Fine-grained label learning in object detection with weak supervision of captions

Wang, X.; Du, Y.; Verberne, S.; Verbeek, F.J.

Multimedia Tools and Applications (2023) 82:6557–6579
https://doi.org/10.1007/s11042-022-13592-7

Fine-grained label learning in object detection with weak supervision of captions

Xue Wang[1,2] · Youtian Du[1] · Suzan Verberne[2] · Fons J. Verbeek[2]

Received: 19 April 2021 / Revised: 30 June 2022 / Accepted: 18 July 2022 /
Published online: 6 August 2022

Abstract

This paper addresses the task of fine-grained label learning in object detection with the weak supervision of auxiliary information attached to images. Most of the recent work focused on the label prediction for objects in the same category space as in training data under the fully-supervised learning framework and cannot be expanded to the learning of more fine-grained categories that have not been defined in training sets. In this paper, we propose a new weakly-supervised learning approach, called label inference curriculum network (LICN), to detecting objects and learning their fine-grained category labels based on supervision of captions via curriculum learning. First, we build a semantic mapping based on embedding techniques and a knowledge base to measure the correspondence between coarse labels and fine-grained label proposals; second, we introduce a label inference curriculum network, which ranks the order of training samples by the complexity of samples. We construct two datasets, namely FG-COCO and FGs-COCO, consisting of both coarse and fine-grained labels based on MS COCO and Visual Genome to train and test our approach. Experimental results demonstrate the effectiveness of our proposed LICN model, and LICN-E2C achieves an improvement of 1.7% mAP with 0.5:0.05:0.95 IoU compared with the LICN-C2E on the FG-sCOCO test dataset.

Keywords Fine-grained label learning · Object detection · Weakly-supervised learning · Semantic mapping · Curriculum learning

✉ Youtian Du
duyt@mail.xjtu.edu.cn

Xue Wang
nimowangxue1989@stu.xjtu.edu.cn; x.wang@liacs.leidenuniv.nl

Suzan Verberne
s.verberne@liacs.leidenuniv.nl

Fons J. Verbeek
f.j.verbeek@liacs.leidenuniv.nl

[1] Xi'an Jiaotong University, No.28, Xianning West Road, Xi'an, Shaanxi, 710049, China

[2] Leiden Institute of Advanced Computer Science, Leiden University, Niels Bohrweg 1, 2333 CA, Leiden, The Netherlands

1 Introduction

Visual object detection is a fundamental problem in computer vision research and has a wide range of applications in real life, such as self-driving vehicles [4] and scene understanding [1]. The performance of an object detector may crucially affect these applications. For instance, as a downstream task of object detection, scene understanding requires an object detector to correctly provide the locations of visual objects present in an image and their class labels. With the renaissance of deep neural networks in recent years, object detection has been revolutionized by a series of groundbreaking works, including Faster-RCNN [26], Mask-RCNN [7] and YOLO [25].

Despite these achievements, most deep learning-based methods suffer from an important limitation: they need to be trained with exhaustive and clean human annotations. These annotations are expensive as they require humans to mark class labels and locations of visual objects in images. In order to address the data limitation, researchers seek to relax this requirement of exhaustively labeled data with a weakly-supervised learning paradigm. A typical problem of weakly-supervised object detection (WSOD) is to learn an object detector under the supervision of only a set of class labels assigned to each image (called image-level labels) [5, 33], where the correspondence between each individual visual object and a class label (called instance-level labels) is unavailable in the training procedure. In general, it requires the labels to be precise. More specially, each visual object in the image has a correct class label belonging to the image-level labels.

Another type of WSOD is to learn object detectors with the supervision of user-generated data on web sources, e.g., social media services like Flickr and Twitter [6, 8]. More and more people tend to share pictures with user-generated tags (or captions) on social media. The user-generated textual data can be seen as natural annotations of the images, providing a weak supervision, for the WSOD problem. It is a cheap way to address the constraint of annotations and increase the scale of datasets near-infinitely. Different from image-level labels, the user-generated natural annotations (tags or captions) consist of a lot of inaccurate or irrelevant terms to supervise the learning of object detectors. Previous works have shown that the weakly-supervised learning can be performed well based on the noisy labels [10, 31]. Misra et al. [23] and Zhang et al. [38] addressed the WSOD problem based on the supervision of captions associated with images. However, the user-generated textual data tend to consist of a diversity of words (of the same semantics), different from the ground-truth label, to describe the same visual object in images. In addition, the words in captions may possess more fine-grained semantics than the class labels in some public datasets, because the latter are determined in a pre-defined label space. For example, Fig. 1 shows multiple image captions that describe the same object (marked by a red bounding box) in the image with different key words (i.e., person and man) than the predefined class label (i.e., person). It is clear that the word "man" is more fine-grained than "person" in describing this object. In addition, we also find in other examples that there are often multiple visual objects of the same class existing in an image, which may cause the ambiguity of correspondence between visual objects and key words in captions.

In this work, we focus on a new WSOD problem, called fine-grained label learning, different from the typical problems introduced above. Suppose we have a set of data consisting of the paired images and captions, as shown in Fig. 1, as well as a coarse label assigned to each visual object as ground truth in training sets. We aim to detect objects and learn the fine-grained labels under the joint supervision of the coarse label for an object and captions for an image. The problem has the following two characteristics. The fine-grained

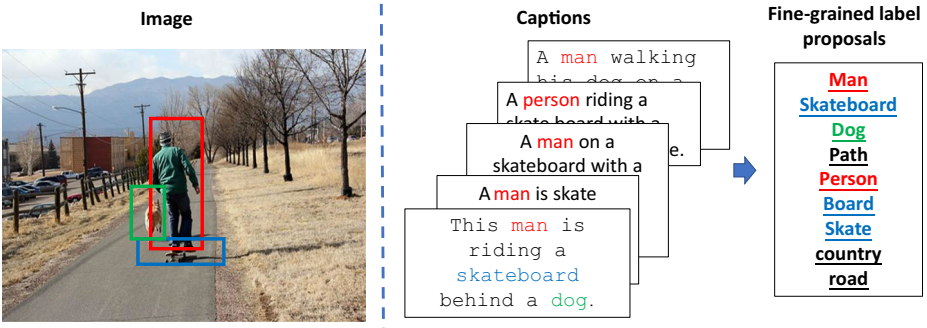

Fig. 1 An illustration of the image-caption pair. For an image, the location of objects (bounding boxes), the corresponding coarse labels, and the attached captions are provided in the datasets for training. In general, the captions consist of a set of fine-grained label proposals for the objects in the image

labels extracted from captions are considerably weak, noisy and ambiguous. Second, the uncertainty of correspondence between visual objects and fine-grained labels, caused by noise and ambiguity in the supervision of captions, is different among the examples. This uncertainty results in different difficulties in learning the fine-grained labels for different examples. Thus, the order of data sequence being fed to models may affect the learning performance. To address the problem, this paper formulates the task of fine-grained label learning with the joint supervision of coarse labels and captions and proposes a novel approach called label inference curriculum network (LICN). First, we build a semantic mapping that provides a correspondence between the coarse labels and fine-grained label proposals coming from captions based on embedding techniques and a knowledge base. Furthermore, we design a curriculum learning process for the Faster R-CNN backbone to detect visual objects and learn the fined-grained labels. To determine the order of training data in the curriculum learning process, we define a term, called the complexity of samples (CoS), that measures the difficulty of learning fine-grained labels for each example.

In summary, our contributions are four-fold.

1. We introduce and formulate the problem of fine-grained label learning with the joint supervision of the coarse category labels and captions.
2. We build a semantic mapping between the coarse labels and fine-grained label proposals coming from captions based on embedding techniques and a knowledge base.
3. We propose a novel approach called LICN and design the weakly-supervised curriculum learning process for improving the learning performance, where the complexity of samples (CoS) is defined to determine the order of training data in the curriculum learning process.
4. We construct two datasets, namely FG-COCO and FGs-COCO, consisting of both coarse and fine-grained labels based on MS COCO and Visual Genome to train and test our approach. Experimental results demonstrate the effectiveness of our proposed LICN model, and LICN-E2C achieves an improvement of 1.7% mAP with 0.5:0.05:0.95 IoU compared with the LICN-C2E on the FG-sCOCO test dataset.

The rest of this paper is organized as follows. Section 2 presents a brief overview of related work. Section 3 formulates the problem of fine-grained label learning and introduces our approach in details. Section 4 provides the experimental results and analysis, and Section 5 concludes the paper.

2 Related work

We review the related work in terms of lexico-semantic analysis, weakly-supervised object detection and curriculum learning.

2.1 Lexico-semantic analysis

In the widely-used public image datasets, there is typically a semantic gap between the human-written captions and the categorical annotations of the objects in the images. For example, as shown in Fig. 1, the annotation of the object in the red box is "person" while the caption uses the word "man". A variety of lexico-semantic methods have been proposed to bridge this semantic gap. These methods can be divided into two categories: knowledge-based methods and corpus-based methods [27, 30]. Knowledge-based methods rely on external semantic resources (thesauri or lexical knowledge bases) to identify similarities between two words. For example, WordNet [32] is used to measure the semantic distance between a pair of words. Although these semantic bases are interpretable and effective, they lack the consideration of the context information and work only for the words present in the lexicon.

