FoodLogoDet-1500: A Dataset for Large-Scale Food Logo Detection via Multi-Scale Feature Decoupling Network

Qiang Hou
School of Information Science and Engineering, Shandong Normal University
Shandong, China
2019309052@stu.sdnu.edu.cn

Weiqing Min*
Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
minweiqing@ict.ac.cn

Jing Wang
School of Information Science and Engineering, Shandong Normal University
Shandong, China
2018020875@stu.sdnu.edu.cn

Sujuan Hou*
School of Information Science and Engineering, Shandong Normal University
Shandong, China
sujuanhou@sdnu.edu.cn

Yuanjie Zheng
School of Information Science and Engineering, Shandong Normal University
Shandong, China
zhengyuanjie@gmail.com

Shuqiang Jiang
Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
sqjiang@ict.ac.cn

ABSTRACT

Food logo detection plays an important role in the multimedia for its wide real-world applications, such as food recommendation of the self-service shop and infringement detection on e-commerce platforms. A large-scale food logo dataset is urgently needed for developing advanced food logo detection algorithms. However, there are no available food logo datasets with food brand information. To support efforts towards food logo detection, we introduce the dataset FoodLogoDet-1500, a new large-scale publicly available food logo dataset, which has 1,500 categories, about 100,000 images and about 150,000 manually annotated food logo objects. We describe the collection and annotation process of FoodLogoDet-1500, analyze its scale and diversity, and compare it with other logo datasets. To the best of our knowledge, FoodLogoDet-1500 is the first largest publicly available high-quality dataset for food logo detection. The challenge of food logo detection lies in the large-scale categories and similarities between food logo categories. For that, we propose a novel food logo detection method Multi-scale Feature Decoupling Network (MFDNet), which decouples classification and regression into two branches and focuses on the classification branch to solve the problem of distinguishing multiple food logo categories. Specifically, we introduce the feature offset module, which utilizes the deformation-learning for optimal classification offset and can effectively obtain the most representative features of classification in detection. In addition, we adopt a balanced feature pyramid in MFDNet, which pays attention to global information, balances the multi-scale feature maps, and enhances feature extraction capability. Comprehensive experiments on FoodLogoDet-1500 and other two popular benchmark logo datasets demonstrate the effectiveness of the proposed method. The code and FoodLogoDet-1500 can be found at https://github.com/hq03/FoodLogoDet-1500-Dataset.

CCS CONCEPTS

• **Computing methodologies** → **Image representations**; **Object detection**.

KEYWORDS

food logo detection; food logo datasets; multi-scale; feature decoupling

ACM Reference Format:
Qiang Hou, Weiqing Min, Jing Wang, Sujuan Hou, Yuanjie Zheng, and Shuqiang Jiang. 2021. FoodLogoDet-1500: A Dataset for Large-Scale Food Logo Detection via Multi-Scale Feature Decoupling Network. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475289

1 INTRODUCTION

Logo detection research has always been extensively studied in the field of multimedia [9, 12, 22, 43, 45]. As one significant task in logo detection, food logo detection can be applied for healthy diet recommendation, food trademark infringement dispute, food advertising placement and supermarket self-checkout system. For example, with the rapid development of e-commerce platforms, many food businesses ignore copyright awareness in pursuit of profits, resulting in irreparable losses. Through food logo detection, we can avoid trademark infringement by detecting the new food logo and comparing their similarity with existing food logos. Furthermore, when we detect the logo from food products,

*Corresponding Author.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475289

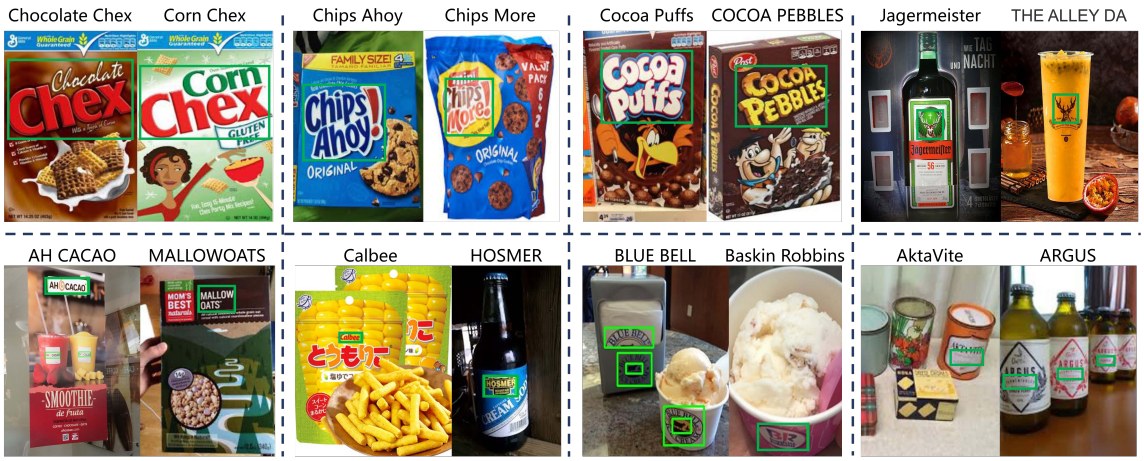

Figure 1: Some samples from FoodLogoDet-1500. Green boxes: ground-truth boxes.

we can further conduct various health-related and sell advertising applications [36].

Despite its great potential applications, food logo detection is still a challenging task, and the challenge mainly derives from two aspects:

- **There is a lack of large-scale food logo dataset for food logo detection.** Existing works mainly focus on messy logo datasets for logo detection, such as FlickrLogos-32 [46] and QMUL-OpenLogo [55]. For example, Romberg *et al.* [46] introduce the FlickrLogos-32 dataset with 32 categories but only very few classes belong to food logos. Su *et al.* [55] release one logo dataset with 352 categories, and full annotations. However, it is not all about food logos. Existing logo datasets only contain a tiny number of food logo categories. Therefore, they are probably not sufficient to construct more complicated deep learning models for food logo detection.
- **There are multi-scale and similar logos from food logo images, which are harder to detect in many cases.** Compare with other logo images, the multi-scale and similar food logos of the food logo images are more complicated, and make it difficult to accurately extract effective features. As shown in Fig. 1, the first row represents eight different food logo images. However, two classes of food logos are so similar that they are difficult to distinguish, such as 'Chips Ahoy' and 'Chips More'. Different brands of the same food may have similar food logos, which makes detection more difficult. Some food logos look like text, and they also have occlusion problems. Different food logos also have the characteristics of multi-scale, the second row illustrates this, such as 'MALLOW OATS' and 'Calbee'. These characteristics lead to the difficulty of food logo detection.

