Weakly Supervised Object Localization Through Inter-class Feature Similarity and Intra-class Appearance Consistency

Jun Wei[1,2,3], Sheng Wang[6], S. Kevin Zhou[1,4,5], Shuguang Cui[1,2,3], and Zhen Li[1,2,3(✉)]

[1] The Future Network of Intelligence Institute, School of Science and Engineering, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong (Shenzhen), Shenzhen, China
junwei@link.cuhk.edu.cn, lizhen@cuhk.edu.cn
[2] Shenzhen Research Institute of Big Data, Shenzhen, China
[3] The Future Network of Intelligence Institute, Shenzhen, China
[4] School of Biomedical Engineering and Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, China
[5] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[6] Shanghai Key Laboratory of Metabolic Remodeling and Health, Institute of Metabolism and Integrative Biology, Fudan University, Shanghai, China

Abstract. Weakly supervised object localization (WSOL) aims at detecting objects through only image-level labels. Class activation maps (CAMs) are the commonly used features for WSOL. However, existing CAM-based methods tend to excessively pursue discriminative features for object recognition and hence ignore the feature similarities among different categories, thereby leading to CAMs incomplete for object localization. In addition, CAMs are sensitive to background noise due to over-dependence on the holistic classification. In this paper, we propose a simple but effective WSOL model (named **ISIC**) through **I**nter-class feature **S**imilarity and **I**ntra-class appearance **C**onsistency. In practice, our ISIC model first proposes the inter-class feature similarity (ICFS) loss against the original cross entropy loss. Such an ICFS loss sufficiently leverages the shared features together with the discriminative features between different categories, which significantly reduces the model over-fitting risk to background noise and brings more complete object masks. Besides, instead of CAMs, a non-negative matrix factorization mask module is applied to extract object masks from multiple intra-class images. Thanks to intra-class appearance consistency, the achieved pseudo masks are more complete and robust. As a result, extensive experiments confirm that our ISIC model achieves state-of-the-art on both CUB-200 and ImageNet-1K benchmarks *i.e.,* **97.3%** and **70.0%** GT-Known localization accuracy, respectively.

Keywords: Weak supervision · Object localization · Inter-class similarity

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13690, pp. 195–210, 2022.
https://doi.org/10.1007/978-3-031-20056-4_12

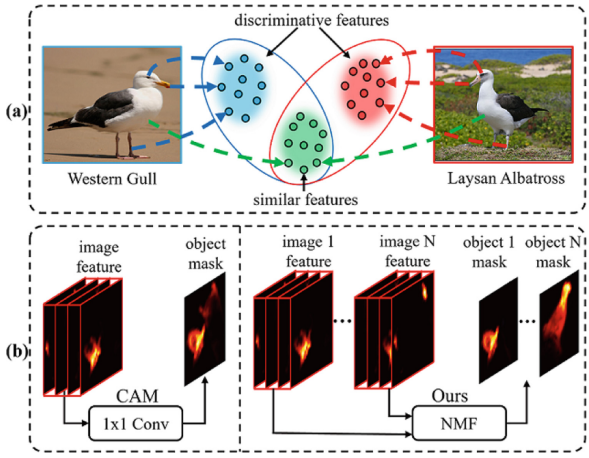

Fig. 1. (a) Visualization of discriminative (eye, beak and foot) and similar (body and feather) features between two categories. (b) The comparison between the CAM-based method and non-negative matrix factorization (NMF)-based one. NMF utilizes features of multiple images of the same category to assist the object mask prediction, while CAM performs on the single image features and highly relies on the classification layer.

1 Introduction

Thanks to the breakthrough of deep learning, recent years have witnessed great progress in object detection [7]. However, training a high-performance object detector requires massive bounding box annotations. These annotations are expensive and sometimes unavailable. To alleviate the model's thirst for annotations, weakly-supervised object localization (WSOL) has gained lots of attentions [1–4,9,12,15,16,24,27,28,31,32,34,35] as it aims at predicting objects' bounding boxes through cheap image-level annotations. Therefore, it largely reduces the annotation cost and is of great practical significance. Previous WSOL methods mainly rely on the class activation maps (CAMs) [35]. However, CAMs tend to cover only small discriminative regions of an object, causing incomplete predictions. Hence, lots of approaches have been proposed to improve CAMs, such as erasing based methods [3,14,19,26,32], feature refining based methods [4,16,24,27,29] and regression based methods [6,12,31]. All these methods have achieved remarkable localization performance. However, these methods mostly are based on classification models, whose goals are inconsistent with object localization due to the following two defects.

First, the similarity between classes has been ignored. Previous classification models [8,18,20] usually adopt cross entropy loss for model training, thus finding the discriminative features of each category. However, for WSOL, focusing too much on differences between images will lead to incomplete predictions. Because images from different categories might share highly similar features. Forcing the model to distinguish between them causes the model to focus only on the most