Due to the limitations of knowledge-based methods, corpus-based methods are then proposed to utilize context information around the center words. Current corpus-based methods seek to learn vector representations (called embeddings) based on the contexts of words in a large text collection. The word embedding learning research mostly uses a statistical description of the context [2, 16]. Word2Vec [21, 22] is a popular model for text representation, which transforms words into a K-dimensional embedding space based on the context and measures the semantic similarity of two words using their distance in the embedding space. Li et al. [18] proposed to utilize a transferred vector for the representation of a word to reveal its semantics better, not just relying on its own embedding. In our work, we jointly utilize WordNet and Word2Vec to build a semantic mapping between the pre-annotated coarse labels and the fine-grained label proposals coming from captions.

2.2 Weakly-supervised object detection

Object detection is a fundamental task in a lot of applications, such as scene semantic recognition [37] and self-driving vehicles [4]. In recent years, more and more researchers have paid much attention to weakly-supervised object detection. In general, this task aims to detect objects from images based on the supervision of a set of image-level labels [11, 29, 39]. Most existing methods formulate this task as a multiple instance learning (MIL) problem. In this case, MIL considers the visual objects in an image as a bag of instances associated with a label (a set of labels). Oquab et al. [24] and Zhou et al. [40] proposed a global average (max) pooling layer to learn class activation maps. Bilen et al. [5] proposed a weakly-supervised deep detection network (WSDDN) containing classification and detection data streams, where the detection stream weighs the results of the classification predictions. Kantorov et al. [15] improved WSDDN by considering the context information. Tang et al. [28, 29] jointly trained multiple refining models together with WSDDN and showed the benefit from the online iterative refinement. Diba et al. [7] and Wei et al. [35] applied a segmentation map and Wan et al. [33] incorporated saliency to improve the performance of weakly-supervised object detection.

To the best of our knowledge, the above existing WSOD methods have not involved captions, a type of weaker supervisory information than exact image-level labels, in object

detection. Ye et al. [36] harvested detection models from free-form text and used a label inference module to amplify signals in the free-formed texts to supervise the learning of a multiple instances detection network. Jerbi et al. [14] proposed a learning procedure that extracts textual scene graphs from captions and use them within a weak supervision framework to learn object detectors.

Our work is similar to the above works in terms of the MIL weighted representation for the visual objects in an image. However, we go one step further to successfully adopt a more challenging supervision scenario where the captions are utilized as the weak supervision for learning fine-grained labels in the task of object detection.

2.3 Curriculum learning

Bengio et al. [3] proposed the curriculum learning method that designs a learning order by measuring the complexity of data in feature space. Guo et al. [12] proposed a method called CurriculumNet, which allows for an efficient implementation of curriculum learning on large-scale web images, resulting in a high-performance CNN model. The order of the curriculum learning reduces the negative impact of noisy labels substantially. Wang et al. [34] addressed the object detection problem by learning an effective object detector using weakly-annotated images with curriculum learning. Hacohen et al. [13] analyzed the effect of curriculum learning, which involves the non-uniform sampling of mini-batches, on the training of deep networks. In this paper, we determine the order in the curriculum learning process by measuring the degree of complexity of samples in fine-grained label learning.

3 Methodology

3.1 Overview

In this paper, we are given a set of data pairs, each consisting of an image and its captions. Formally, we have $\mathcal{D}_{tr} = \{(I_i, \mathcal{R}_i, \mathcal{L}_i^I, C_i)\}_{i=1}^{M_{tr}}$ and $\mathcal{D}_{te} = \{(I_i, \mathcal{L}_i^I, C_i)\}_{i=1}^{M_{te}}$ as the training set and test set, respectively, where I_i and C_i denote the i-th image and caption, respectively, and $\mathcal{L}_i^I = \{l_{i1}^I, l_{i2}^I, \cdots, l_{im_i}^I\}$ refers to the annotations of I_i, each considered as a coarse category label and assigned to one of the m_i visual object regions $\mathcal{R}_i = \{r_{i1}, r_{i2}, \cdots, r_{im_i}\}$ segmented from this image. The caption C_i consists of a set of entities that generally provide more fine-grained category information than $\mathcal{L}_i^I$ for the visual object regions $\mathcal{R}_i$. We extract them from captions as fine-grained label proposals, denoted by $\mathcal{L}_i^C = \{l_{i1}^C, l_{i2}^C, \cdots, l_{in_i}^C\}$. In this manner, we have a coarse label vocabulary $\mathcal{V}_I$ and a fine-grained label proposal vocabulary $\mathcal{V}_C$ that consist of all coarse labels and fine-grained label proposals, respectively, where $l_{i\cdot}^I \in \mathcal{V}_I$ and $l_{i\cdot}^C \in \mathcal{V}_C$. Regarding the labels, we make two observations: 1) The label proposals $\mathcal{L}_i^C$ from captions are generally more fine-grained than the coarse category labels $\mathcal{L}_i^I$ preassigned to the visual object regions; 2) the correspondence at the granularity of instances (i.e., between a fine-grained label proposal $l_{i\cdot}^C$ and a visual object region $r_{i\cdot}$) is missing. An example can be seen in the second image of Fig. 2a. It is in this image unknown which region corresponds to the fine-grained label "man" or "woman" extracted from the captions.

We aim to learn and infer the fine-grained label $l_{i\cdot} \in \mathcal{V}_I \cup \mathcal{V}_C$ for each visual object region based on the supervision from the training data $\mathcal{D}_{tr}$. As illustrated in Fig. 2, our framework includes two main processes: semantic mapping and curriculum learning-based

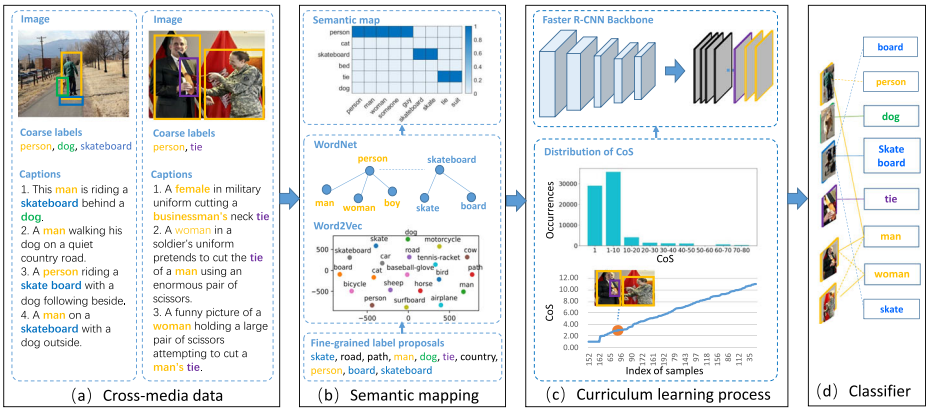

Fig. 2 The framework of the proposed LICN approach. (a) The approach takes cross-media data in the form of image-captions pairs as the input, where each coarse label is individually assigned to a visual object and captions are associated with an image. (b) The semantic mapping module builds a correspondence between a coarse label and a set of fine-grained label proposals extracted from captions based on embedding techniques and knowledge bases. (c) We use the Faster R-CNN as the backbone of object detection, and introduce the curriculum learning process in the fine-grained label learning with the consideration of the complexity of samples. (d) The classifier is used to predict fine-grained labels

fine-grained label learning. In the semantic mapping, we extract the entities from captions as the fine-grained label proposals $l_{i\cdot}^{C} \in \mathcal{V}_C$, and measure the semantic similarity between the extracted label proposals and the coarse labels $l_{i\cdot}^{I} \in \mathcal{V}_I$ based on the combination of the knowledge base WordNet and data-driven embedding techniques. To learn the fine-grained label for each object, we propose a curriculum learning-based method to train the model by adding data in ascending order of example complexity.

3.2 Semantic mapping

The purpose of semantic mapping is to build the relationship between the coarse label $l_{i\cdot}^{I}$ and the fine-grained label proposals $l_{i\cdot}^{C}$ by measuring their similarity over the training set. We extract all nouns from captions with the CoreNLP toolkit [20] as the candidates for the fine-grained label proposals. In order to get a semantic mapping, we pass three steps.

3.2.1 Semantic mapping based on knowledge base

We employ WordNet as the knowledge base to measure the semantic similarity between the coarse labels and fine-grained label proposals. WordNet can represent relations between word senses with an ontology. For a coarse label $l_{i\cdot}^{I}$, we obtain the synset $\mathcal{W}_{kb}(l_{i\cdot}^{I})$ from WordNet in the form of:

$$\mathcal{W}_{kb}(l_{i\cdot}^{I}) = \{H_{per}(l_{i\cdot}^{I}), H_{pon}(l_{i\cdot}^{I}), S_{non}(l_{i\cdot}^{I})\}, \tag{1}$$

where $H_{per}(\cdot)$, $H_{pon}(\cdot)$ and $S_{non}(\cdot)$ refer to the hypernym, hyponym and synonym, respectively, for a given word in the WordNet.