In this work, we address data limitations by building a large-scale dataset FoodLogoDet-1500 with 1,500 categories, 99,768 images and 145,400 objects. As the largest food logo detection dataset so far, FoodLogoDet-1500 brings great opportunities and challenges for food logo detection in general and sophisticated scenarios. To solve another challenge, we propose a Multi-scale Feature Decoupling Network (MFDNet) to improve food logo detection. This is achieved by two main modules named Feature Offset Module (FOM) and Balanced Feature Pyramid (BFP). FOM firstly decouples classification and regression into two branches, and then utilizes the deformation-learning for optimal classification offset. Finally, the optimal classification offset is merged with the original features of the network. The experiment proves that FOM can improve the classification accuracy in the food logo detection. In addition, we adopt BFP in MFDNet, which pays attention to global information and has a good performance on multi-scale food logos.

To summarize, our paper main contributions are as follow:

- We first introduce a large-scale and highly diverse food logo dataset FoodLogoDet-1500 with 1,500 categories, 99,768 images and 145,400 objects.
- We propose a Multi-scale Feature Decoupling Network for food logo detection by decoupling shared heads of classification and regression at the same time. In this network, we further introduce a balanced feature pyramid to ensure the detection of multi-scale food logos.
- We conduct extensive evaluation on three datasets, including FoodLogoDet-1500 and other two standard logo datasets QMUL-OpenLogo, FlickrLogos-32, and verify the effectiveness of our proposed method.

2 RELATED WORK

This section presents related work in the areas of logo datasets and logo detection.

2.1 Logo Datasets

The large-scale datasets play an indispensable role in current object detection algorithms, and it is no exception in food logo detection. In object detection, MS COCO [31] and PASCAL VOC [10] are the most commonly used datasets. In logo detection, FlickrLogos-32 [46] is the most popular dataset. However, it only consists of 32 logo categories with 70 images in each category. Similarly, Top-Logo-10 [54] contains fewer logo categories and fewer images. Logo-2K+ [60] belongs to image-level dataset and cannot be used for logo detection. QMUL-OpenLogo [55] consists of 352 logo categories, but it is

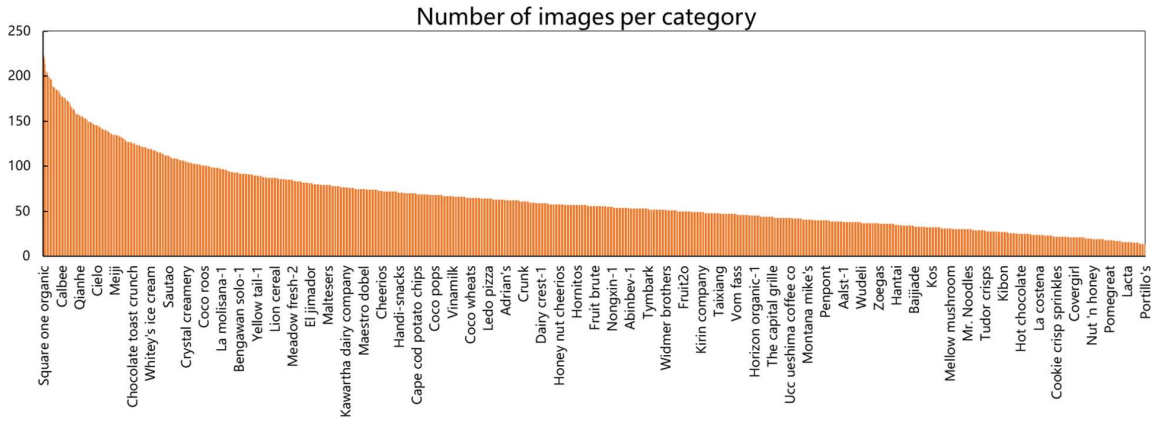

Figure 2: Sorted distribution of the number of images from each food logo in the FoodLogoDet-1500.

Table 1: Comparison between FoodLogoDet-1500 and existing logo datasets.

Dataset	♯Logos	♯Images	♯Objects	Availability
BelgaLogos [21]	37	10,000	-	Yes
FlickrLogos-27 [22]	27	1,080	4,671	Yes
FlickrLogos-32 [46]	32	8,240	5,644	Yes
Top-Logo-10 [54]	10	700	-	Yes
WebLogo-2M [53]	194	1,867,177	-	Yes
QMUL-OpenLogo [55]	352	27,083	-	Yes
Logos-in-the-Wild [58]	871	11,054	32,850	Yes
Logo-2K+ [60]	2,341	167,140	-	Yes
LogoDet-3K [59]	3,000	158,652	194,261	Yes
MICC-Logos [48]	13	720	-	No
FlickrBelgaLogos [26]	34	10,000	2,695	No
Logo-18 [17]	18	8,460	16,043	No
Logo-160 [17]	160	73,414	130,608	No
Logos-32plus [1]	32	7,830	12,302	No
Video SportsLogo [28]	20	2,000	-	No
CarLogo-51 [63]	51	11,903	-	No
Open Brands [20]	1,216	1,437,812	3,113,828	No
SynthLogo [35]	604	280,000	-	No
PL2K [11]	2,000	295,814	-	No
FoodLogoDet-1500	**1,500**	**99,768**	**145,400**	**Yes**

an all-encompassing logo dataset (e.g., Foods, Clothes, Necessities), and it can not be used for food logo detection. Wang *et al.* introduce LogoDet-3K [59], which has 3,000 categories, where only 932 classes belong to food logos. In order to promote the development of food logo detection in the multimedia community, we make supplementary improvements on food logos of LogoDet-3K. And then FoodLogoDet-1500 was built completely. In addition, some researchers construct other logo datasets, such as Logo-160 [17] and Open Brands [20]. These logo datasets are not currently available to the public, and are not helpful to the development of logo detection research.