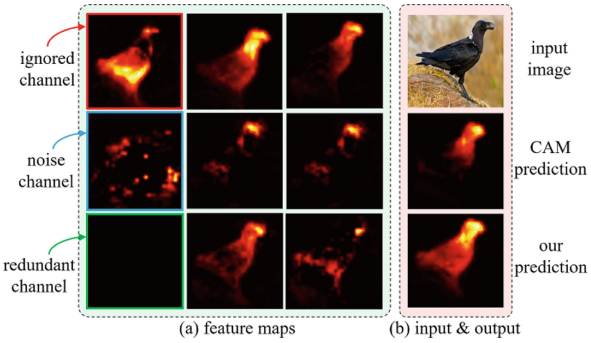

Fig. 2. Visualization of feature maps and predictions. (a) shows some feature maps before the classification layer, where red rectangle in the first row shows the unused features by CAM. (b) compares the predictions between CAM and our model. (Color figure online)

discriminative object regions. To better elaborate this statement, we present an illustrative example in Fig. 1(a). Two images come from different categories and each of them could be regarded as a bag of features, such as beak, eye, body, feather and foot. Among these two images, beak, eye and foot show different colors or shapes, as shown in blue and red areas. These features are usually regarded as the discriminative features and will be extracted to assist the classification model in making decisions. However, there also exist some similar features in the two images, namely body and feather, as shown in the green area. Overemphasizing the differences between images results in these similar features being ignored, which leads to incomplete predictions for WSOL. Besides, due to the classification supervision, even though there are no discriminative features in the image, the model still is forced to learn the differences between images, thus overfitting to the background noise.

Second, as shown in Fig. 1(b), the generation of CAMs is highly dependent on the final classifier (*i.e.*, the last fully connected layer), suffering from false positives (*i.e.*, noise) and false negatives (*i.e.*, content missing) in the final predictions. Because CAM-based methods [32,35] generally adopt parameters of the classification layer as the coefficients to combine feature maps for final prediction. However, these parameters are optimized for classification, where only the most discriminative feature maps are selected out for combination and the rest maps are ignored. But we argue that those overlooked feature maps actually contain helpful information for WSOL. To explain it, we visualize some feature maps extracted from the CAM-based model, as shown in Fig. 2. The body region (red rectangle) of the bird has been activated in feature maps. But in the final prediction, the body region is ignored and only the head region stands out. Namely, CAM-based methods could not make full use of the extracted feature maps. Besides, CAMs are generated based on a single image, which is not robust to background noise. As shown in the blue rectangle in Fig. 2, many background areas are also activated, which interferes with the model's predictions.

To address the above concerns, we propose the **I**nter-class feature **s**imilarity and **I**ntra-class appearance **C**onsistency (named ISIC) model, which improves WSOL from two aspects: supervision and object mask generation. For supervision, we introduce the inter-class feature similarity (ICFS) loss to supplement the widely used cross entropy (CE) loss. In practice, CE loss focuses on the discriminative features of each category (*i.e.*, red and blue areas in Fig. 1(a)), and ICFS loss focuses on the similarities between different categories (green area in Fig. 1(a)). These two losses work against each other and eventually reach an equilibrium. Therefore, ISIC can better balance the localization task and the classification task and is not easy to overfit to the background noise, resulting in more complete predictions.

For object mask generation, instead of relying on the classification layer, we apply the non-negative matrix factorization module (NMFM) to obtain the object mask. NMFM is based on features of multiple images from the same category, which achieves the object mask by extracting the commonness of these images, as shown in Fig. 1(b). Compared with previous methods, NMFM does not rely on the high-level classification layer, so it will not ignore the body region in Fig. 2 and fully exploit all the feature maps. Besides, NMFM is based on multiple images, which is more robust to background noise than that based on a single image. After getting the predicted mask, we follow [24] to train a class-agnostic segmentation model to get the final mask and apply a bounding box extractor to obtain the final object localization. In summary, our contributions fall into three parts:

- Opposite to classification, we propose the ICFS loss to constrain and maintain the similarity between classes. Such ICFS loss can largely reduce the model risk of over-optimizing the discriminative features, thus more complete regions of the object can be activated.
- We propose to replace the original CAMs with non-negative matrix factorization for object mask generation, which avoids the over-discriminative effect of the classification layer and the background noise.
- With negligible computational cost overheads, our proposed methods achieve consistent and substantial gains, *i.e.*, state-of-the-art on both CUB-200–2011 and ImageNet-1K benchmarks for WSOL.

2 Related Works

2.1 Class Activation Maps (CAMs) Based WSOL

Weakly supervised object localization (WSOL) is a challenging task, aiming to localize objects with inexpensive image-level annotations. Zhou *et al.* [35] firstly propose the class activation maps (CAMs) to extract the object location. But restricted by the classification mechanism, CAMs only cover the discriminative object parts. To make CAMs complete, HaS [19] proposes the random erasure of image patches to force the model to mine more object regions. ACoL [32], ADL [3], EIL [14] and AE [26] follow the erasure paradigm and drop the most

discriminative features to reduce the model's dependency on them. CutMix [30] assembles patches from different images to guide model to learn more object parts. These methods greatly improve the quality of CAMs, but have the risk of spreading to the background regions when discriminative features are insufficient.

2.2 Pseudo Label Based WSOL

[6,12,31] take object localization as a regression task. Specifically, GCNet [12] utilizes a detector to regress the object bounding box, and produces the object mask by a generator to maximize the score of the classifier. But the indirect supervision brings unstable predictions. Inspired by GCNet, SLTNet [6] supervises the regressor to learn through the pseudo bounding box generated by a newly designed locator. On the contrary, PSOL [31] divides WSOL into two separate tasks, classification and localization. It applies DDT [25] to produce pseudo bounding boxes from the pre-trained model, which are exploited afterward to train a detector. However, these pseudo labels come from the pre-trained model, which are inexact and lower the upper limit of the detector. Different from pseudo bounding box label, SPOL [24] proposes to generate the pseudo mask to train a lass-agnostic segmentation model and achieves higher performance.