3.2.2 Semantic mapping based on embedding

We use Word2Vec as the embedding technique to measure the similarity of labels in $\mathcal{V}_I \cup \mathcal{V}_C$. We fine-tune the pre-trained Word2Vec model [21] on all captions in the data. By our analysis, the coarse labels preassigned to visual object regions all appear in captions of the public dataset used in experiments. Thus, we can obtain the feature vector of each coarse label in the embedding space generated using Word2Vec. As the fine-grained label proposals are extracted from the captions, we can obtain the feature vectors of fine-grained labels as well. For a coarse label $l^I_{i\cdot}$ and a fine-grained label proposal $l^C_{i\cdot}$, we achieve their d_e-dimensional feature vectors $\mathbf{l}^I_{i\cdot}$ and $\mathbf{l}^C_{i\cdot}$, respectively. The similarity between two vectors in the embedding space is measured by the cosine similarity $S(\cdot, \cdot)$.

3.2.3 Building the semantic mapping

As analyzed above, we build a semantic mapping between the coarse label $l^I_{i\cdot}$ and the fine-grained label proposal $l^C_{i\cdot}$ with the following matrix:

$$W(l^I_{i\cdot}, l^C_{i\cdot}) = \begin{cases} 1, \ if \ l^C_{i\cdot} \in \mathcal{W}_{kb}(l^I_{i\cdot}) \ and \ S(\mathbf{l}^I_{i\cdot}, \mathbf{l}^C_{i\cdot}) > \varepsilon \\ 0, \ otherwise \end{cases}, \tag{2}$$

where ε is a threshold in $[0, 1]$. With (2), we can find one or multiple fine-grained label proposals that are semantically similar to a preassigned coarse label. Since a visual object region strictly corresponds to a coarse label in the dataset, we can achieve a weak correspondence between visual object regions and fine-grained label proposals.

3.3 Fine-grained label learning based on curriculum learning

In the following subsection, we will find that the examples are of different difficulties to learn and infer the fine-grained labels. Therefore, we perform the fine-grained label learning based on the curriculum learning framework.

3.3.1 Backbone for object detection

Based on the semantic mapping introduced in Section 3.2, we have achieved the correspondence between each visual object region $r_{i\cdot}$ in the i-th image and one or multiple fine-grained label proposals (i.e., a subset of $\mathcal{L}^C_i$). Without ambiguity, we re-denote them by r_k and $\widetilde{\mathcal{L}}^C_k$ by removing the subscript i indicating the index of images, where k is the index of a visual object region in the dataset, $r_k \in \mathcal{R}_i$ and $\widetilde{\mathcal{L}}^C_k \subset \mathcal{L}^C_i$. Thus, our objective is to localize the visual object and learn its fine-grained label with the weak supervision of a set of fine-grained label proposals $\widetilde{\mathcal{L}}^C_k$ to the visual object region r_k.

We use the Faster R-CNN model [26], denoted by $F_{det}(I_i)$, as the backbone of our work. Faster R-CNN consists of three modules: a convolutional neural network for generating the feature map of an image, a region proposal network (RPN) for generating a set of rectangular object proposals based on the feature map, and a classifier for learning the category label of each region. The output of the backbone can be described as follows:

$$(\mathbf{P}_i, \mathcal{R}_i) = F_{det}(I_i), \tag{3}$$

where $\mathcal{R}_i = \{r_{ij}\}_{j=1}^{m_i}$ denotes the set of m_i visual object regions extracted from the image I_i, in which the location of each region is described by four coordinates of the bounding box, and $\mathbf{P}_i = [\mathbf{p}_{i,1}, \mathbf{p}_{i,2}, \cdots, \mathbf{p}_{i,m_i}]$ denotes the probabilities that all object

regions in $\mathcal{R}_i$ are predicted to categories. Without ambiguity, we rewrite $\mathbf{p}_{i,j}$ as $\mathbf{p}_k = [p_{k,1}, p_{k,2}, \cdots, p_{k,C_C}]^T$ by ignoring the index of images, where $p_{k,c}$ denotes the probability that a visual object region r_k is categorized into the c-th class and C_C denotes the cardinality of $\mathcal{V}_C$ (the same as the cardinality of $\mathcal{V}_I \cup \mathcal{V}_C$ since the coarse labels all appear in the fine-grained label proposals). In our work, we define the space of categories with the fine-grained label proposals, i.e., $\mathcal{V}_C$.

3.3.2 Curriculum learning

Curriculum learning [3] aims to formalize the learning process of humans and animals and organize it in a meaningful order to train models. The basic idea is to start learning easier aspects of the task or easier subtasks, and then gradually raise the level of difficulty. By using the concepts of "easy" and "hard" instances, curriculum learning is an efficient learning framework that imposes a structure on the training set. Dealing with noise and outliers has recently been the subject of curriculum learning. Designing or defining an effective learning order for each sample in the training dataset is the most crucial challenge. In this paper, we define the complexity of samples and employ curriculum learning to learn the fine-grained labels of objects.

3.3.3 The complexity of samples

Different samples have different difficulties in fine-grained label learning. For example, if there is only an object region annotated with the coarse label "person" in an image and only a fine-grained label proposal "man" in the caption is related to the coarse label according to the semantic mapping in (2), it is easy to infer the fined-grained label for the object region. In contrast, if there are multiple fined-grained label proposals corresponding to the coarse label according to the semantic mapping, it is much more difficult to discriminate which one is the true fine-grained label of the object region. We introduce a term called ***the complexity of samples*** (CoS) to describe the difficulty in the task. We define the CoS of a sample $D_i \in \mathcal{D}$ as follows:

$$H_{CoS}(D_i) = -\sum_{l_{i\cdot}^I} \sum_{l_{i\cdot}^C} \Pr(l_{i\cdot}^C | l_{i\cdot}^I) \log(\Pr(l_{i\cdot}^C | l_{i\cdot}^I)), \tag{4}$$

where $\Pr(l_{i\cdot}^C | l_{i\cdot}^I)$ is the conditional probability of the fine-grained label proposal $l_{i\cdot}^C$ given the coarse label $l_{i\cdot}^I$ and can be achieved by:

$$\Pr(l_{i\cdot}^C | l_{i\cdot}^I) = \frac{W(l_{i\cdot}^I, l_{i\cdot}^C)}{\sum_{l_{i\cdot}^C \sim l_{i\cdot}^I} W(l_{i\cdot}^I, l_{i\cdot}^C)}, \tag{5}$$

where $l_{i\cdot}^C \sim l_{i\cdot}^I$ denotes all fine-grained label proposals $l_{i\cdot}^C$ related to the coarse label $l_{i\cdot}^I$ according to (2). As shown in (4), CoS is defined based on the Shannon's Entropy that is mainly used to measure the uncertainty of a discrete random variable. In this work, we consider $l_{i\cdot}^I$ as the random variable and $l_{i\cdot}^C$ as its values with non-zero probability. If more fine-grained label proposals are related to the coarse label, the correspondence between them is more uncertain and the fine-grained label learning is thus more intractable. Moreover, if there are multiple visual objects detected in an image, the CoS tends to be a larger value accordingly based on (4).

3.3.4 Curriculum learning process

Based on the semantic mapping, we have obtained the fine-grained label proposals $\widetilde{\mathcal{L}}_k^C$ for each visual object region r_k. Here we transform $\widetilde{\mathcal{L}}_k^C$ to a binary vector $\mathbf{y}_k = [y_{k,1}, y_{k,2}, \cdots, y_{k,C_C}]^T \in \{0,1\}^{C_C}$. $y_{k,c} = 1$ ($y_{k,c} = 0$) means the c-th fine-grained label proposal of $\mathcal{V}_C$ is present (absent) in $\widetilde{\mathcal{L}}_k^C$.

In the curriculum learning process, the training data are fed to Faster R-CNN in the order of easy samples (i.e., with low CoS) to hard samples (i.e., with high CoS). The loss of fine-grained label learning is defined as follows:

$$L_{ws}^k = \sum_{c=1}^{C_C} y_{k,c} \cdot \log p_{k,c} + (1 - y_{k,c}) \cdot (1 - \log p_{k,c}), \tag{6}$$

where L_{ws}^k refers to the weakly supervised loss. The difference from the original Faster R-CNN is that the ground truth of label vector, i.e., $\mathbf{y}_k$, may consist of multiple ones corresponding to multiple fine-grained label proposals, rather than being a one-hot vector.

4 Experimental results and discussion

In this section, we evaluate the effectiveness of the proposed model LICN by answering the following two questions. Q1: How reasonable is the semantic mapping for this weakly-supervised object detection model? Q2: How effective the proposed LICN approach is in terms of the fine-grained label learning based on weakly-supervised paradigm learning?

4.1 Experimental setup

For the experimental setup, we first describe the datasets and then the implementation details.

4.1.1 Datasets

In order to compare with previous works, our experiments are conducted on three widely used public datasets: the MS COCO 2017 dataset [19], Visual Genome [17], and the Pascal VOC 2007 test dataset [9]. Table 1 shows an overview of these datasets.

