores are more likely to be sampled at the discriminative branch, and those with lower scores are more likely to be sampled at the complementary branch. We compare our probabilistic partition with another two reasonable sparse attention partition strategies: 1) **All**. Both the discriminative branch and the complementary branch sampling all the sparse attention. 2) **Random**. The two branches randomly select in the whole sparse attention set; thus, they use different sparse attention. As is shown in Tab. 4, our probabilistic partition performs well since we sample strong and weak visual evidence separately at two branches to enhance the strong evidence and maintain the weak evidence at the same time.

Method	Backbone	CUB	Aircraft	Cars
B-CNN [14]	VGG-16	84.1	84.1	91.3
Low-rank B-CNN [9]		84.2	87.3	90.9
HIHCA [1]		85.3	88.3	91.7
Boosted CNN [18]		85.6	88.5	92.1
RA-CNN [3]	VGG-19	85.3	-	92.5
MA-CNN [33]		86.5	89.9	92.8
DT-RAM [12]	ResNet-50	86.0	-	93.1
FT ResNet [5]		86.0	89.9	92.6
DPL-CNN [26]		87.1	-	93.1
DFL-CNN [27]		87.4	91.7	93.1
NTS [29]		87.5	91.4	93.9
S3N (Ours)		**88.5**	**92.8**	**94.7**

Table 5. Comparison of our approach to recent results on CUB-200-2011, FGVC-Aircraft and Stanford Cars.

4.3. Fine-grained Image Classification

In this section, we compare the performance of the proposed S3N with existing methods on three popular fine-grained image recognition datasets.

Numerical results: The fine-grained image classification is evaluated by the top-1 classification accuracy (%). As shown in Tab. 5, our model significantly outperforms the ResNet-50 baseline (FT ResNet) by 2.5% (86.0% vs. 88.5%), 2.9% (89.9% vs. 92.8%) and 2.1% (92.6% vs. 94.7%) on three challenging dataset respectively, which shows the ability of our S3N to learn rich representation for fine-grained image recognition. Our S3N also outperforms the state-of-the-art by a margin of 1% (87.5% vs. 88.5%), 1.1% (91.7% vs. 92.8%), 0.8% (93.9% vs. 94.7%) on the three datasets, which are significant for the current task. This further validates the benefit of sampling a flexible set of sparse attention for each image over fixed hyper-parameters of the part number by cropping parts.

Qualitative results: In Fig. 7, we visualize the intermediate outputs of our method and the ResNet-50 baseline to interpret why and how our approach can give the correct predictions when the baseline fails. As shown in the first row, our S3N correctly predicts the class *Horned Grebe* while the baseline mistakes it for the similar category *Eared Grebe*. The key to distinguishing the two fine-grained spicies is that the *Eared Grebe* has a fan-shaped splay of golden head feathers whereas *Horned Grebe* has golden tufts that run straight back across the head. Although correctly localizing the “bird head” as discriminative evidence, the baseline failed because of the disability to encode the rich information lying in the “head feather”. In contrast, S3N enhances the visual evidence of “bird head” at the discriminative branch, encoding the information of “head feather”. In the second row of Fig. 7, when using “bird head” as visual evidence, baseline failed to identify