2.3 Attention Based WSOL

SPG [33] adopts a stage-wise manner to refine object mask, which regards high confident object regions as the foreground seeds and uses the self-produced guidance maps to progressively expand these seeds. SPOL [24] focuses on shallow features and proposes a multiplication feature fusion to combine the complementary features of different layers. To capture the long-range feature dependency, TS-CAM [4] proposes to generate the token semantic coupled attention map by visual transformer [22], which extracts both semantics and positioning information. Similarly, SPA [16] proposes the self-correlation to capture long-range structural information of objects. All these methods have achieved great progress in WSOL. However, the similarity between categories has been ignored. In this paper, we explicitly use inter-class similarity to boost WSOL performance.

3 Methodology

3.1 Pipeline

Figure 3 depicts the pipeline of our proposed ISIC model, which consists of two stages (*i.e.*, object mask generation and class-agnostic segmentation). During training, both stages are involved, where the object mask generation stage is used to generate pseudo masks for the input images, and the class-agnostic segmentation stage adopts these pseudo masks as labels to train a binary (*i.e.*, object or no-object) segmentation model. But during inference, only the class-agnostic segmentation stage is involved, where we directly derive the segmentation mask

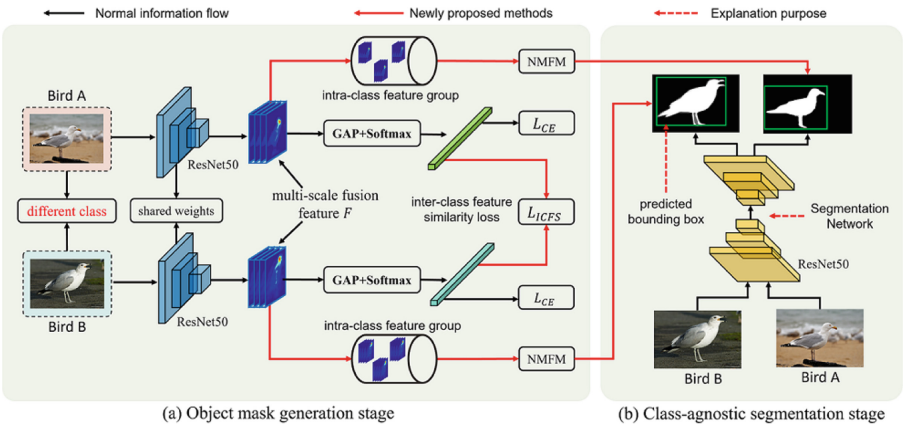

Fig. 3. Pipeline for the proposed ISIC model. In the object mask generation stage, to improve the similarity between different categories, inter-class feature similarity loss $\mathcal{L}_{\text{ICFS}}$ is applied. Besides, based on non-negative matrix factorization, we design NMFM module to generate the object masks instead of CAM, which flows into the subsequent segmentation stage as the pseudo labels. After training, a class-agnostic segmentation model is achieved, which is adopted as the final model to predict the object bounding boxes during inference.

for each input image and extract the object bounding box from the mask. This decoupled design brings three benefits. First, the complex design (*i.e.*, ICFS and NMFM) in object mask generation will not be brought into the inference phase. Hence, the time complexity of the model depends entirely on the segmentation network and is not affected by ICFS or NMFM. Second, unlike the CAM-based methods, which deal with the classification task and the location task at once, our class-agnostic segmentation model focuses only on localization and is not disturbed by the classification task, thus it can derive more complete object regions. Third, the bounding box extraction from a segmentation mask is much easier and less sensitive to the threshold selection than from a class activation map, because values in the segmentation mask are more consistent (tending to 0 or 1), compared with the class activation map. After getting the bounding box, we follow SPOL [24] to use a separate classification network(SPOL adopts the EfficientNet-B7 [21]) to predict the category of the input image. Combining the bounding box and the category, we derive the final results. In fact, this step of obtaining the object category can be omitted, if we focus only on object localization without category information.

3.2 Baseline

As shown in the left part of Fig. 3, our proposed methods (*i.e.*, ICFS and NMFM) are concentrated in the object mask generation stage, which aims to improve the accuracy of the pseudo masks. Before introducing the specific methods, let's

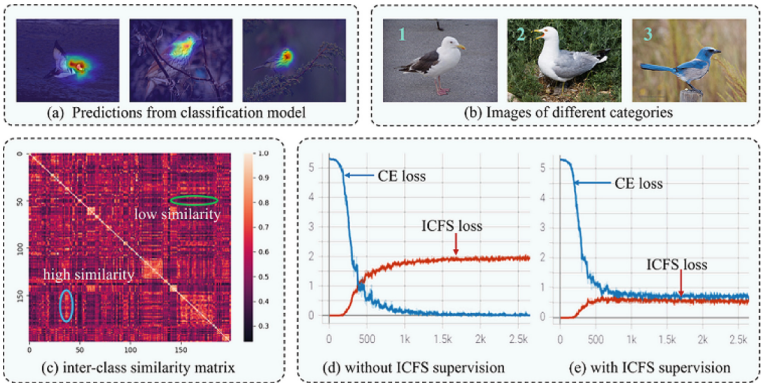

Fig. 4. (a) visualizes incomplete predictions of CAM-based models. (b) shows images from three categories. Obviously, the appearance similarity between image 1 and 2 is larger than that between image 1 and 3. (c) shows the inter-class similarity matrix, where the horizontal and vertical axes both represent the category index. The bright areas (*i.e.*, blue circle) and dark areas (*i.e.*, green circle) indicate the high similarity and low similarity between categories, respectively. (d) shows the loss curve of cross-entropy (CE), where ICFS loss is not adopted as the supervision. (e) shows that ICFS loss is adopted as the supervision.