Table 1 An overview of the datasets

Datasets	# of images	# of categories	# of objects
Visual Genome	107,228	80,138	3,909,697
MS COCO	118,287	80	860,001
FG-COCO	118,287	169	860,001
sCOCO training	76,631	69	200,962
FG-sCOCO training	76,631	150	200,962
FG-sCOCO test	13,175	150	29,169
FG-sCOCO val.	2,000	150	14,090
Visual Genome test	54,212	150	496,809

- The *MS COCO 2017* dataset contains 118,287 training images and 5,000 validation images. It provides 5 captions per image and a total of 80 category labels for the object regions segmented from all the images. The category labels are utilized as the coarse labels $\mathcal{L}_i^I$ and the captions are used for extracting fine-grained label proposals $\mathcal{L}_i^C$ for image I_i and building the semantic mapping.
- *Visual Genome* contains 107,228 images, 3,909,697 objects from 80,138 categories, and other information such as the relationships between objects. Visual Genome consists of much more fine-grained categories than MS COCO and is thus employed for testing the performance of fine-grained label learning with the category labels as the ground truth.
- The *Pascal VOC 2007 test* dataset has 4,952 images and 20 categories of objects. It is utilized as the test dataset.

As MS COCO 2017 dataset contains both the annotations of visual objects and the captions associated with each image, we employ it as the training dataset. The Visual Genome dataset [17] provides a large number of fine-grained category labels, and more than one half of images in the dataset also appear in the MS COCO 2017 dataset. We construct our test dataset based on the images appearing in both the MS COCO 2017 and Visual Genome datasets to evaluate the performance of our approach. We construct the following datasets for training and testing our approach:

- *FG-COCO*: We replace the coarse category labels of the objects in each image in MS COCO with the fine-grained label proposals appearing in the corresponding caption based on the semantic mapping and thus obtain FG-COCO. A total of 169 fine-grained category labels (including the original coarse labels from MS COCO and new fine-grained category labels) are generated for the objects in the dataset.
- *FG-sCOCO test dataset*: For an image appearing in both MS COCO and Visual Genome, if the Intersection over Union (IoU) between two bounding boxes separately provided by the two datasets is larger than 0.90, we keep the image as an example of the FG-sCOCO test dataset. The bounding box surrounding a visual object is provided by MS COCO and the corresponding fine-grained labels are provided by Visual Genome as the ground truth of locations and labels, respectively. In the generation of ground-truth labels, we only keep those fine-grained labels in Visual Genome that also appear in the captions in MS COCO. We randomly choose 2000 images from the set for validation (called FG-sCOCO val. as shown in Table 1), and the rest is for the test. As a result, the FG-sCOCO test dataset consists of 13,175 images and 29,169 objects of 150 fine-grained categories.
- *FG-sCOCO training dataset*: It is a subset of FG-COCO, which excludes all the images appearing in the FG-sCOCO test and FG-sCOCO val. dataset. This dataset consists of 76,631 images and 200,962 objects of 150 fine-grained categories. To make the learning robust, we keep only the categories consisting of more than 200 examples of object regions in the dataset.
- *sCOCO training dataset*: As a subset of MS COCO, it consists of all the images in FG-sCOCO training dataset, and the bounding boxes and category labels are from MS COCO. As a result, the dataset consists of 76,631 images and 69 coarse labels for 200,962 objects.
- *Visual Genome test dataset*: Different from the FG-sCOCO test dataset, the Visual Genome test dataset is the subset of Visual Genome that excludes all the images appearing in MS COCO. In this dataset, we only keep those objects whose category labels

appear in the FG-sCOCO training dataset. As a result, the dataset consists of 54,212 images and 496,809 objects of 150 fine-grained categories.

The following is an analysis of the building of the FG-sCOCO test dataset. In general, for the same object in an image, we generally have a high IoU between the bounding boxes separately provided by Visual Genome and MS COCO. As shown in Fig. 3, the paired bounding boxes with a high IoU value has the same semantics, but may have different object labels. As Visual Genome has 80K category labels that contain all fine-grained categories, we use these labels as ground truth to evaluate the semantic mapping (Q1). We illustrate a few results for different IoU thresholds in Fig. 4. Table 2 shows the effect of different IoU thresholds on the data. For example, there are 19,702 paired objects with an IoU larger than 0.90 from 15,529 images, and these objects belong to 74 categories in MS COCO and 413 categories in Visual Genome. Considering the count of test data and the object categories, we will evaluate our model on the FG-sCOCO test dataset with IoU in [0.90, 1]. For the object detection (Q2), we found that an image in MS COCO has a little difference from the same one in Visual Genome in terms of the size. We resize the Visual Genome images to the size of images in MS COCO.

4.1.2 Implementation details

We train the proposed model on two different datasets: FG-COCO and FG-sCOCO, and thus generate the following four configurations:

- LICN-E2C$_{FG-COCO}$: learned on the FG-COCO dataset by feeding training examples from easy to complex;
- LICN-C2E$_{FG-COCO}$: learned on the FG-COCO dataset by feeding training examples from complex to easy;
- LICN-E2C$_{FG-sCOCO}$: learned on the FG-sCOCO training dataset by feeding training examples from easy to complex;
- LICN-C2E$_{FG-sCOCO}$: learned on the FG-sCOCO training dataset by feeding training examples from complex to easy.

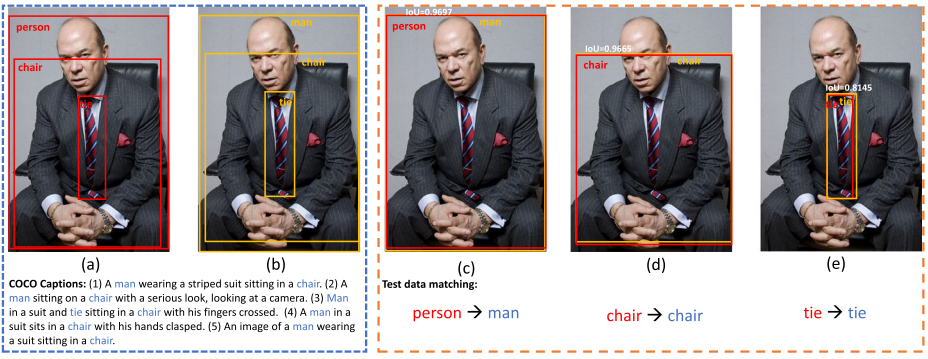

Fig. 3 Test data example: (a) shows an example from MS COCO with object bounding boxes and the associated category labels (red color); (b) shows the same image in the Visual Genome dataset with object bounding boxes and the associated category labels (blue color); (c), (d) and (e) show the matching between the object regions from MS COCO and Visual Genome with an IoU value larger than 0.90. We see that "person" matches to "man", "chair" to "chair" and "tie" to "tie"

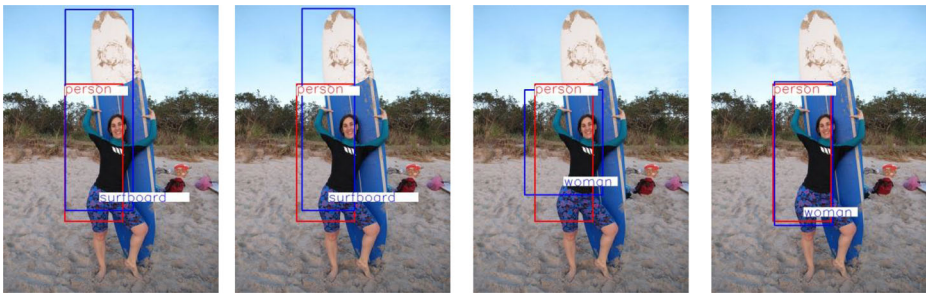

Fig. 4 IoU example: The red color box and blue color box come from MS COCO and Visual Genome, respectively, and the IoU value of two different boxes of the same object should be high

In this work, we employ VGG-16 as the basic model of Faster R-CNN due to the fewest feature memory requests for VGG-16 among the commonly used deep networks. VGG-16 is pre-trained on ImageNet and then fine-tuned on our training datasets. In the process of fine-grained label learning, we use the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a learning rate of 0.01. We set the maximum epoch to 20 for the convergence of the learning process. The minibatch size is set to 1 for the flexible feeding of the examples of different complexity. All experiments are conducted on a platform of 8 Nvidia Titan V GPUs with Pytorch.