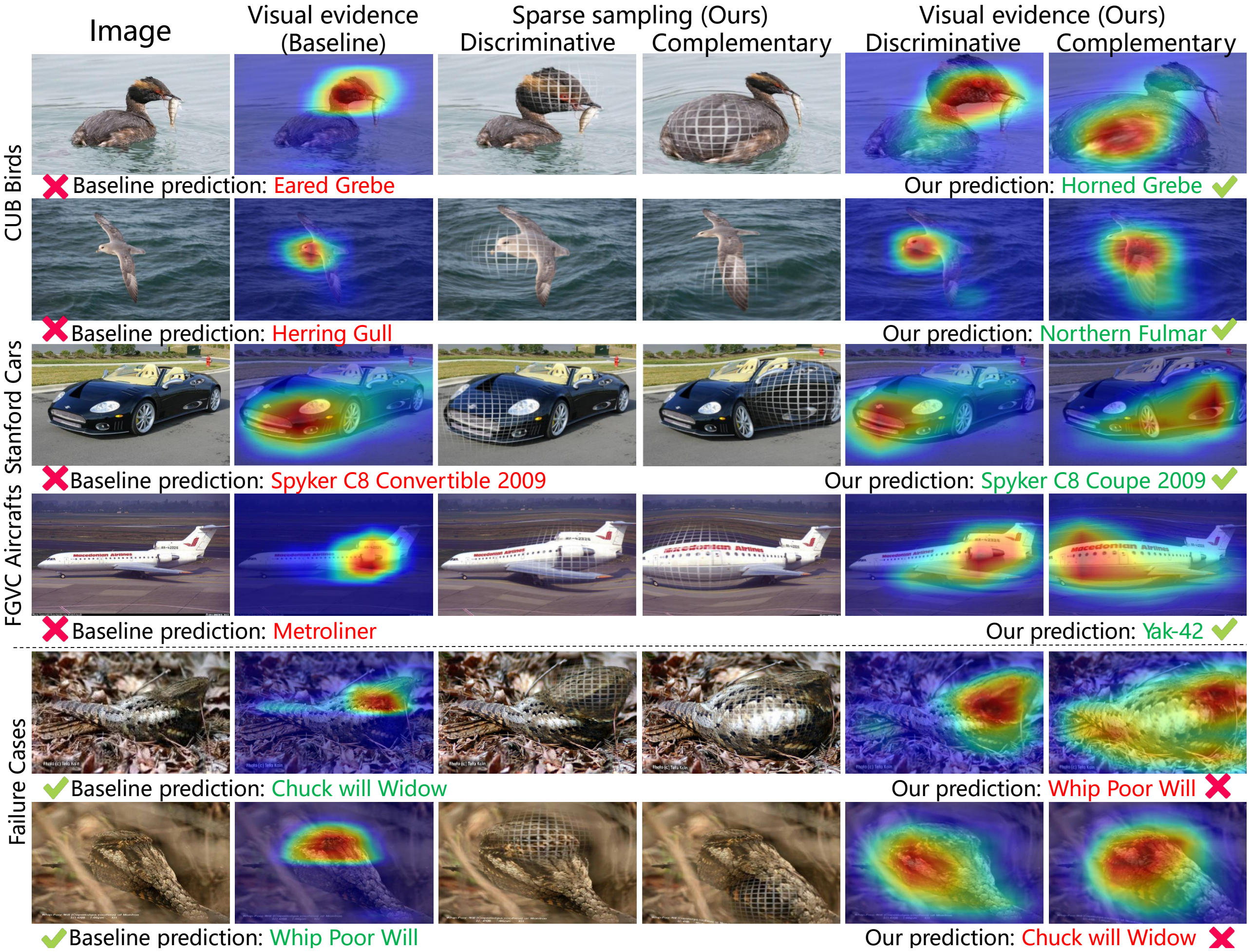
Image
Visual evidence (Baseline)
Sparse sampling (Ours)
Discriminative
Complementary
Visual evidence (Ours)
Discriminative
Complementary
CUB Birds
Stanford Cars
FGVC Aircrafts
Failure Cases
✗ Baseline prediction: Eared Grebe
Our prediction: Horned Grebe ✓
✗ Baseline prediction: Herring Gull
Our prediction: Northern Fulmar ✓
✗ Baseline prediction: Spyker C8 Convertible 2009
Our prediction: Spyker C8 Coupe 2009 ✓
✗ Baseline prediction: Metroliner
Our prediction: Yak-42 ✓
✓ Baseline prediction: Chuck will Widow
Our prediction: Whip Poor Will ✗
✓ Baseline prediction: Whip Poor Will
Our prediction: Chuck will Widow ✗

Figure 7. Visualization of the sparse sampling and the learned visual evidence of our method. Our method can not only enhance the discriminative visual evidence but also explore the complementary visual evidence, *e.g.*, "body" and "wing" for birds, "headstock" and "door" for cars, as well as "nose" and "tail" for aircraft. Our method may make mistakes when texture in objects are particularly similar to background and species under the same genus, i.e. images shown in the last two rows. Best viewed in color.

Northen Fulmar. Our S3N gives right label combining the visual evidence of "bird wing" provided by the complementary branch, since *Herring Gull* has more contrasting plumage than *Northen Fulmar*. Our method's ability to capture subtle inter-class differences generalizes well across domains. As can be seen from the third and fourth rows in Fig. 7, our method can discover more fined-detailed visual evidence for aircraft and cars when the baseline only notices the most significant ones. As for the failure cases, we infer the reason is that features learned by the complementary branch are useless or even harmful when object are similar to background and other species.

5. Conclusions

In this paper, we propose Selective Sparse Sampling, a simple yet effective framework targeting at addressing the challenge of fine-grained image recognition. The framework is implemented with convolutional neural networks, referred to as Selective Sparse Sampling Networks (S3Ns). With image-level supervision, S3Ns estimate sparse attention and implements spatial and semantic sampling. In this way, it selectively aggregates fine-detailed visual evidence from a dynamic number of informative regions conditioned on the image content and with surrounding context. S3Ns consistently improve the baselines and yield superior performance over the state-of-the-art on multiple popular fine-grained recognition benchmarks. The underlying reality is that the Selective Sparse Sampling is consistent with the mechanisms of the human visual system, which provides fresh insights for the field of image recognition.