first introduce the baseline model that we used. Our model is based SPOL [24], which combines the complementarity of deep and shallow features and designs the multiplication fusion strategy to improve the completeness of the object regions. Specifically, SPOL adopts the ResNet50 [8] as the backbone network. For each input image with the size $H \times W$, SPOL extracts its features at five scales (denoted as $\{f_i | i = 1, ..., 5\}$) with the resolutions $[\frac{H}{2^i}, \frac{W}{2^i}]$. Considering the calculation cost, SPOL only uses the last three scale features (*i.e.*, f_3, f_4 and f_5). These features are firstly upsampled to the same scale $[\frac{H}{8}, \frac{W}{8}]$ and then aggregated by element-wise multiplication. In this way, the details of the shallow features and the semantics of the deep features are combined, both of which are helpful for WSOL. We call these aggregated features as the multi-scale fusion features, as shown in Fig. 3. More than that, SPOL also introduces the Gaussian prior pseudo label, self-distillation and auxiliary loss to further enhance the WSOL model. Readers can refer to the specific paper [24] for more details. But to keep the model simple, only the most effective multiplication strategy is involved in our baseline model and the other parts are directly ignored.

3.3 Inter-class and Intra-class Features Analysis

For WSOL, most of previous methods [1,24,32,35] rely on classification models to predict the object masks and then obtain the bounding boxes. Unfortunately, limited by classification models, these masks only cover the most discriminative object regions while other less discriminative ones are ignored. As shown in Fig. 4(a), only the head regions of the birds are highlighted but the body parts

are ignored. Because classification models focus only on the differences (*i.e.*, head parts) between classes. To maximize the classification accuracy, features that have the similar appearance (*i.e.*, body parts) will be discarded. But for WSOL, classification accuracy is not the only goal. Overemphasis on the inter-class differences leads to incomplete object masks. Thus, we argue that WSOL models should also consider the inter-class similarity.

Figure 4(b) shows three images of three categories. From appearance, the similarity between image 1 and 2 is larger than that between image 1 and 3. To quantify the similarities between different categories, we use the pretrained ResNet50 to extract a 128 dimensional vector for each image in both CUB-200 [23] and ImageNet-1k [17], then average the vectors of i-th category as its class representation c_i. For any two representations c_i, c_j, we calculate their cosine similarity $s_{ij} = \frac{c_i c_j}{||c_i||_2||c_j||_2}$. Bringing all s_{ij} together, we get the similarity matrix S. As shown in Fig. 4(c), S is not evenly distributed. The highlighted areas (*e.g.*, blue circle) show the high similarity between categories and the dark areas (*e.g.*, green circle) show the low similarity between categories. However, previous methods ignore the inter-class similarity, thus leading to incomplete predictions.

3.4 Inter-class Similarity Feature Loss

To address the above concerns, we propose the inter-class feature similarity (ICFS) loss, which aims at reducing the feature distance between similar categories. Specifically, we first derive the representation c_i for i-th category according to Sect. 3.3 and then find the similar categories of c_i by $M_i = \{j \mid S_{ij} > \gamma\}$, where i and j are the category indexes, γ is a threshold, M_i is the index collection of categories that similar to c_i and S is the inter-class similarity matrix, as shown in Fig. 4(c). Finally, we could define the distance $\mathcal{D}_i$ between c_i and $c_j (j \in M_i)$ and derive ICFS loss $\mathcal{L}_{\text{ICFS}}$ by Eq. 1

$$\mathcal{D}_i = \frac{1}{N_i} \sum_{j \in M_i} ||c_i - c_j||_2^2, \quad \mathcal{L}_{\text{ICFS}} = \frac{1}{N_k} \sum_{i=1}^{N_k} D_i \tag{1}$$

where N_i is element number of M_i and N_k is the total number of categories. The challenge is how to get the representation c for each category during training. The naive way is feeding the entire training set into the model at each iteration and calculate the class representation for each category, which is totally unacceptable due to the high cost of computation and storage. Alternatively, we regard c_i, c_j as the expectations of the image vectors of i-th and j-th categories, respectively.

$$c_i = E[X_i], \quad c_j = E[X_j] \tag{2}$$

where $E[\cdot]$ is the expectation and X_i, X_j are the image vectors corresponding to i-th and j-th categories, respectively. Hence, we derive the upper bound of $\mathcal{D}_i$.

$$\mathcal{D}_i = \frac{1}{N_i} \sum_{j \in M_i} ||E[X_i] - E[X_j]||_2^2 \leq \frac{1}{N_i} \sum_{j \in M_i} E||X_i - X_j||_2^2 \tag{3}$$

In Eq. 4, we use Monte Carlo sampling to approximate the upper bound, where i and j are the category indexes. p and q are the sample indexes. x_i^p and x_j^q are specific vectors. N_{ip} and N_{iq} are the numbers of x_i^p and x_j^q, respectively.