4.2 Evaluation metrics

4.2.1 Semantic mapping

We define a weighted semantic mapping Jaccard index (SMJI) for measuring the closeness between the fine-grained labels extracted based on semantic mapping and the fine-grained label ground truth provided in the FG-sCOCO validation set. The weighted SMJI is defined as follows:

$$\text{W\_SMJI} = \sum_{k} W_k \cdot \frac{\mid \mathcal{L}_k^{SM} \cap \mathcal{L}_k^{GT} \mid}{\mid \mathcal{L}_k^{SM} \cup \mathcal{L}_k^{GT} \mid}, \tag{7}$$

Table 2 The characteristics of MS COCO and Visual Genome with different IoU threshold values

IoU	# of images	# of objects	# of categories in MS COCO	# of categories in Visual Genome
0.50	30,983	96,529	79	2,004
0.55	29,337	85,468	79	1,680
0.60	27,621	75,772	79	1,407
0.65	26,118	67,248	79	1,143
0.70	24,890	59,503	78	940
0.75	23,591	51,222	77	787
0.80	21,958	41,848	76	654
0.85	19,603	31,303	76	537
0.90	15,529	19,702	74	413
0.95	7,306	7,957	72	281

where W_k is a weight, $|\cdot|$ denotes the cardinality of a set, $\mathcal{L}_k^{SM}$ and $\mathcal{L}_k^{GT}$ denote the sets of fine-grained labels extracted based on semantic mapping and fine-grained label ground truth provided in the FG-sCOCO validation set, respectively, corresponding to the k-th coarse category label, and the operators $\cup$ and $\cap$ denote the union and intersection of two sets, respectively. For example, for the coarse category label "person", we have $\mathcal{L}_k^{SM}$= {"guy", "man", "person", "woman", "someone"} and $\mathcal{L}_k^{GT}$= {"guy", "man", "person", "skateboarder", "surfer", "woman"}. The weight W_k in (7) is defined as follows:

$$W_k = \frac{|\mathcal{L}_k^{GT}|}{\sum_k |\mathcal{L}_k^{GT}|} \tag{8}$$

Figure 5 reports the weighted SMJI on the FG-sCOCO validation set as the threshold ε changes. From the figure, we observe that the performance of semantic mapping in mining the fine-grained labels is optimal when $\varepsilon = 0.72$. Thus, we choose $\varepsilon = 0.72$ in the following experiments. Figure 6 illustrates the result of semantic mapping that consists of 69 coarse category labels and 81 fine-grained category labels appearing in the FG-sCOCO validation set. From the figure, we observe that most fine-grained label proposals extracted from captions are semantically similar to the coarse labels, while a few noises are introduced by the semantic mapping. For example, the generated "chicken", "meat", "pasta", "rice" and "sauce" are not semantically similar to the coarse label "broccoli". The effect caused by the noises will be reduced with the curriculum learning process.

In Fig. 7, we illustrate the occurrence frequencies of the category labels (including the coarse and fine-grained labels) in the FG-sCOCO training dataset and the sCOCO training dataset, which correspond to the categories with and without semantic mapping, respectively. Due to the large difference in the occurrence frequencies of these categories, we report the results separately in three subfigures. From the figure, we find that a large number of fine-grained label proposals are introduced from captions based on semantic mapping.

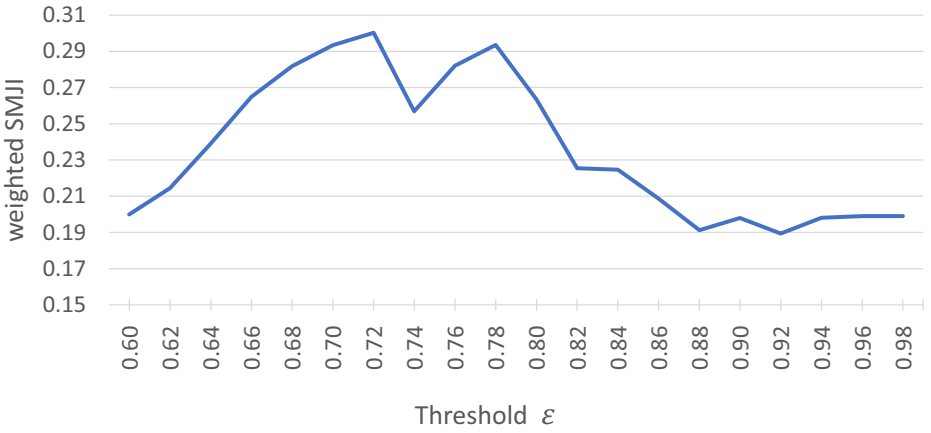

Fig. 5 The effect of ε in (2) on the performance of semantic mapping in terms of weighted SMJI on the FG-sCOCO validation set

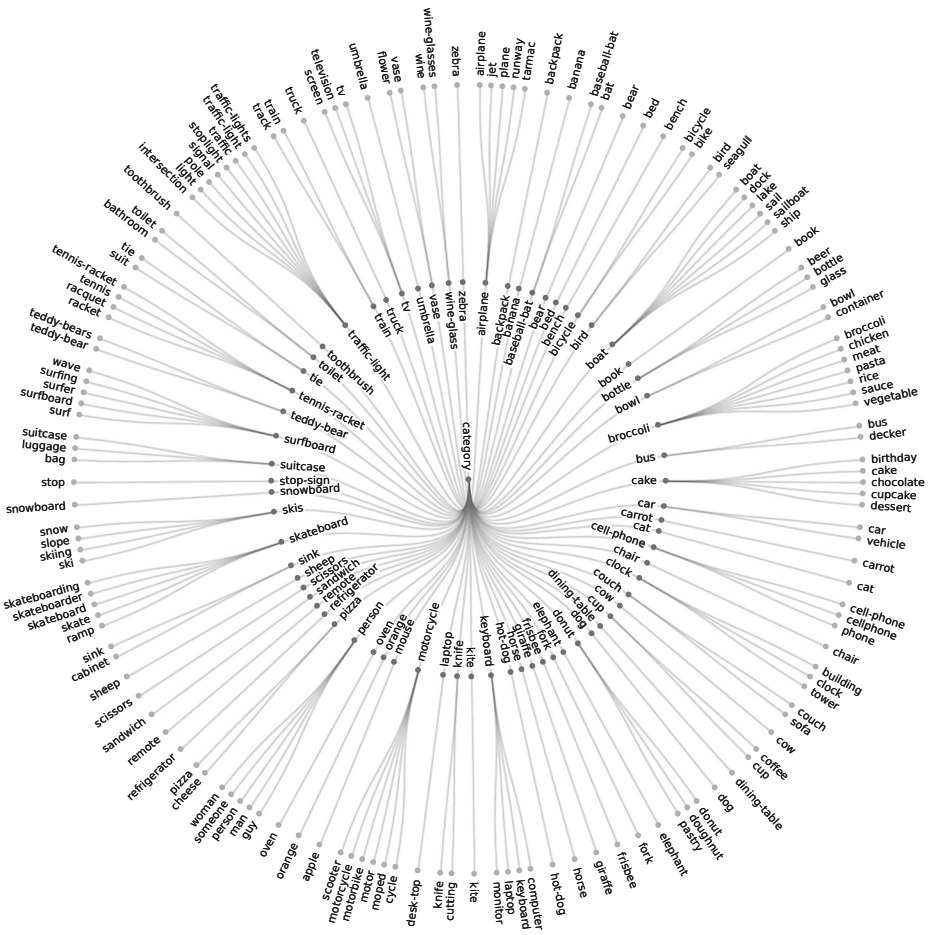

Fig. 6 The result of semantic mapping for the FG-sCOCO validation set, which consists of 69 coarse category labels (points on the inner circle) and 81 fine-grained category labels (points on the outer circle)

4.2.2 Object detection

We utilize a widely-used metric, namely average precision (AP), to evaluate the performance of weakly-supervised object detection. AP is defined as the average detection precision under different recalls and usually evaluates the performance in a category specific manner. To compare performance in overall object categories, the mean AP (mAP) averaged over all object categories is usually used as the final metric of performance. To measure the object localization accuracy, we need to check whether the IoU between the predicted bounding box and the ground-truth bounding box is greater than a predefined threshold of 0.5. Instead of using a fixed IoU threshold, we evaluate our model with AP over the IoU thresholds from 0.5 (coarse localization) to 0.95 (perfect localization).

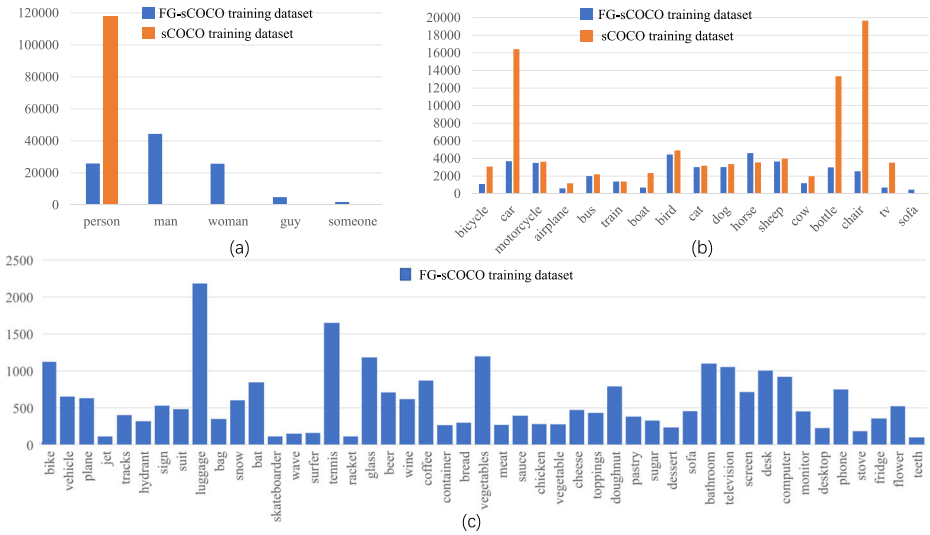

Fig. 7 The comparison of occurrence frequencies of category labels between before and after semantic mapping, where the orange bars indicate the occurrence frequencies of the coarse labels in sCOCO training dataset and blue bars indicate the occurrence frequencies of the labels (either the original coarse labels or the generated fine-grained label proposals) in our constructed FG-sCOCO training dataset after semantic mapping. a) Comparison between the coarse label of category "person" and the corresponding fine-grained labels, b) comparison on 17 coarse categories, and c) comparison on the generated fine-grained categories

4.3 Performance and analysis

4.3.1 FG-sCOCO

We first evaluate our method on the FG-sCOCO validation dataset to analyze the importance of curriculum learning, where the proposed LICN model is trained on the FG-sCOCO training dataset.