Acknowledgments. The authors are very grateful to the support by NSFC grant 61836012, 61771447, and 61671427, and Beijing Municipal Science and Technology Commission grant Z181100008918014.

References

[1] Sijia Cai, Wangmeng Zuo, and Lei Zhang. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *IEEE International Conference on Computer Vision*, pages 511–520, 2017.

[2] Tianshui Chen, Wenxi Wu, Yuefang Gao, Le Dong, Xiaonan Luo, and Liang Lin. Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. *arXiv preprint arXiv:1808.04505*, 2018.

[3] Jianlong Fu, Heliang Zheng, and Tao 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*, pages 4476–4484, 2017.

[4] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–326, 2016.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[6] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked CNN for fine-grained visual categorization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1173–1182, 2016.

[7] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *neural information processing systems*, pages 2017–2025, 2015.

[8] . I. Jhannesson, I. M. Thornton, I. J. Smith, A Chetverikov, and Kristjnsson. Visual foraging with fingers and eye gaze. *i-Perception*, 7(2):2041669516637279, 2016.

[9] Shu Kong and Charless C Fowlkes. Low-rank bilinear pooling for fine-grained classification. *computer vision and pattern recognition*, pages 7025–7034, 2017.

[10] Jonathan Krause, Michael Stark, Jia Deng, and Fei Fei Li. 3d object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops*, 2013.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *International Conference on Neural Information Processing Systems*, 2012.

[12] Zhichao Li, Yi Yang, Xiao Liu, Feng Zhou, Shilei Wen, and Wei Xu. Dynamic computational time for visual attention. *international conference on computer vision*, pages 1199–1209, 2017.

[13] Di Lin, Xiaoyong Shen, Cewu Lu, and Jiaya Jia. Deep LAC: deep localization, alignment and classification for fine-grained recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1666–1674, 2015.

[14] Tsungyu Lin, Aruni Roychowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. *international conference on computer vision*, pages 1449–1457, 2015.

[15] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *HAL - INRIA*, 2013.

[16] M. B. Mirza, R. A. Adams, C Mathys, and K. J. Friston. Human visual exploration reduces uncertainty about the sensed world. *Plos One*, 13(1):e0190429, 2018.

[17] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2204–2212, 2014.

[18] Mohammad Moghimi, Serge J. Belongie, Mohammad J. Saberian, Jian Yang, Nuno Vasconcelos, and Li-Jia Li. Boosted convolutional neural networks. In *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*, 2016.

[19] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008*, pages 722–729, 2008.

[20] Guo Pei and Ryan Farrell. Fine-grained visual categorization using pairs: Pose and appearance integration for recognizing subcategories. *CoRR*, 2018.

[21] Adria Recasens, Petr Kellnhofer, Simon Stent, Wojciech Matusik, and Antonio Torralba. Learning to zoom: A saliency-based sampling layer for neural networks. *european conference on computer vision*, pages 51–66, 2018.

[22] Pierre Sermanet, Andrea Frome, and Esteban Real. Attention for fine-grained categorization. *arXiv: Computer Vision and Pattern Recognition*, 2014.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *international conference on learning representations*, 2015.

[24] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[25] Dequan Wang, Zhiqiang Shen, Jie Shao, Wei Zhang, Xiangyang Xue, and Zheng Zhang. Multiple granularity descriptors for fine-grained categorization. In *ICCV 2015*, pages 2399–2406, 2015.

[26] Yaming Wang, Vlad I Morariu, and Larry S Davis. Weakly-supervised discriminative patch learning via cnn for fine-grained recognition. *arXiv: Computer Vision and Pattern Recognition*, 2016.

[27] Yaming Wang, Vlad I Morariu, and Larry S Davis. Learning a discriminative filter bank within a cnn for fine-grained recognition. *computer vision and pattern recognition*, pages 4148–4157, 2018.

[28] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *IEEE Conference*

on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pages 842–850, 2015.

[29] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei 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*, pages 438–454, 2018.

[30] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed M. Elgammal, and Dimitris 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*, pages 1143–1152, 2016.

[31] Ning Zhang, Jeff Donahue, Ross B. Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 834–849, 2014.

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

[33] Heliang Zheng, Jianlong Fu, Mei Tao, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *IEEE International Conference on Computer Vision*, 2017.

[34] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2921–2929, 2016.

[35] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *CVPR*, 2018.