$$\mathcal{D}_i \leq \mathcal{U}_i = \frac{1}{N_i N_{ip} N_{iq}} \sum_{j \in M_i} \sum_{p=1}^{N_{ip}} \sum_{q=1}^{N_{iq}} ||x_i^p - x_j^q||_2^2 \tag{4}$$

Finally, we replace $\mathcal{D}_i$ with its upper bound $\mathcal{U}_i$ in $\mathcal{L}_{\text{ICFS}}$ and get Eq. 5.

$$\mathcal{L}_{\text{ICFS}} = \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{1}{N_i N_{ip} N_{iq}} \sum_{j \in M_i} \sum_{p=1}^{N_{ip}} \sum_{q=1}^{N_{iq}} ||x_i^p - x_j^q||_2^2 \tag{5}$$

The total training loss consists of cross-entropy loss (*i.e.*, $\mathcal{L}_{\text{CE}}$) and $\mathcal{L}_{\text{ICFS}}$, as shown in Eq. 6, where λ is a hyper-parameter. $\mathcal{L}_{\text{CE}}$ supervises the model to learn the discriminative features between categories. In contrast, $\mathcal{L}_{\text{ICFS}}$ forces the model to learn the similarities between categories. These two losses work against each other so that the model will not go to extremes and eventually reach an equilibrium. Figure 4(d) shows the loss curves for $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{ICFS}}$ when $\lambda = 0$ in CUB-200. The model minimizes the $\mathcal{L}_{\text{CE}}$ as much as possible, and the inter-class difference gradually becomes large. However, when $\lambda = 1$, as shown in Fig. 4(e), inter-class difference is constrained and the model does not go to extremes for classification, thus could get more complete predictions. Note that, ICFS loss aims at improving the integrity of pseudo masks and does not care about the classification performance. Following SPOL [24], a separate classification model is adopted to predict the object category.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{ICFS}} \tag{6}$$

3.5 Intra-class Appearance Consistency

Most of the previous WSOL methods obtain the object mask based on class activation maps, where the parameters of the classifier play an important role. Specifically, given a group of feature maps $\{F_1, F_2, ..., F_N\}$ (extracted before the classifier) with spatial size $W \times H$ and the parameters L of the final classifier with shape $N \times C$, where N and C is the number of maps and categories, respectively. Then the class activation map M_c for the c-th class is derived as Eq. 7. With a threshold, M_c can be binarized to extract the object bounding box.

$$M_c = \sum_{i=1}^{N} L_{i,c} F_i \tag{7}$$

However, CAM-based methods are flawed in two ways. First, the goals of classification and localization are inconsistent. Directly using the parameters of the final classifier to generate the class activation maps is harmful. As shown in Fig. 2,

although the bird's body have been included in the feature maps, the final prediction suffers from the under-utilization of feature maps and get incomplete predictions. Second, for CAM-based methods, each image is processed separately, which is exposed to the risk of accidental noise. Namely, some cluttered background may lead to the prediction failure. In contrast, predictions based on multiple images (Fig. 1(b)) are statistically more robust to noise. By extracting the commonality of multiple images of the same category, accidental risk is reduced and the complementarity between images is fully explored.

Given the above concerns, we propose the non-negative matrix factorization mask (NMFM) module to generate object masks. Different from CAMs [35], NMFM does not rely on the final classifier. Instead, it achieves the object mask based on the appearance consistency of multiple images from the same category. Specifically, NMFM utilizes the non-negative matrix factorization (NMF) to extract the commonalities between images. NMF was first proposed in [10] and has been widely used in face recognition [5], recommender system [13] and data compression [11]. Given a non-negative matrix $V \in R^{m \times n}$, NMF finds two non-negative matrices $P \in R^{m \times c}$ and $Q \in R^{c \times n}$, so that $V \approx PQ$. The specific optimization function is shown in the Eq. 8.

$$\min_{P,Q} \quad f(P,Q) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} (V_{ij} - (PQ)_{ij})^2 \tag{8}$$

$$\text{subject to} \quad P_{ia} \geq 0, Q_{bj} \geq 0, \quad \forall i, a, b, j$$

Instead of relying on the classifier, we apply NMF to compress the feature maps $F \in R^{W \times H \times N}$ into the object mask $M \in R^{W \times H}$. Namely, we find a project direction vector $S \in R^{N \times 1}$ so that $M = F \cdot S$ (dot production), where S is derived from the statistics of multiple F of the same category rather than the parameters of the classifier. Specifically, we split F into $W \times H$ vectors, each of which has N dimensions. Supposing there are T images in each category, then we could get $T \times W \times H$ vectors. Lining up these vectors together, we get a big matrix $\Theta \in R^{TWH \times N}$. To find the optimal projection direction S, we use NMF to decompose Θ into two small matrices $\theta_1 \in R^{TWH \times 1}$ and $\theta_2 \in R^{1 \times N}$ so that $\Theta \approx \theta_1 \cdot \theta_2$(dot production), where θ_1 represents the set of vectors reduced in dimension, which is discarded. θ_2 is what we need, which represents the projection direction and combines the commonality of multiple images. Namely, $S = \theta_2^T$. According to $M = F \cdot S$, the object mask could be derived by $M = F \cdot \theta_2^T$.