Figure 8 shows the results of LICN with two different settings, i.e., LICN-E2C and LICN-C2E, for the FG-sCOCO validation dataset. We find that the LICN-E2C setting improves the performance of fine-grained label learning. As shown in Fig. 8a, in terms of the mean AP with 0.5:0.05:0.95 IoU, LICN-E2C performs approximately 0.02 mAP improvement better than the LICN-C2E model. However, in the Fig. 8b, there is not a clear

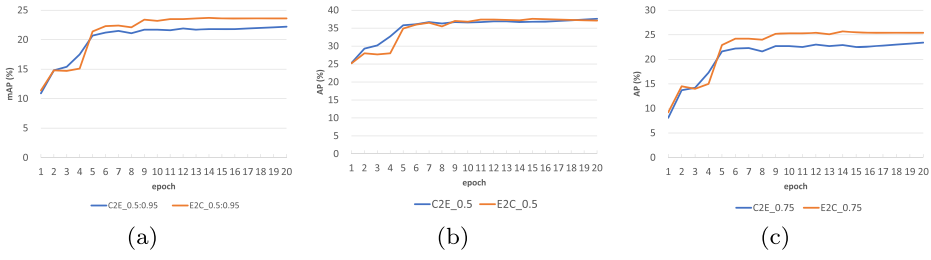

Fig. 8 Results of LICN-E2C and LICN-C2E on the FG-sCOCO validation set for different training epochs. (a) mAP with 0.50:0.95 IoU in steps of 0.05, (b) mAP with IoU over 0.5, and (c) mAP with IoU over 0.75

difference between the LICN-E2C and LICN-C2E model after 7 epochs in terms of the mAP with IoU over 0.5. Figure 8c reports the results in terms of mAP with IoU over 0.75, which shows that LICN-E2C performs approximately 0.03 AP better than the LICN-C2E model. As IoU means the object location accuracy, IoU close to 1 means that the predicted object location is close to the ground truth. Thus, Fig. 8 indicates that the improvement is brought by considering the ascending order of CoS in curriculum learning as the IoU increases.

Table 3 shows a more detailed experimental result on the FG-sCOCO test dataset. In Table 3, "Avg. Precision, Area S M L" means the average precisions for small ($area < 32^2$), medium ($32^2 < area < 96^2$), and large ($area > 96^2$) objects, respectively, where the area is measured with the number of pixels in the segmentation mask. The table shows that in the case of 0.75 IoU, the LICN-E2C setting improves the performance by 2.6% compared with LICN-C2E, which also demonstrates that it is better to learn the fine-grained labels with the consideration of CoS defined in the "Methodology" section.

4.3.2 VOC 2007

We train our model on FG-COCO and the FG-sCOCO training dataset and test the learned models on the VOC 2007 test dataset to evaluate the object detection performance. Correspondingly, the Faster R-CNN baseline is trained on MS COCO and the sCOCO training dataset. Table 4 shows the experimental results for the 20 coarse categories in the VOC 2007 test dataset, where only 18 categories are shown for our model learned on the FG-sCOCO training dataset as the categories of "diningtable" and "pottedplant" do not appear in the training set. The table shows a term called *ratio*, which is defined as the ratio of the number of occurrences for a category in the training set FG-COCO (FG-sCOCO training) to that in the training set MS COCO (sCOCO training), and describes the degree of how many objects in a coarse category have not been re-assigned to a corresponding fine-grained category with the semantic mapping. The ratio equal to 1 means that no object in MS COCO (sCOCO training) is re-assigned to a fine-grained category and its coarse label is kept in constructing FG-COCO (FG-sCOCO training). From the table, we observe that for most of the categories with the ratio close to 1, such as "car", "chair", "dog" and "train", the proposed LICN-E2C has better performance than the Faster R-CNN baseline. For these categories, the training examples are almost the same between FG-COCO (FG-sCOCO training) and MS-COCO (sCOCO training). The result demonstrates that our approach improves the label inference performance in the image detection problem. For the categories with a ratio much lower than 1, such as "aero" and "person", LICN has a lower performance than Faster-RCNN. We note that in this case, there is a large difference between the training sets for LICN and Faster R-CNN: FG-COCO (FG-sCOCO training) has a much larger label space and lower training examples for many categories than MS COCO (sCOCO training), which significantly increases the difficulty of label learning and inference and thus results in the drop of AP of

Table 3 Average precision (AP) (%) results of LICNs trained on FG-sCOCO training dataset

Method	Avg. Precision, IoU			Avg. Precision, Area		
	0.5:0.95	0.5	0.75	S	M	L
LICN-C2E	21.90	37.00	22.80	15.40	16.80	24.00
LICN-E2C	23.60	37.40	25.40	13.10	19.10	25.30

The results are reported on the FG-sCOCO test dataset

Table 4 Average precision (AP) (%) results for all the 20 categories of the Pascal VOC 2007 test dataset

	Method	Aero	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Diningtable	Dog	Horse	Motorbike	Person	Pottedplant	Sheep	Sofa	Train	TVmonitor	Mean
FG-COCO	ratio	0.48	0.89	0.98	0.90	0.96	0.96	1.00	0.99	1.00	0.99	1.00	1.00	1.00	0.94	0.69	1.00	0.95	1.00	0.96	0.92	
	Faster R-CNN[26]	**84.0**	**83.1**	**76.5**	**58.9**	**67.7**	**87.4**	77.1	85.6	61.0	**83.9**	**66.3**	78.4	86.3	86.6	**86.2**	50.9	81.7	68.1	86.1	**78.8**	**76.7**
	LICN-C2E	76.8	71.7	74.3	52.7	62.4	87.2	79.7	85.3	60.6	82.5	65.0	79.3	85.5	85.7	70.5	50.2	81.4	68.1	86.5	74.2	74.0
	LICN-E2C	71.6	69.9	75.2	52.9	64.1	87.0	**80.0**	**86.5**	**62.0**	83.6	65.6	**81.0**	**86.4**	**86.7**	68.2	**54.2**	**83.2**	**70.7**	**86.8**	76.9	74.6
FG-sCOCO	ratio	0.26	0.51	0.97	0.67	0.38	0.72	1.00	1.00	1.00	1.00	–	1.00	1.00	0.86	0.11	–	0.93	1.00	0.91	0.59	
	Faster R-CNN[26]	**77.4**	**79.7**	71.5	**58.9**	**52.3**	**85.2**	74.4	**86.3**	38.4	77.5	–	80.4	85.6	81.9	**83.9**	–	81.2	**64.1**	85.2	**64.3**	**73.8**
	LICN-C2E	69.9	76.1	68.6	50.9	41.8	81.4	73.4	85.9	37.3	74.4	–	78.3	84.0	81.2	46.3	–	76.1	63.7	84.1	59.6	68.5
	LICN-E2C	71.7	77.3	**73.8**	48.0	42.7	79.3	**75.4**	86.2	**39.4**	**79.5**	–	**80.5**	**86.2**	**82.7**	47.1	–	**81.4**	63.7	**86.1**	61.7	70.1

Faster R-CNN was trained on MS COCO and the sCOCO training dataset consisting of 80 coarse labels and 69 coarse labels, respectively, and LICN was trained on the FG-COCO dataset and FG-sCOCO training dataset with the expanded labels consisting of both the coarse and fine-grained labels. The score with bold font indicates the best result of the three compared methods for each category

LICN. It is noteworthy that our LICN-E2C achieves improvements of 0.6% and 1.6% compared with LICN-C2E with the training on FG-COCO and the FG-sCOCO training dataset, respectively. The results indicate that it is important to train the model in ascending order of CoS in improving the object detection performance.

4.3.3 Visual Genome

In this subsection, we evaluate the performance of our approach on the Visual Genome test dataset, where LICN and Faster R-CNN are trained on FG-COCO and MS COCO, respectively.