Food logo detection is an important branch of logo detection. However, those have not publicly available food logo datasets with brand information at present. Therefore, we introduce a new large-scale food logo dataset FoodLogoDet-1500 with 1,500 food logo categories. Table 1 summarizes the statistics of existing logo datasets and FoodLogoDet-1500. To the best of our knowledge, FoodLogoDet-1500 is the first largest publicly available high-quality dataset for food logo detection. It helps to promote the development of food logo detection research.

2.2 Logo Detection

Typically, logo detection is performed by adapting object detection methods in the domain of commercial logos [18] (i.e., treating each logo as a different object or class). Traditionally, the use of handcrafted features, such as SIFT [34] and textures [14], along with statistical classifiers, such as Support Vector Machines (SVM) [7], have been the main approaches for object detection. In the last few years, deep learning has shown its good performance in object detection, and Convolutional Neural Networks (CNNs) [25, 49] was the core element of deep learning methods. Multiple methods for object detection using CNNs have been presented. In general, object detectors could be divided into two types: two-stage detector and one-stage detector. Two-stage means that the object detection algorithm needs to be completed in two steps. First, candidate regions need to be obtained and then classified, such as R-CNN series like Fast R-CNN [13] and Faster R-CNN [42]. On the other hand, one-stage detector, which can be understood as one-step detection, does not need to search for candidate regions separately, typically including SSD [33] and YOLO [2, 39–41]. Recently, anchor-free methods [24] and transformers [5] for object detection are widely used.

Multi-scale feature fusion is one of the most important research hotspots in deep networks. Low-level features generally lack semantic information but rich in keeping geometric details, which is the opposite for high-level features. FPN [29] first built a top-down architecture with lateral connections to extract features across multiple layers. PANet [32] directly created a short path for low-level feature maps since detecting large objects also needs the assistance of location-sensitive feature maps. Libra R-CNN [38] improved the level of feature fusion by adding a non-local block to fine-tune the combined feature maps.

Detection heads are also one of the focuses of research. Mask R-CNN [15] brought in an extra head for instance segmentation.

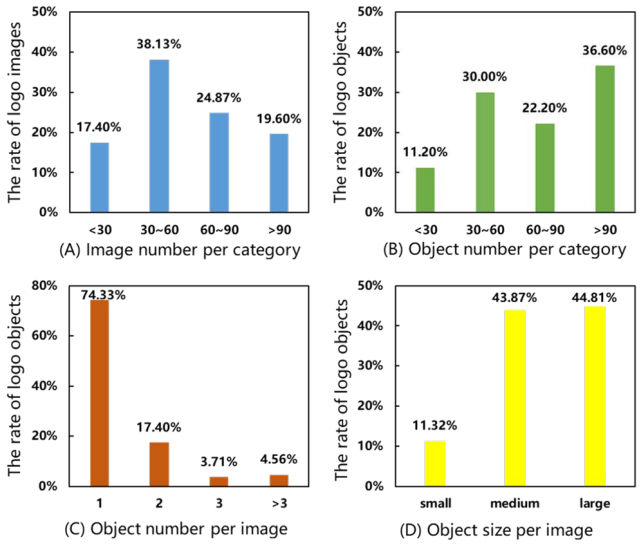

Figure 3: The detailed statistics of FoodLogoDet-1500.

IoU-Net [19] introduced a branch to predict IoUs between detected bounding boxes and their corresponding ground truth boxes. FCOS [57] added a single-layer branch, which is parallel to the classification branch, to predict the centerness position. Double-Head R-CNN [62] proposed to disentangle the sibling head into two independent branches for classification and localization. Song *et al.* [52] also dealt with the classification branch and the regression branch and obtained a relatively good detection result.

Different from these works, our work decouples classification and regression into two branches and focuses on the classification branche to solve the problem of distinguishing multiple food logo categories. At the same time, we consider the multi-scale and similar characteristics of food logos, and use multi-layer features for food logo detection.

3 FOODLOGODET-1500

In order to obtain a high-quality food logo dataset with high diversity and high coverage. We build FoodLogoDet-1500 from the following three steps: **(1) Constructing the Food Logo Category List.** In order to guarantee wide coverage of the food logo category list, we resort to the widely used shopping application Taobao and Jingdong, also with Wikipedia to construct the food logo category list. **(2) Collecting Food Logo Images.** Using a query term from the constructed food logo category list, we crawled candidate images from various search engines (i.e., Google, Bing and Baidu) for broader coverage and higher diversity of food logo images compared with other datasets from only one data source. At the same time, we also added some scene words to ensure more complexity and better diversity of the captured food logo images, such as Coca Cola + Supermarket and Heineken + Bar. **(3) Cleaning and Labeling Food Logo Images.** We checked each category manually to ensure that each image contained the corresponding food logo. It is worth noting that we focused on the food logo rather than the food itself. We also deleted repetitive images and images with incomplete RGB channels. Labeling is not only the most important step in creating a dataset, but also the most complicated step. Every food logo object needs to be annotated, regardless of which image it is placed on. We kept the low-resolution, incomplete food logo images to enhance the challenge of the dataset. After labeling, we then conduct manual verification by crowd-sourcing the task to 13 Lab members. In addition, a food brand may have two or more different types of logos, such as graphic logos and textual logos. We treat different logo variations of the same brand as distinct food logo classes, similar to [58]. Note that the suffix '-1', '-2' is added to the logo name as the new logo category, such as 'Maruchan-1' presents the 'Maruchan' graphic logo while 'Maruchan-2' presents its textual logo for the brand 'Maruchan'.