Compared with CAMs, NMFM does not rely on the classifier, hence making better use of feature maps, as shown in Fig. 2. Besides, NMFM extracts the commonality of a category of images, which is more robust to background noise. Note that, NMFM is not involved in the training or inference phase. It is just applied to generate the pseudo masks after the classification model has been trained. Thus, it is called only once and will not bring any time complexity for the training or inference phase. With these pseudo masks, we train a class-agnostic segmentation model. The final object bounding boxes are extracted from the predictions of the class-agnostic segmentation model rather than the pseudo masks generated by NMFM.

3.6 Class-Agnostic Segmentation Stage

Although NMFM generates accurate object masks, too many modules are involved in the object mask generation stage, which brings a lot of computation and complexity. To make the inference faster and easier, we use the object masks (generated by NMFM) as the pseudo labels to train a separate class-agnostic segmentation model for prediction. Specifically, we use ResNet50 as the backbone network to extract the features of five scales (denoted as $\{f_i|i = 1, ..., 5\}$) for each image. Similar to the baseline model in Sect. 3.2, only features of the last three scales are utilized, namely f_3, f_4, f_5. We upsample these features to the same scale and aggregate them by element-wise multiplication. Finally, we send the aggregated features to a 1×1 convolutional layer to generate the binary object mask, which is supervised by the pseudo labels derived from NMFM. During inference, for each image, we use the segmentation model to get the object mask and the complex object mask generation stage is discarded. Hence, the whole inference process is simple and quick. Besides, compared with the class activation maps, the predictions of the segmentation are already binary. So precise threshold adjustment for bounding box extraction is no longer required.

4 Experiments

4.1 Experimental Setup

Datasets. CUB-200 [23] and ImageNet-1K [17] are adopted for model evaluation, where CUB-200 consists of 200 categories, with 5,994 training images and 5,794 testing images. ImageNet-1K consists of 1000 categories, with 1,281,197 training images and 50,000 testing images. All the training images have only image-level labels, but the testing images have bounding box annotations.

Metrics. Following [3,24,35], three metrics are adopted to quantify the model performance. 1) Top-1 localization (*Top-1 Loc*): top-1 prediction is exactly the right image class and the IoU (Intersection over Union) between the predicted bounding box and the ground truth one is larger than 0.5. 2) Top-5 localization (*Top-5 Loc*): top-5 predictions contain the right image class and the IoU between the predicted bounding box and the ground truth one is larger than 0.5. 3) GT-known localization (*GT-known Loc*): the IoU between the predicted bounding box and the ground truth one is larger than 0.5.

Data Augmentation and Training Settings. During training, we follow previous methods [3,24,31,32,35] to first resize each input image to 256×256 then randomly crop it to 224×224. Also, a random flip is adopted to increase the diversity of input images. During inference, the random cropping is replaced by the center cropping and the random flip is removed [3,31]. We use the SGD optimizer to train our model, where the learning rates for both CUB-200 [23] and ImageNet-1K [17] are 0.02 and remain constant throughout the training process. Besides, due to the difference in dataset size, the training epochs for CUB-200 and ImageNet-1K are set to 32 and 5, respectively.

Table 1. Performance comparison between the state-of-the-art methods. '–' means no given. The highest scores are highlighted in bold.

Model	Backbone	CUB-200			ImageNet-1K		
		Top-1 Loc	Top-5 Loc	GT-known	Top-1 Loc	Top-5 Loc	GT-known
CAM [35]	VGG16	36.13	–	–	42.80	54.86	59.00
ACoL [32]	VGG16	45.92	56.51	62.96	45.83	59.43	62.96
SPG [33]	InceptionV3	46.64	57.72	–	48.60	60.00	64.69
ADL [3]	VGG16	52.36	–	73.96	44.92	–	–
I^2C [34]	InceptionV3	65.99	68.34	72.60	53.11	64.13	68.50
GC-Net [12]	InceptionV3	58.58	71.10	75.30	49.06	58.09	–
PSOL [31]	InceptionV3	65.51	83.44	–	54.82	63.25	65.21
SPA [16]	VGG16	60.27	72.5	77.29	49.56	61.32	65.05
ORNet [27]	VGG16	67.74	80.77	86.2	52.05	63.94	68.27
TS-CAM [4]	Deit-S	71.3	83.8	87.7	53.4	64.3	67.6
RCAM [1]	ResNet50-SE	58.39	–	74.51	51.96	–	64.40
ADL [3]	ResNet50-SE	62.29	–	71.99	48.53	–	–
SPOL [24]	ResNet50	80.12	93.44	96.46	59.14	67.15	69.02
SLT-Net [6]	ResNet50	72.3	–	90.7	56.2	–	68.5
PSOL [31]	ResNet50	70.68	86.64	90.00	53.98	63.08	65.44
FAM [15]	ResNet50	73.74	–	85.73	54.46	–	64.56
ISIC (Ours)	ResNet50	**80.68**	**94.08**	**97.32**	**59.61**	**67.84**	**70.01**

4.2 Comparison with State-of-the-Arts

Quantitative Comparison. To evaluate the performance of the proposed ISIC, we train it both on CUB-200 [23] and ImageNet-1k [17], as shown in Table 1. Many state-of-the-art methods [1,3,4,12,15,16,24,27,31–35] are also included in Table 1 for comparison. The highest scores are highlighted in bold. Among all these methods, ISIC achieves the highest accuracy on both CUB-200 and ImageNet-1K in terms of **Top-1 Loc**, **Top-5 Loc** and **GT-Known Loc** metrics. Especially for **GT-Known Loc** metric, ISIC achieves a pronounced performance boost, demonstrating its superiority in object localization.

