mathematics

MDPI

Article

Discriminative Semantic Feature Pyramid Network with Guided Anchoring for Logo Detection

Baisong Zhang [1], Sujuan Hou [1,*], Awudu Karim [2], Jing Wang [1], Weikuan Jia [1] and Yuanjie Zheng [1]

[1] School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China
[2] School of Engineering, Beijing University of Technology, Beijing 101303, China
* Correspondence: sujuanhou@sdnu.edu.cn

Abstract: Logo detection is a technology that identifies logos in images and returns their locations. With logo detection technology, brands can check how often their logos are displayed on social media platforms and elsewhere online and how they appear. It has received a lot of attention for its wide applications across different sectors, such as brand identity protection, product brand management, and logo duration monitoring. Particularly, logo detection technology can offer various benefits for companies to help brands measure their logo coverage, track their brand perception, secure their brand value, increase the effectiveness of their marketing campaigns and build brand awareness more effectively. However, compared with the general object detection, logo detection is more challenging due to the existence of both small logo objects and large aspect ratio logo objects. In this paper, we propose a novel approach, named Discriminative Semantic Feature Pyramid Network with Guided Anchoring (DSFP-GA), which can address these challenges via aggregating the semantic information and generating different aspect ratio anchor boxes. More specifically, our approach mainly consists of two components, namely Discriminative Semantic Feature Pyramid (DSFP) and Guided Anchoring (GA). The former is proposed to fuse semantic features into low-level feature maps to obtain discriminative representation of small logo objects, while the latter is further integrated into DSFP to generate large aspect ratio anchor boxes for detecting large aspect ratio logo objects. Extensive experimental results on four benchmarks demonstrate the effectiveness of the proposed DSFP-GA. Moreover, we further conduct visual analysis and ablation studies to illustrate the strength of the proposed DSFP-GA when detecting both small logo objects and large aspect logo objects.

Keywords: object detection; discriminative semantic features; small logo; large aspect ratio logo; logo detection

MSC: 68T45

check for updates

Citation: Zhang, B.; Hou, S.; Karim, A.; Wang, J.; Jia, W.; Zheng, Y. Discriminative Semantic Feature Pyramid Network with Guided Anchoring for Logo Detection. *Mathematics* **2023**, *11*, 481. https://doi.org/10.3390/math11020481

Academic Editor: Liangxiao Jiang

Received: 10 December 2022
Revised: 5 January 2023
Accepted: 9 January 2023
Published: 16 January 2023

1. Introduction

Researches related to logo detection have been widely carried out in multimedia and beyond [1–7]. Logo detection is an important task for its various applications, such as vehicle logo recognition for intelligent transportation and protection of intellectual property for commercial research [8,9] to mention but a few.

Many current logo detectors directly adopt object detection methods, and thus lack the refinement to the issues of logo detection based on the characteristics of logos. For example, many logo detection methods directly use feature maps extracted by CNNs [10,11]. As a result, the semantic information of low-level feature maps for detecting small logo objects is insufficient. Because the influence of low-level feature maps is low, and the semantic information is thus not fully extracted. Moreover, existing models use the preset anchor mechanism [12–14], and thus cannot effectively deal with the different aspect ratio (the ratio of maximum side to minimize side of object) logo objects, making it difficult to detect large aspect ratio logo objects.

In feature representation, high-level feature maps have detailed information to detect large logo objects, but may miss small logo objects due to their huge influence. However, low-level feature maps contain less semantic information [15,16], which makes it difficult to distinguish between foreground and background, resulting in insufficient training of small logo objects. Although Feature Pyramid Network (FPN) [15] as shown in Figure 1a, is proposed to build a feature pyramid by sequentially combining two adjacent layers via top-down and lateral connections for object detection, the top-down pathway doesn't fully integrate rich semantic information into low-level feature maps.