After completing the construction of the FoodLogoDet-1500, in order to show the details of our dataset, we provide the statistics at the category levels. Fig. 2 shows sorted distribution of the number of images from sampled classes, we can see that imbalanced distribution across different food logo categories are one characteristic of FoodLogoDet-1500, posing a challenge for effective food logo detection with few samples. In addition, we also conduct data statistics on images and objects in the FoodLogoDet-1500 as shown in Fig. 3. Fig. 3 (A) shows the distribution of the number of images for each category, where each category represents each food logo. Fig. 3 (B) shows the distribution of the number of objects for each category. As we can see, there is the imbalance between images and objects in different food logo categories. Fig. 3 (C) provides the number of objects per image. We can draw the conclusion that most images contain one or two logo objects, which is similar to what happens in our real world. Fig. 3 (D) gives the number of objects size in each image. In FoodLogoDet-1500, the large percentage of small and medium food logo objects (~ 56%) will pose another challenge to food logo detection, since smaller food logos are harder to detect.

4 METHODOLOGY

In this section, we will introduce the proposed Multi-scale Feature Decoupling Network (MFDNet) for food logo detection. Fig. 4 illustrates the architecture of MFDNet, which contains two main components, namely Balanced Feature Pyramid (BFP) and Feature Offset Module (FOM). Specifically, the features of one input food logo image are extracted by ResNet-50 [16]. Then FPN is employed to fuse multi-scale features and BFP is used for feature refinement in the feature maps. Feature fusion and feature refinement are more effective for multi-scale food logo detection. The region proposal generation step yields a set of region of interests (RoIs) using Region Proposal Network (RPN). Then the RoIs are fed into RoI Pooling layer, in which each RoI is pooled into a fixed-size feature map. Finally, it is divided into classification and regression branches by feature decoupling. FOM is used to disentangle the classification and regression. In the classification branch, FOM utilizes the deformation-learning for optimal offset, which helps us to obtain the most representative features of classification in food logo detection. Then the optimal classification offset is merged with the original features of the network. Finally, feature maps are mapped to a feature vector by a fully connected layer (FC), which is followed by training the final object classifiers and bounding box regressors.

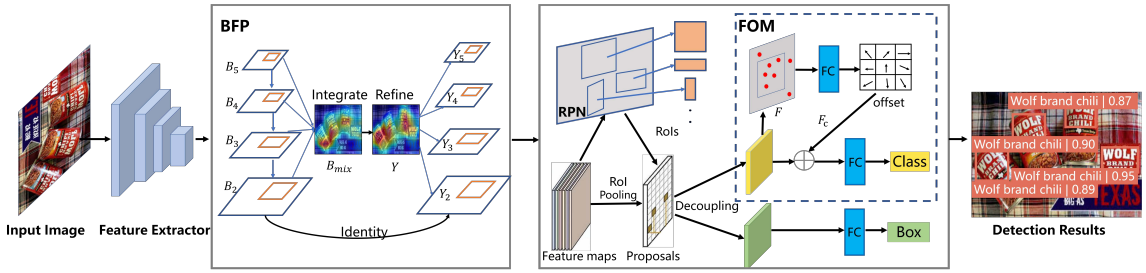

Figure 4: Overview of proposed Multi-scale Feature Decoupling Network (MFDNet) for food logo detection. BFP: Balanced Feature Pyramid. FOM: Feature Offset Module. RPN: Region Proposal Network. FC: Full-Connected layer.

Next, we will focus on two main modules in the MFDNet, namely BFP and FOM.

4.1 BFP

In object detection, multi-scale features fusion has been a hot topic of research. Deep high-level features in backbones are with more semantic information while the shallow low-level features are more descriptive content [65]. On that basis, we fuse BFP into MFDNet to better integrate multi-scale features. Different from former methods that integrate multi-level features using lateral connections, the BFP uses the same deeply feature maps to integrate balanced semantic features to strengthen the multi-level features.

To integrate multi-level features while maintaining their semantic hierarchy, we first adjust the multi-level features of FPN outputs $\{B_2, B_3, B_4, B_5\}$ to the same size as B_4. This is achieved using interpolation and maxi-pooling on the other levels to prepare for integration. And then the integrated semantic information is obtained by Eq. 1.

$$B_{mix} = \frac{1}{L} \sum_{l_{min}}^{l_{max}} B_l \tag{1}$$

where B_l is l-th feature maps. The number of multi-level features are denoted as L. l_{min} and l_{max} are the lowest and highest feature levels, respectively.

Then, the BFP uses non-local module to further refine balanced semantic features. The refining step helps us to enhance the integrated features which are more discriminative. The non-local module is adopted as follows:

$$Y_i = \frac{1}{C(x)} \sum_{\forall j} f(B^i_{mix}, B^j_{mix}) g(B^j_{mix}) \tag{2}$$

where i is the index of an output position whose response is to be computed and j is the index that enumerates all possible positions in feature map B_{mix}. Y is the output of the same size as B_{mix}. $f(\cdot,\cdot)$ computes a scalar between i and all j. $g(\cdot)$ computes a representation of the input at the position j. $C(x)$ is the normalization parameter.

After BFP, we can use the feature information of different layers more effectively.

4.2 FOM

The challenge of food logo detection lies in the large-scale categories and similar food logos. Thus, we focus on the problem of classification on food logo detection with large-scale categories. In food logo detection, we are more inclined to extract more expressive semantic regional features in images for large-scale food logo classification. As shown in Fig. 4, different from the original detection head, in FOM, we propose an auto-learned anchor region proposal network for pixel wise offset. FOM is used to search for the best feature extraction for food logo classification.

We used the deformable learning manner to achieve this goal. As shown in Fig. 4, F is the output feature map of the RoI pooling layer. RoI pooling divides the RoI into $k \times k$ bins and output a $k \times k$ feature map F_c. From the F, a fully connected layer generates the normalized offsets $\triangle\widehat{C}_{ij}$ which are then transformed to the offsets $\triangle C_{ij}$ by element-wise product with the RoI's width and height by Eq. 3. For (i, j)-th bin, the translation $\triangle C$ is performed on the sample points in it to obtain the new sample points for F_c. This procedure can be formulated as follows:

$$\triangle C_{ij} = \alpha \triangle \widehat{C}_{ij} \cdot (w, h) \tag{3}$$

where α is a predefined scalar to modulate the magnitude of the $\triangle C_{ij}$, and (w, h) is the width and height of F.