Visual Comparison. Figure 5 shows some localization maps for CUB-200 and ImageNet-1k, where the bottom row and the middle row visualize the predictions derived from CAM [35] and our proposed ISIC, respectively. Obviously, compared with CAM, ISIC could cover more complete object regions rather than only focus on the most discriminative ones. Besides, ISIC predictions preserve sharper object boundaries and more detailed shapes.

4.3 Ablation Studies

Ablation Study for Each Component. We use CUB-200 to evaluate each component of the proposed ISIC. As shown in Table 2, ICFS loss largely improve the localization accuracy of the baseline model by 4.9% in the GT-Known Loc metric, surpassing a lot of SOTA methods, which proves the significance of inter-class similarity for WSOL. Compared with the excessive pursuit of inter-class

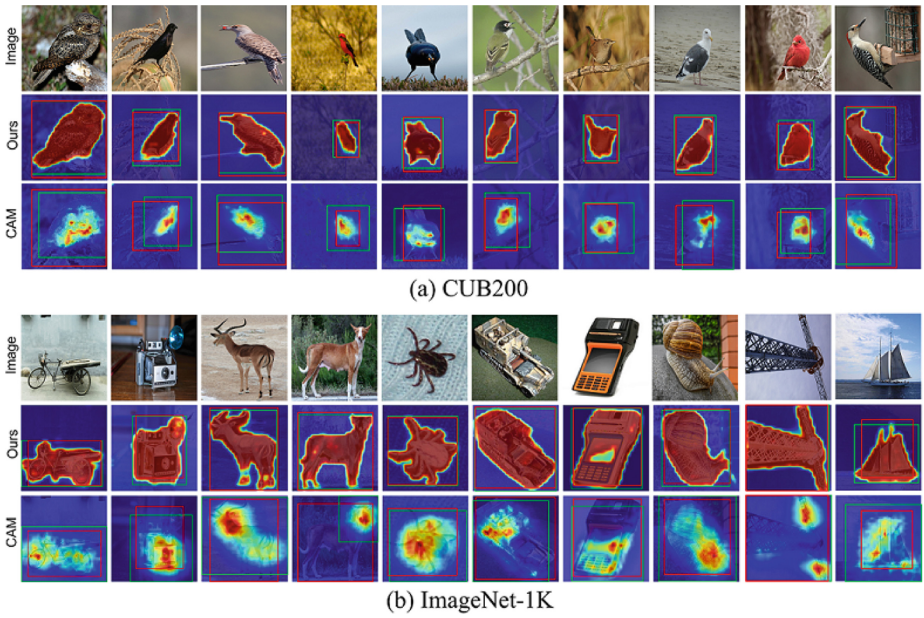

Fig. 5. Visualization of the localization maps with CAM [35] (bottom row) and the proposed ISIC (middle row). Ground truth bounding boxes and the predicted bounding boxes are shown in red and green color, respectively. (Color figure online)

difference in classification models, ICFS loss guides the model to a better balance between inter-class similarity and inter-class difference, thus achieving more complete predictions. Besides, NMFM also boosts the model performance by suppressing noise and improving feature utilization. With all components, the object localization capability of ISIC is largely enhanced.

Ablation Study for λ. In Eq. 6, λ is set to balance $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{ICFS}}$. To study its disturbance with the performance, different values are chosen for CUB-200, as shown in Table 3. $\lambda = 0$ means no ICFS supervision. When $\lambda = 1.0$, the model reaches an equilibrium between inter-class similarity and inter-class difference, achieving the best performance. However, when λ keeps increasing, the balance is broken and the model degrades.

Ablation Study for γ. In Sect. 3.4, we set a threshold γ to find the similarity categories. Table 4 shows its effect at different values. When $\gamma = 0.3$, our model achieves the best performance.

Visualization of the similar categories. Figure 6 shows some images of the similar categories (Sect. 3.4). As shown, category similarity is widespread both in the fine-grained dataset (CUB-200) and the general dataset (ImageNet-1k).

(a) CUB-200 (b) ImageNet-1k

Fig. 6. Images from the similar categories. One row represents a group of categories.

Table 2. Ablation studies for each component of ISIC. BASE is the baseline model. ICFS and NMFM are the proposed components of ISIC. SEG means the class-agnostic segmentation model. CUB-200 is adopted for evaluation.

BASE	ICFS	NMFM	SEG	CUB-200		
				Top-1	Top-5	GT-Known
✓				73.0	85.2	88.3
✓	✓			77.3	90.1	93.2
✓		✓		75.7	88.6	91.5
✓	✓	✓		77.4	90.8	94.1
✓	✓	✓	✓	**80.7**	**94.1**	**97.3**

Table 3. Ablation studies for λ.

λ	0	0.5	1.0	1.5
Top-1	73.0	76.5	77.3	75.4
Top-5	85.2	89.2	90.1	89.8
Gt-Known	88.3	92.3	93.2	92.7

Table 4. Ablation studies for γ.

γ	0.2	0.3	0.4	0.5
Top-1	76.9	77.3	76.7	75.6
Top-5	89.7	90.1	89.5	88.7
Gt-Known	92.8	93.2	92.6	91.9

5 Conclusion

In this paper, we investigate the effect of inter-class similarity on WSOL and propose the ICFS loss against the widely used cross entropy loss. Besides, considering predictions from the classifier are biased to classification task, we propose to abandon CAMs and apply the non-negative matrix factorization to generate object masks. All the proposed modules greatly improve the WSOL performance.