Figure 9 reports the comparison results of different methods on the test dataset in three cases: a) Fig. 9a shows the results for the fine-grained categories that do not appear in MS COCO and do come from the semantic mapping; b) Fig. 9b is for the coarse categories that have no corresponding fine-grained labels in the semantic mapping, i.e., the information for these categories in the training set MS COCO is the same as that in FG-COCO, and $ratio = 1$; and c) Fig. 9c is for the coarse categories, where different proportions of object samples with these category labels in MS COCO are re-labeled by new fine-grained labels based on semantic mapping in building FG-COCO, i.e., $ratio \in (0, 1)$. In Fig. 9a, we see that the proposed LICN-E2C performs better than LICN-C2E for some fine-grained categories, such as "guy", "fighter", "subway", "branch", "skateboarder", "wave", "tennis", "bear" and "television". The mean AP of the LICN-E2C model over all categories in Fig. 9a is 10.73, which achieves 0.61 mAP improvement over LICN-C2E (10.12). However, the Faster R-CNN baseline trained on the coarse categories cannot detect the new fine-grained categories, i.e., $AP = 0$ for these categories (the gray bars are not visible for that reason). From Fig. 9b, we can see that for those coarse categories that have not been re-annotated with fine-grained category labels, there are no obvious differences between these three models. As shown in Fig. 9c, for each coarse category in which a proportion of object samples have been re-annotated with fine-grained labels, Faster R-CNN has a better performance because its training dataset, i.e., MS COCO, consists of fewer categories and more examples in each of these categories than the training set of LICN. With the ratio decreasing, LICN obtains a little performance drop for the coarse categories. A direct reason is that the number of objects re-assigned from the coarse categories to the fine-grained categories increases in FG-COCO used for the training of LICN. For those coarse categories with a ratio close to 1, our LICN model can achieve a performance close to Faster R-CNN.

Actually, the problem of fine-grained label learning with the weak supervision of captions resolved by our approach is more challenging than the object detection and label inference resolved by the compared method, i.e., Faster R-CNN. The main reason is that the category space coming from captions in our problem (e.g., 150-dim as shown in Table 1) is much larger and consists of much more labeling noise than that in the latter problem.

4.3.4 Example illustrations

Figure 10 shows 5 fine-grained categories, namely "man", "woman", "plane", "bike" and "bat", predicted in object detection with our approach. For each category, we show 4 representative images with top confidence in category prediction. The illustration shows that our LICN approach can truly predict the fine-grained category labels with the weak supervision of captions.

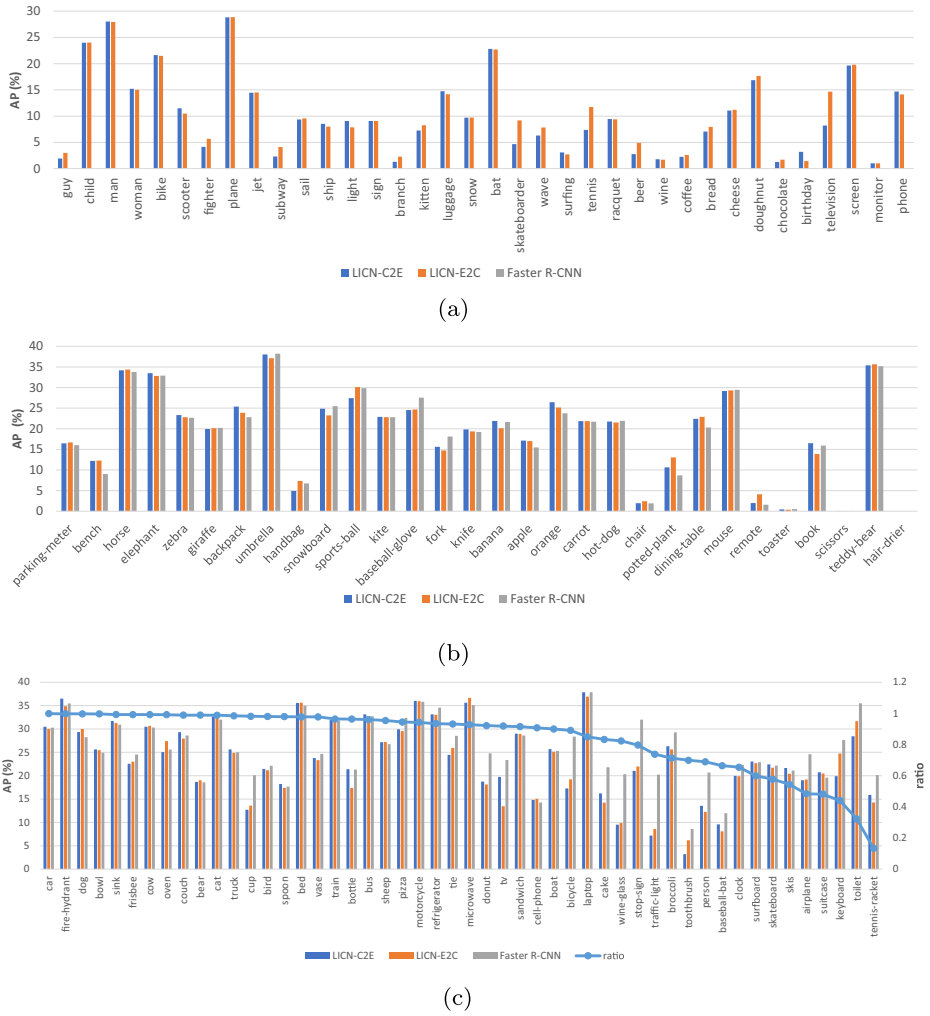

Fig. 9 The comparison of LICNs and Faster R-CNN, where the former is trained on FG-COCO and the latter on MS COCO. As introduced in Subsection 4.1.1, both datasets consist of the same images. The testing results are reported for the Visual Genome test dataset. (a) shows the results for the fine-grained categories whose labels are not appearing in MS COCO. (b) shows results for the coarse categories that have no corresponding fine-grained labels in the semantic map, i.e., $ratio = 1$. (c) shows the results for the coarse categories where different proportions of object samples are re-labeled by new fine-grained labels with semantic mapping, i.e., $ratio \in (0, 1)$

Figure 11 shows multiple failure cases of the categories of fruit, animals, etc. From this figure, we observe that the detected regions can attain a high accuracy, while the fine-grained label prediction does not work well. We consider that there are two possible factors resulting in this failure: the appearance of objects and the noise in introducing fine-grained labels. For example, for the top-left example, we incorrectly predict the object "apple" to the "orange" category due to the similar appearance between them.

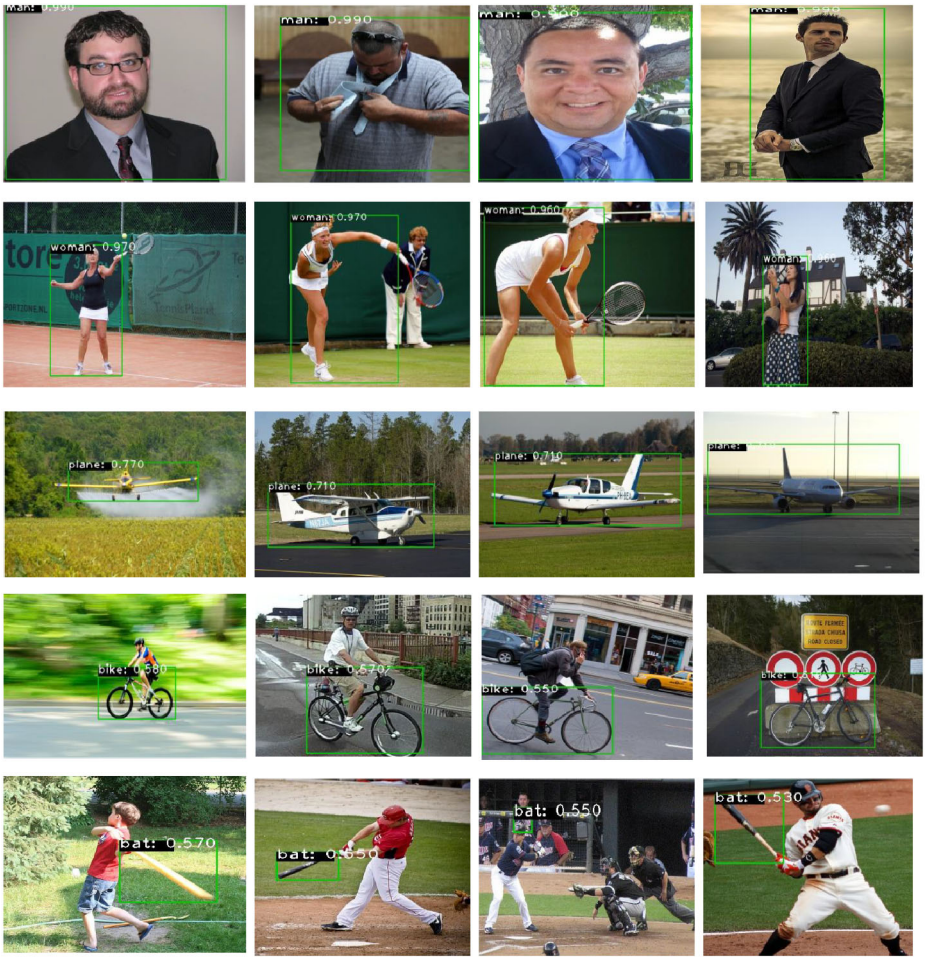

Fig. 10 Example illustration of 5 fine-grained categories: "man", "woman", "bike", "plane" and "bat", which correspond to the coarse categories: "person", "person", "airplane", "bicycle" and "baseball bat", respectively. The values next to bounding boxes indicate the confidences of fine-grained label prediction