For generating feature maps by irregular F_c, we use the deformable RoI pooling as:

$$F_c(i,j) = \sum_{p \in bin(i,j)} \frac{G(p_0,\ p_0 + p_n + \triangle C_{ij})}{n_{ij}} \tag{4}$$

where p_0 is the top-left corner and p_n enumerates all integral spatial locations in the feature map. n_{ij} is the number of pixels in the bin. As the offset $\triangle C_{ij}$ is typically fractional, $G(\cdot,\cdot)$ is implemented via the bilinear interpolation.

By disentangling the shared proposal for the classification and regression, FOM is used to search for the best feature for food logo classification. It allows classification tasks to adaptively seek the optimal location in space. This has excellent detection accuracy for large-scale categories and similar food logos.

4.3 Loss Function

In MFDNet, the final loss function is as follows:

$$L = L_{rpn} + L_{cls} + L_{loc} + L_{FOM} \tag{5}$$

where L_{rpn} , L_{cls} and L_{loc} are the losses for RPN, classification and localization, respectively. L_{FOM} is the loss for FOM.

Among them, the loss function of the FOM is as follows:

$$L_{FOM} = L_F(C(f(F, F_c)), y) \tag{6}$$

where L_F is achieved through the cross-entropy loss function. $f(\cdot)$ is the feature extractor and $C(\cdot)$ is a function for transforming features to predict specific category. y is the logo category.

The cross-entropy classification loss function is adopted as follows:

$$L_F = -\frac{1}{N}\sum_{i}^{M}\log y_{ic}\left(p_{ic}\right) \tag{7}$$

where N is the number of training samples and M is number of food logo categories. y_{ic} is the indicator variable. If the sample category is the same as the category of sample i, then y_{ic} is 1. p_{ic} is the probability which predict the whether a sample belongs to category c.

5 EXPERIMENT

5.1 Experimental Setup

Dataset and evaluation metrics. To evaluate the effectiveness of the proposed MFDNet, we conduct extensive experiments on our introduced FoodLogoDet-1500 and two standard logo detection datasets, including the FlickrLogos-32 and the QMUL-OpenLogo.

For evaluation, we adopt the widely used mean Average Precision (mAP) [10] and the IoU threshold is 0.5, which means that a detection will be considered as positive if the IoU between the predicted box and ground-truth box exceeds 50%. We also use AP_{25} and AP_{75} as evaluation standards, which represent the IoU threshold of 0.25 and 0.75, respectively.

Implementation details. We implement our method based on the publicly available mmdetection toolbox [6]. The Double-Head R-CNN based on ResNet-50 is adopted as the baseline network.

In our experiment, the base detection networks are trained with stochastic gradient descent (SGD). The input images are resized to 800×600 pixels. We train detectors end-to-end with 2 GPUs (2 images per GPU) for 12 epochs. The initial learning rate is set to 2.5×10^{-3}. The weight decay of 0.0001 and the momentum of 0.9 are used. Other hyperparameters follow the settings in mmdetection unless otherwise specified.

Table 2: Evaluation on individual modules and two modules of MFDNet on FoodLogoDet-1500 (%).

FOM	BFP	mAP	AP_{25}	AP_{75}
		84.5	84.4	82.1
✓		86.0$\uparrow_{1.5}$	85.9$\uparrow_{1.5}$	84.4$\uparrow_{2.3}$
	✓	85.1$\uparrow_{0.6}$	85.0$\uparrow_{0.6}$	83.1$\uparrow_{1.0}$
✓	✓	**86.6**$\uparrow_{2.1}$	**86.4**$\uparrow_{2.0}$	**85.0**$\uparrow_{2.9}$

5.2 Experiment on FoodLogoDet-1500

To enable the benchmark research, we follow the standard setup for data partitions in our experiments. 80%, 20% of images are randomly selected for training and testing in each food logo category.

Ablation Study. For the ablation study, we conduct a comprehensive analysis of the effects of two modules from the MFDNet in the FoodLogoDet-1500. Table 2 shows an ablation study on the effects of different combinations of FOM and BFP in the FoodLogoDet-1500. Two modules are added to Double-Head R-CNN, and the results respectively improve the mAP by 1.5%, 0.6% and 2.1%. These results prove the effectiveness of the FOM in large-scale food logo dataset.

Next, we perform the visualization of the ablation study and analyze the existing detail problem in the Double-Head R-CNN. And then we provide more typical examples in Fig. 5, including the regression bounding box and the classification accuracy. The red box represents the prediction box and the green box is the ground-truth box. Clearly, MFDNet can accurately detect objects with occluded, ambiguous and smaller cases, and it obtains more accurate bounding box regression and classification score. The Double-Head R-CNN makes some detection mistakes, such as misclassified similar food logo categories, and mistaking words into food logo categories. Such as the word 'Good' is used as a food logo 'coors', because two words are similar. The reason for the above errors is that the large-scale categories of the food logo and the similarity between the food logos are not considered. In contrast, for the detected logos in the middle two images in Fig. 5, our method has an advantage in classification accuracy, and also detect smaller food logos. This shows that FOM can search the best feature for classification, and BFP can fusion of multi-scale information for detection.

Table 3: Performance comparison on FoodLogoDet-1500 (%).