Acknowledgement. This work was supported in part by NSFC-Youth 61902335, by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen HK S&T Cooperation Zone, by the National Key R&D Program of China with grant No. 2018YFB1800800, by Shenzhen Outstanding Talents Training Fund, by Guangdong Research Project No. 2017ZT07X152 and No. 2019CX01X104, by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), by zelixir biotechnology company Fund, by Tencent Open Fund, and by ITSO at CUHKSZ.

References

1. Bae, W., Noh, J., Kim, G.: Rethinking class activation mapping for weakly supervised object localization. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12360, pp. 618–634. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58555-6_37
2. Bai, H., Zhang, R., Wang, J., Wan, X.: Weakly supervised object localization via transformer with implicit spatial calibration. In: ECCV (2022)
3. Choe, J., Shim, H.: Attention-based dropout layer for weakly supervised object localization. In: CVPR, pp. 2219–2228 (2019)
4. Gao, W., et al.: Ts-cam: token semantic coupled attention map for weakly supervised object localization. In: ICCV, pp. 2886–2895 (2021)
5. Guillamet, D., Vitria, J.: Non-negative matrix factorization for face recognition. In: CCAI, pp. 336–344 (2002)
6. Guo, G., Han, J., Wan, F., Zhang, D.: Strengthen learning tolerance for weakly supervised object localization. In: CVPR, pp. 7403–7412 (2021)
7. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: ICCV, pp. 2961–2969 (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
9. Kim, J., Choe, J., Yun, S., Kwak, N.: Normalization matters in weakly supervised object localization. In: ICCV, pp. 3427–3436 (2021)
10. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
11. Liu, W., Zheng, N., Lu, X.: Non-negative matrix factorization for visual coding. In: ICASSP, pp. 111–293 (2003)
12. Lu, W., Jia, X., Xie, W., Shen, L., Zhou, Y., Duan, J.: Geometry constrained weakly supervised object localization. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12371, pp. 481–496. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58574-7_29
13. Luo, X., Zhou, M., Xia, Y., Zhu, Q.: An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems, pp. 1273–1284 (2014)
14. Mai, J., Yang, M., Luo, W.: Erasing integrated learning: a simple yet effective approach for weakly supervised object localization. In: CVPR, pp. 8766–8775 (2020)
15. Meng, M., Zhang, T., Tian, Q., Zhang, Y., Wu, F.: Foreground activation maps for weakly supervised object localization. In: ICCV, pp. 3385–3395 (2021)
16. Pan, X., et al.: Unveiling the potential of structure preserving for weakly supervised object localization. In: CVPR, pp. 11642–11651 (2021)
17. Russakovsky, O.: ImageNet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR, pp. 1–14 (2015)
19. Singh, K.K., Lee, Y.J.: Hide-and-seek: forcing a network to be meticulous for weakly-supervised object and action localization. In: ICCV, pp. 3544–3553 (2017)
20. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the Inception architecture for computer vision. In: CVPR, pp. 2818–2826 (2016)
21. Tan, M., Le, Q.: Efficientnet: rethinking model scaling for convolutional neural networks. In: ICML, pp. 6105–6114 (2019)
22. Vaswani, A., et al.: Attention is all you need. In: NIPS, pp. 5998–6008 (2017)

23. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD birds-200-2011 dataset. Technical Rep. CNS-TR-2011-001, California Institute of Technology (2011)
24. Wei, J., Wang, Q., Li, Z., Wang, S., Zhou, S.K., Cui, S.: Shallow feature matters for weakly supervised object localization. In: CVPR, pp. 5993–6001 (2021)
25. Wei, X.S., Zhang, C.L., Wu, J., Shen, C., Zhou, Z.H.: Unsupervised object discovery and co-localization by deep descriptor transformation, vol. 88, pp. 113–126 (2019)
26. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: a simple classification to semantic segmentation approach. In: CVPR, pp. 1568–1576 (2017)
27. Xie, J., Luo, C., Zhu, X., Jin, Z., Lu, W., Shen, L.: Online refinement of low-level feature based activation map for weakly supervised object localization. In: ICCV, pp. 132–141 (2021)
28. Xue, H., Liu, C., Wan, F., Jiao, J., Ji, X., Ye, Q.: Danet: divergent activation for weakly supervised object localization. In: ICCV, pp. 6589–6598 (2019)
29. Yang, S., Kim, Y., Kim, Y., Kim, C.: Combinational class activation maps for weakly supervised object localization. In: WACV, pp. 2941–2949 (2020)
30. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: regularization strategy to train strong classifiers with localizable features. In: CVPR, pp. 6023–6032 (2019)
31. Zhang, C.L., Cao, Y.H., Wu, J.: Rethinking the route towards weakly supervised object localization. In: CVPR, pp. 13460–13469 (2020)
32. Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.S.: Adversarial complementary learning for weakly supervised object localization. In: CVPR, pp. 1325–1334 (2018)
33. Zhang, X., Wei, Y., Kang, G., Yang, Y., Huang, T.: Self-produced guidance for weakly-supervised object localization. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11216, pp. 610–625. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01258-8_37
34. Zhang, X., Wei, Y., Yang, Y.: Inter-image communication for weakly supervised localization. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12364, pp. 271–287. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58529-7_17
35. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR, pp. 2921–2929 (2016)