5 Conclusion and future work

This paper seeks to answer the question of how to learn the fine-grained labels in object detection with the help of auxiliary information attached to images. In this paper, we propose a novel approach called label inference curriculum network (LICN) to the problem of fine-grained label learning with the weak supervision of captions. First, we construct a semantic mapping that builds a correspondence between the coarse category labels provided by public datasets and the fine-grained category labels extracted from captions based on the combination of embedding techniques and knowledge bases. Second, we present the label inference curriculum network with the consideration of the complexity of samples that describes the difficulty of fine-grained label learning. To evaluate the performance of fine-grained object label learning in different aspects, we construct multiple datasets based on widely-used

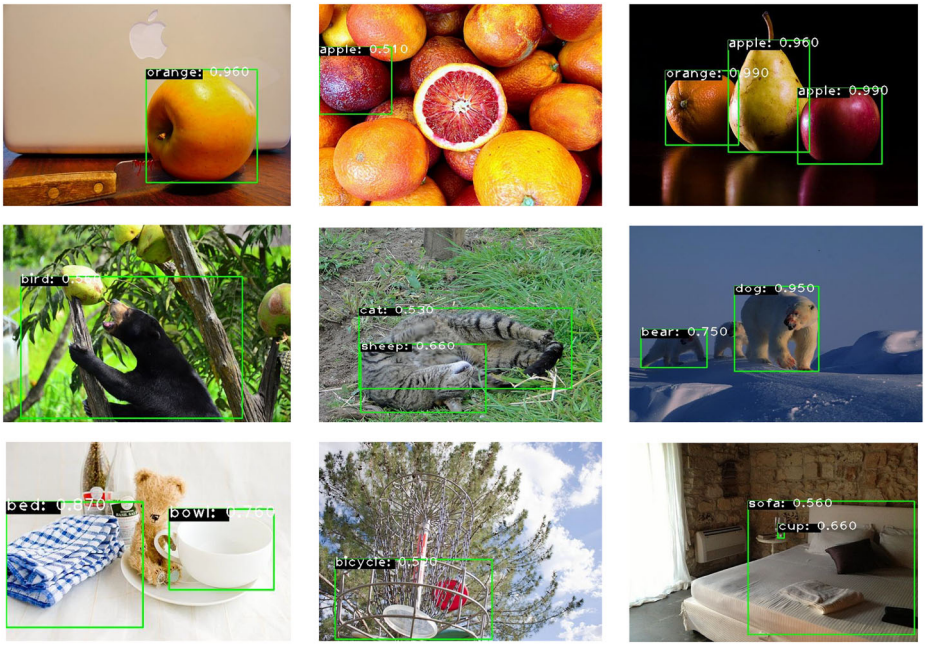

Fig. 11 Example illustration of failure cases

public datasets. Experimental results demonstrate the effectiveness of our proposed LICN model, and LICN-E2C achieves an improvement of 1.7% mAP with 0.5:0.05:0.95 IoU compared with the LICN-C2E on the FG-sCOCO test dataset. This improvement demonstrates that it is useful to consider the complexity of samples with curriculum learning in the fine-grained label learning. For the new fine-grained categories, LICN-E2C achieves the result of 10.73% mAP, while the Faster R-CNN baseline cannot work in this case. The experimental results show the effectiveness of our weakly-supervised learning approach to fine-grained label learning by considering the complexity of samples with the curriculum learning.

Acknowledgments This research is supported in part by China Scholarship Council (No. 20190628 0464), the National Key R&D Program (No. 2018AAA0101501) and the National Natural Science Foundation (61375040, 61772415), of China.

Funding China Scholarship Council (No. 201906280464), the National Key R&D Program (No. 2018AAA0101501) and the National Natural Science Foundation (61375040, 61772415), of China.

Data Availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of Interests The authors declare that there is no conflict of interests regarding the publication of this paper.

References

1. Ahmed A, Jalal A, Kim K (2021) Multi-objects detection and segmentation for scene understanding based on texton forest and kernel sliding perceptron. J Electr Eng Technol 16(2):1143–1150
2. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6077–6086
3. Bengio Y, Louradour J, Collobert R, Weston J (2009) Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning, pp 41–48
4. Bhujade S, Kamaleshwar T, Jaiswal S, Babu DV (2022) Deep learning application of image recognition based on self-driving vehicle. In: International conference on emerging technologies in computer engineering, Springer, pp 336–344
5. Bilen H, Vedaldi A (2016) Weakly supervised deep detection networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2846–2854
6. Buonviri A, York M, LeGrand K, Meub J (2019) Survey of challenges in labeled random finite set distributed multi-sensor multi-object tracking. In: 2019 IEEE Aerospace Conference, IEEE, pp 1–12
7. Diba A, Sharma V, Pazandeh A, Pirsiavash H, Van Gool L (2017) Weakly supervised cascaded convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 914–922. https://doi.org/10.1109/CVPR.2017.545
8. Du W, Phlypo R, Adalı T (2019) Adaptive feature selection and feature fusion for semi-supervised classification. J Signal Process Syst 91(5):521–537
9. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88(2):303–338
10. Fang H, Gupta S, Iandola F, Srivastava RK, Deng L, Dollár P, Gao J, He X, Mitchell M, Platt JC, et al. (2015) From captions to visual concepts and back. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1473–1482
11. Ge W, Yang S, Yu Y (2018) Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1277–1286
12. Guo S, Huang W, Zhang H, Zhuang C, Dong D, Scott MR, Huang D (2018) CurriculumNet: weakly supervised learning from large-scale web images. In: Proceedings of the european conference on computer vision (ECCV), pp 135–150
13. Hacohen G, Weinshall D (2019) On the power of curriculum learning in training deep networks. arXiv:190403626
14. Jerbi A, Herzig R, Berant J, Chechik G, Globerson A (2020) Learning object detection from captions via textual scene attributes. arXiv:200914558
15. Kantorov V, Oquab M, Cho M, Laptev I (2016) ContextLocNet: context-aware deep network models for weakly supervised localization. In: European conference on computer vision, Springer, pp 350–365
16. Krause J, Johnson J, Krishna R, Fei-Fei L (2017) A hierarchical approach for generating descriptive image paragraphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 317–325
17. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li LJ, Shamma DA, et al. (2017) Visual genome: connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis 123(1):32–73
18. Li C, Ma T, Zhou Y, Cheng J, Xu B (2017) Measuring word semantic similarity based on transferred vectors. In: International conference on neural information processing, Springer, pp 326–335
19. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: European conference on computer vision, Springer, pp 740–755
20. Manning CD, Surdeanu M, Bauer J, Finkel JR, Bethard S, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations, pp 55–60
21. Mikolov T, Chen K, Corrado G, Dean J (2013a) Efficient estimation of word representations in vector space. arXiv:13013781
22. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013b) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp 3111–3119
23. Misra I, Lawrence Zitnick C, Mitchell M, Girshick R (2016) Seeing through the human reporting bias: visual classifiers from noisy human-centric labels. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2930–2939

24. Oquab M, Bottou L, Laptev I, Sivic J (2015) Is object localization for free?-Weakly-supervised learning with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 685–694
25. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
26. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99
27. Song Y, Soleymani M (2019) Polysemous visual-semantic embedding for cross-modal retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1979–1988
28. Tang P, Wang X, Bai X, Liu W (2017) Multiple instance detection network with online instance classifier refinement. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2843–2851
29. Tang P, Wang X, Bai S, Shen W, Bai X, Liu W, Yuille A (2018) PCL: proposal cluster learning for weakly supervised object detection. IEEE Trans Pattern Anal Mach Intell 42(1):176–191
30. Teney D, Anderson P, He X, Van Den Hengel A (2018) Tips and tricks for visual question answering: learnings from the 2017 challenge. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4223–4232
31. Thomas C, Kovashka A (2019) Predicting the politics of an image using webly supervised data. In: Advances in neural information processing systems, pp 3630–3642
32. Tian Jl, Zhao W (2010) Words similarity algorithm based on Tongyici Cilin in semantic web adaptive learning system. J Jilin University (Inf Sci Ed) 28(06):602–608
33. Wan F, Wei P, Jiao J, Han Z, Ye Q (2018) Min-entropy latent model for weakly supervised object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1297–1306
34. Wang J, Wang X, Liu W (2018) Weakly- and semi-supervised Faster R-CNN with curriculum learning. In: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, pp 2416–2421
35. Wei Y, Shen Z, Cheng B, Shi H, Xiong J, Feng J, Huang T (2018) TS2C: tight box mining with surrounding segmentation context for weakly supervised object detection. In: Proceedings of the european conference on computer vision (ECCV), pp 434–450
36. Ye K, Zhang M, Kovashka A, Li W, Qin D, Berent J (2019) Cap2Det: learning to amplify weak caption supervision for object detection. In: Proceedings of the IEEE international conference on computer vision, pp 9686–9695
37. Zakraoui J, Saleh M, Al-Maadeed S, Jaam JM (2021) Improving text-to-image generation with object layout guidance. Multimed Tools Appl 80(18):27423–27443
38. Zhang M, Hwa R, Kovashka A (2018) Equal but not the same: understanding the implicit relationship between persuasive images and text. arXiv:180708205
39. Zhang X, Wei Y, Feng J, Yang Y, Huang TS (2018) Adversarial complementary learning for weakly supervised object localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1325–1334
40. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2921–2929