Method	mAP	AP_{25}	AP_{75}
Faster R-CNN [42]	83.9	83.8	81.7
RetinaNet [30]	77.3	77.0	75.3
DCN [8]	85.2	85.1	84.2
Cascade R-CNN [4]	83.5	83.3	83.2
PANet [32]	83.8	83.6	81.5
Libra R-CNN [38]	77.8	77.7	76.4
FSAF [68]	83.0	83.5	81.0
Dynamic R-CNN [66]	75.9	75.5	65.6
Sparse R-CNN [56]	83.1	-	-
SABL [61]	82.9	83.4	82.0
GRoIE [47]	83.4	83.6	82.1
Generalized Focal Loss [27]	79.2	79.1	78.5
Double-Head R-CNN [62]	84.5	84.4	82.1
ATSS [67]	80.2	80.0	79.8
FoveaBox [23]	75.2	75.0	74.0
Soft-NMS [3]	83.8	83.9	81.8
OHEM [50]	84.1	84.2	82.5
Iou loss [64]	82.2	82.1	79.5
Generalized IoU [44]	83.3	83.2	80.4
SSD [33]	80.4	80.1	78.6
MFDNet	**86.6**	**86.4**	**85.0**

Comparisons with State-of-the-Arts. In this subsection, we compare the results of our method with other works in FoodLogoDet-1500. Table 3 summarizes the clear performance superiority of MFDNet over all state-of-the-arts with significant mAP, AP_{25} and AP_{75} improvement. SSD uses VGG-16 [51] as a backbone, and

Figure 5: Visualization comparison between Double-Head R-CNN and MFDNet on the FoodLogoDet-1500. The first row is Double-Head R-CNN, the second row is MFDNet. Green boxes: ground-truth boxes. Orange boxes: correct detection boxes. Yellow boxes: mistakes detection boxes.

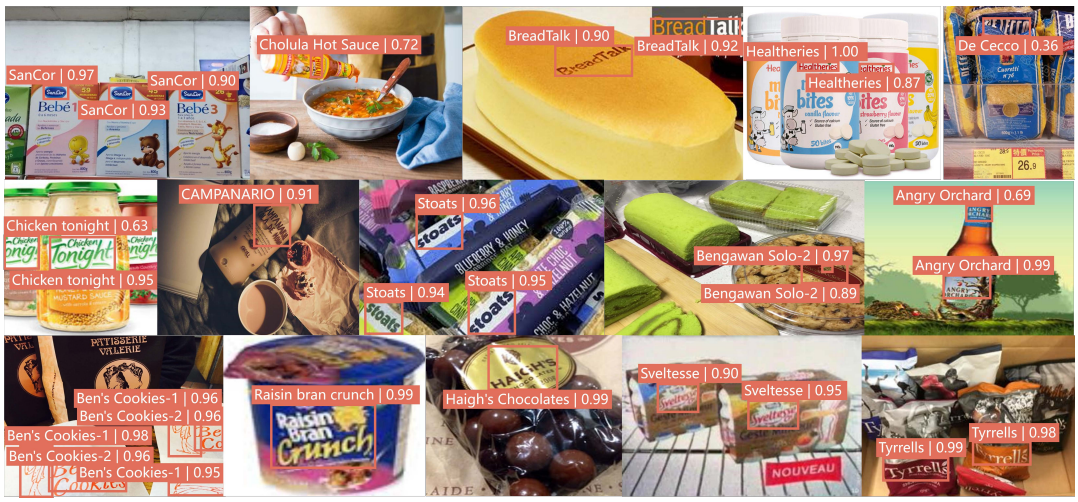

Figure 6: Detection results for our proposed MFDNet on the FoodLogoDet-1500. Orange boxes: correct detection boxes.

other detection models adopt ResNet-50. Compared with existing baselines RetinaNet, Faster R-CNN and Double-Head R-CNN, etc., the proposed method significantly outperforms these state-of-the-art methods. RetinaNet had the limitation in multi-scale object detection. Faster R-CNN adopted the same parameters in two different tasks, which missed the conflict between them in the sibling head, especially the classification of large-scale datasets. MFDNet achieves the best performance with 86.6% mAP. Compared with Double-Head R-CNN, our proposed method gains 2.1% mAP. Specifically, it gains 2.0% AP_{25} and gains 2.9% AP_{75}. This shows that our proposed method still has good performance with the change of the IoU threshold. MFDNet also surpasses methods of two-stage detectors (Faster R-CNN, Sparse R-CNN and GRoIE), boosting the mAP by 2.7%, 3.5% and 3.2%, respectively. These results validate the advantage of our feature decoupling over existing methods. Some detection results of MFDNet are given in Fig. 6, including the regression bounding box and the classification accuracy. The red box represents the prediction box.

We also set different iterations to compare the convergence and accuracy of models. Fig. 7 shows higher performance with increasing iterations. It can be seen that our method converges at about 100,000 iterations and keeps higher accuracy than Double-Head R-CNN in the training process. This shows that feature decoupling can speed up model convergence.

5.3 Experiment on Other Benchmarks

Besides FoodLogoDet-1500, we also conduct the evaluation on other publicly available benchmark datasets, QMUL-OpenLogo and FlickrLogo-32 to further verify the effectiveness of our method. QMUL-OpenLogo contains 27,083 images from 352 logo categories. In each logo category, 70%, 30% of images are randomly selected for training and testing, respectively [55]. FlickrLogos-32 consists

of 2,240 images from 32 logo categories. 80%, 20% of images are randomly selected for training and testing for each logo category. Considering that these baseline experiments only used mAP as the evaluation metric, we also used mAP as the evaluation standard for comparison.

Experiments on QMUL-OpenLogo. We list the experimental results of baselines and our proposed method in Table 4. Our proposed method achieves the best performance with 51.3% mAP. Specifically, MFDNet outperforms the baseline model by 0.4% in mAP. Compared to the anchor-free method FSAF, our method improves the mAP by 6.6%. These results demonstrate the universality of our method on the large-scale logo dataset.

Table 4: Performance comparison on QMUL-OpenLogo (%).

Methods	mAP
YOLO9000 [40]	26.3
ATSS [67]	48.4
Faster R-CNN [42]	51.2
Libra R-CNN [38]	51.2
FSAF [68]	44.7
Dynamic R-CNN [66]	51.2
FoveaBox [23]	35.6
Generalized Focal Loss [27]	46.6
Sparse R-CNN [56]	46.9
Double-Head R-CNN [62]	50.9
MFDNet	**51.3**

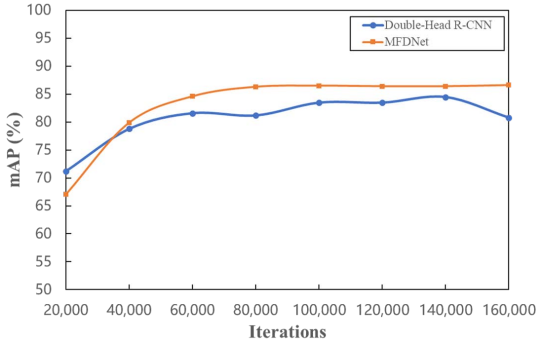

Figure 7: The comparison of MFDNet and Double-Head R-CNN with increasing iterations.

Experiments on FlickrLogos-32. To better prove the effectiveness of our method, we also carry out experiments on the FlickrLogos-32 with fewer images. Table 5 shows that MFDNet still achieves the best performance compared with other methods, surpassing the Double-Head R-CNN by 0.9% in mAP. However, our model achieves a small margin 0.1% in mAP over the best method Generalized Focal Loss. The probable reason is that FlickrLogos-32 contains fewer logo images and the FOM module thus does not play a decisive role in the dataset.

Table 5: Performance comparison on FlickrLogos-32 (%).

Methods	mAP
Bag of Words (BoW) [45]	54.5
Deep Logo [18]	74.4
BD-FRCN-M [37]	73.5
YOLO [39]	68.7
YOLOv3 [41]	71.7
RetinaNet [30]	78.4
Faster R-CNN [42]	83.5
Libra R-CNN [38]	84.6
Dynamic R-CNN [66]	85.8
FoveaBox [23]	84.1
Generalized Focal Loss [27]	86.2
Sparse R-CNN [56]	73.7
Double-Head R-CNN [62]	85.3
MFDNet	**86.2**

5.4 Discussion

Compared with existing methods, our proposed method obtains better detection performance, especially in solving small food logo objects and large-scale classification. However, it can not achieve high detection performance in some cases. In Fig. 5, the fourth image in the second row shows although the MFDNet improves the detection accuracy of small food logos, there are also missed detections for smaller objects. Therefore, the food logo detection on FoodLogoDet-1500 still has great challenges, such as the problem of the small food logo objects. And it meanwhile highlights the comparative difficulty of the FoodLogoDet-1500.

6 CONCLUSIONS

In this paper, we present a new large-scale dataset FoodLogoDet-1500, which is currently the first and largest publicly available food logo detection dataset to the best of our knowledge. In the future, we hope FoodLogoDet-1500 will become a new benchmark food logo dataset, and provide convenience for food logo detection. We then propose a Multi-scale Feature Decoupling Network for food logo detection. Extensive evaluation on FoodLogoDet-1500 and another two standard benchmark logo datasets have verified its effectiveness.

With the rapid development of e-commerce platforms and major food brands, food logo detection will become the trend of future research. We will continue to explore the characteristics of the FoodLogoDet-1500, and generate different benchmarks to evaluate its challenges, such as tiny food logo, serious occlusion and low resolution. Furthermore, we will use transformer [5] and lightweight methods to achieve faster and more accurate performance for food logo detection.

ACKNOWLEDGMENTS

This work was supported in part by the National Nature Science Foundation of China (62072289, 61702313, and 61972378), in part by Postdoctoral Science Foundation of China (2017M612338), in part by Shandong science and technology plan project (J17KB177).

REFERENCES

[1] Simone Bianco, Marco Buzzelli, Davide Mazzini, and Raimondo Schettini. 2017. Deep learning for logo recognition. *Neurocomputing* 245 (2017), 23–30.

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* (2020).

[3] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. 2017. Soft-NMS–improving object detection with one line of code. In *Proceedings of the IEEE International Conference on Computer Vision*. 5561–5569.

[4] Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade R-CNN: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6154–6162.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*. Springer, 213–229.

[6] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. 2019. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155* (2019).

[7] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 764–773.

[9] Christian Eggert, Anton Winschel, and Rainer Lienhart. 2015. On the benefit of synthetic data for company logo detection. In *Proceedings of the ACM International Conference on Multimedia*. 1283–1286.

[10] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*. (2010), 303–338.

[11] István Fehérvári and Srikar Appalaraju. 2019. Scalable logo recognition using proxies. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. IEEE, 715–725.

[12] Yue Gao, Fanglin Wang, Huanbo Luan, and Tat-Seng Chua. 2014. Brand data gathering from live social media streams. In *Proceedings of the International Conference on Multimedia Retrieval*. 169–176.

[13] Ross Girshick. 2015. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*. 1440–1448.

[14] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. 1973. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics* 6 (1973), 610–621.

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*. 2961–2969.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[17] Steven C. H. Hoi, Xiongwei Wu, Hantang Liu, Yue Wu, Huiqiong Wang, Hui Xue, and Qiang Wu. 2015. LOGO-Net: Large-scale Deep Logo Detection and Brand Recognition with Deep Region-based Convolutional Networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 46, 5 (2015), 2403–2412.

[18] Forrest N. Iandola, Anting Shen, Peter Gao, and Kurt Keutzer. 2015. DeepLogo: Hitting Logo Recognition with the Deep Neural Network Hammer. *arXiv preprint arXiv:1510.02131* (2015).

[19] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. 2018. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European Conference on Computer Vision*. 784–799.

[20] Xuan Jin, Wei Su, Rong Zhang, Yuan He, and Hui Xue. 2020. The Open Brands Dataset: Unified brand detection and recognition at scale. In *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 4387–4391.

[21] Alexis Joly and Olivier Buisson. 2009. Logo retrieval with a contrario visual query expansion. In *Proceedings of the ACM International Conference on Multimedia*. 581–584.

[22] Yannis Kalantidis, Lluis Garcia Pueyo, Michele Trevisiol, Roelof van Zwol, and Yannis Avrithis. 2011. Scalable triangulation-based logo recognition. In *Proceedings of the ACM International Conference on Multimedia Retrieval*. 1–7.

[23] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. 2020. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing* 29 (2020), 7389–7398.

[24] Hei Law and Jia Deng. 2018. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision*. 734–750.

[25] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[26] Pierre Letessier, Olivier Buisson, and Alexis Joly. 2012. Scalable mining of small visual objects. In *Proceedings of the ACM International Conference on Multimedia*. 599–608.

[27] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. 2020. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. *arXiv preprint arXiv:2006.04388* (2020).

[28] Yuan Liao, Xiaoqing Lu, Chengcui Zhang, Yongtao Wang, and Zhi Tang. 2017. Mutual enhancement for detection of multiple logos in sports videos. In *Proceedings of the IEEE International Conference on Computer Vision*. 4846–4855.

[29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2117–2125.

[30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 2980–2988.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceeding of the European Conference on Computer Vision*. Springer, 740–755.

[32] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. 2018. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8759–8768.

[33] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. SSD: Single Shot Multibox Detector. In *European Conference on Computer Vision*. Springer, 21–37.

[34] David G Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE International Conference on Computer Vision*, Vol. 2. Ieee, 1150–1157.

[35] MD Mas, L. Qian, A. Jan, and E. J. Delp. 2018. Logo detection and recognition with synthetic images. *Electronic Imaging* 2018, 10 (2018), 3371–3377.

[36] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. 2019. A survey on food computing. *In ACM Computing Surveys* 52, 5 (2019), 1–36.

[37] Gonçalo Oliveira, Xavier Frazão, André Pimentel, and Bernardete Ribeiro. 2016. Automatic graphic logo detection via Fast Region-based Convolutional Networks. In *International Joint Conference on Neural Networks*. 985–991.

[38] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. 2019. Libra R-CNN: Towards balanced learning for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 821–830.

[39] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 779–788.

[40] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7263–7271.

[41] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* (2015).

[43] Jerome Revaud, Matthijs Douze, and Cordelia Schmid. 2012. Correlation-Based Burstiness for Logo Retrieval. In *Proceedings of the ACM International Conference on Multimedia*. 965–968.

[44] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 658–666.

[45] Stefan Romberg and Rainer Lienhart. 2013. Bundle min-hashing for logo recognition. In *Proceedings of the ACM Conference on International Conference on Multimedia Retrieval*. 113–120.

[46] Stefan Romberg, Lluis Garcia Pueyo, Rainer Lienhart, and Roelof Van Zwol. 2011. Scalable logo recognition in real-world images. In *Proceedings of the ACM International Conference on Multimedia Retrieval*. 1–8.

[47] Leonardo Rossi, Akbar Karimi, and Andrea Prati. 2020. A novel region of interest extraction layer for instance segmentation. *arXiv preprint arXiv:2004.13665* (2020).

[48] Hichem Sahbi, Lamberto Ballan, Giuseppe Serra, and Alberto Del Bimbo. 2012. Context-dependent logo matching and recognition. *IEEE Transactions on Image Processing* 22, 3 (2012), 1018–1031.

[49] Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural networks* 61 (2015), 85–117.

[50] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. 2016. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 761–769.

[51] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of the International Conference on Learning Representations*. 1–14.

[52] Guanglu Song, Yu Liu, and Xiaogang Wang. 2020. Revisiting the sibling head in object detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11563–11572.

[53] Hang Su, Shaogang Gong, and Xiatian Zhu. 2017. Weblogo-2m: Scalable logo detection by deep learning from the web. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 270–279.

[54] Hang Su, Xiatian Zhu, and Shaogang Gong. 2017. Deep learning logo detection with data expansion by synthesising context. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. IEEE, 530–539.

[55] Hang Su, Xiatian Zhu, and Shaogang Gong. 2018. Open logo detection challenge. *arXiv preprint arXiv:1807.01964* (2018).

[56] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. 2020. Sparse R-CNN: End-to-end object detection with learnable proposals. *arXiv preprint arXiv:2011.12450* (2020).

[57] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 9627–9636.

[58] Andras Tüzkö, Christian Herrmann, Daniel Manger, and Jürgen Beyerer. 2017. Open set logo detection and retrieval. *arXiv preprint arXiv:1710.10891* (2017).

[59] Jing Wang, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, and Shuqiang Jiang. 2020. LogoDet-3K: A Large-Scale Image Dataset for Logo Detection. *arXiv preprint arXiv:2008.05359* (2020).

[60] Jing Wang, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, Haishuai Wang, and Shuqiang Jiang. 2020. Logo-2K+: A large-scale logo dataset for scalable logo classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 6194–6201.

[61] Jiaqi Wang, Wenwei Zhang, Yuhang Cao, Kai Chen, Jiangmiao Pang, Tao Gong, Jianping Shi, Chen Change Loy, and Dahua Lin. 2020. Side-aware boundary localization for more precise object detection. In *Proceeding of the European Conference on Computer Vision*. Springer, 403–419.

[62] Yue Wu, Yinpeng Chen, Lu Yuan, Zicheng Liu, Lijuan Wang, Hongzhi Li, and Yun Fu. 2020. Rethinking classification and localization for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10186–10195.

[63] Lingxi Xie, Qi Tian, Wengang Zhou, and Bo Zhang. 2014. Fast and accurate near-duplicate image search with affinity propagation on the ImageWeb. *Computer Vision & Image Understanding* 124 (2014), 31–41.

[64] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. 2016. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM International Vonference on Multimedia*. 516–520.

[65] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*. Springer, 818–833.

[66] Hongkai Zhang, Hong Chang, Bingpeng Ma, Naiyan Wang, and Xilin Chen. 2020. Dynamic R-CNN: Towards high quality object detection via dynamic training. In *Proceedings of the European Conference on Computer Vision*. Springer, 260–275.

[67] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. 2020. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9759–9768.

[68] Chenchen Zhu, Yihui He, and Marios Savvides. 2019. Feature selective anchor-free module for single-shot object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 840–849.