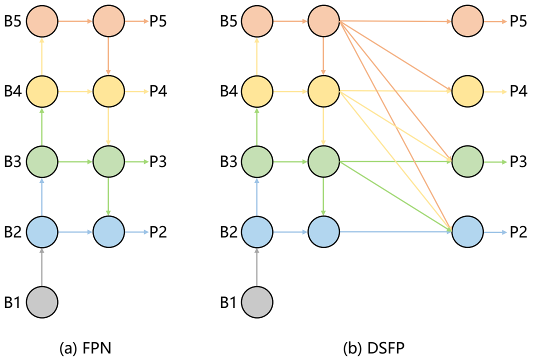


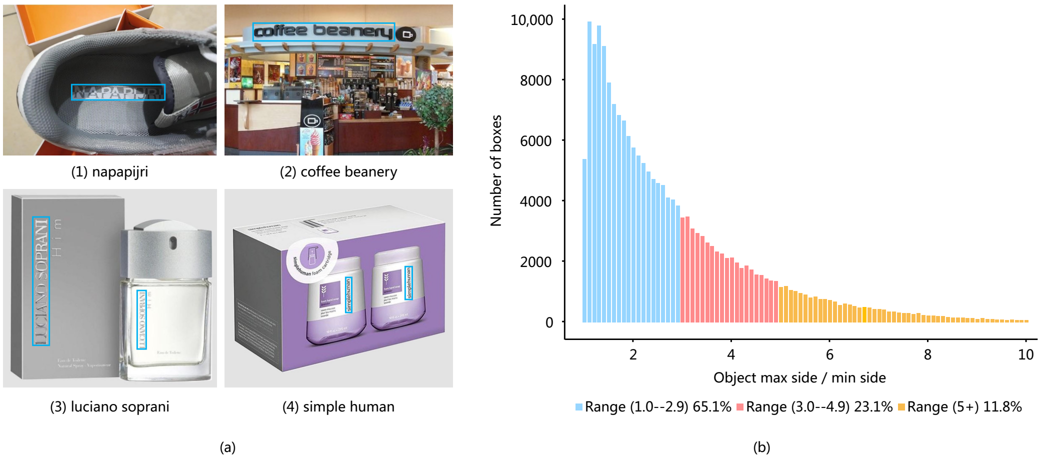
Figure 1. (**a**) FPN introduces a top-down pathway and lateral connections to fuse multi-level features from level 2 to 5 (P2–P5). (**b**) Our DSFP adds high-to-low aggregating pathways and lateral connections, and it mainly can enrich semantic information of low-level feature maps.

In order to obtain efficient feature representation in object detection, existing methods [17–21] usually adopt preset anchor boxes in training. However, large aspect ratio objects exist all the time for logos in reality. As shown in Figure 2a, logos like "napapijri" and "coffee beanery" are extremely wide, while logos like "luciano soprani" and "simple human" are extremely tall, making it challenging for logo detectors to detect these kinds of logo objects. Figure 2b shows LogoDet-3K dataset [22], there are about 35% logo objects with an aspect ratio greater than 3. This challenge leads to the inefficiency of logo detection using the preset anchor boxes, which may produce a large number of regions of negative samples.

The main contributions of this work can be summarized as follows:

- We propose a novel logo detection method DSFP-GA, which can obtain discriminative semantic features and generate large aspect ratio anchor boxes to simultaneously address the issues of detecting small logo objects and large aspect ratio logo objects.
- We design the DSFP to obtain discriminative semantic features for small logos, which can be embedded into any detection models.
- To the best of our knowledge, DSFP-GA is the first work to focus on the issue of large aspect ratio logo objects.
- Extensive evaluations demonstrate the effectiveness of the proposed DSFP-GA over a wide range of state-of-the-art detection models on four logo datasets, namely LogoDet-3K, LogoDet-3K-1000, QMUL-OpenLogo, and FlickrLogos-32.

The remainder of this paper is organized as follows. The related work about object detection and logo detection is described in Section 2. We describe the detailed framework design in Section 3. Experimental results and analysis are reported in Section 4. Finally, we conclude the paper and propose our future research of logo detection in Section 5.



(1) napapijri (2) coffee beanery (3) luciano soprani (4) simple human

(a) (b)

Figure 2. (**a**) Four illustrative large aspect ratio logo images. Categories (1) to (4). Category (1) “napapijri”, max/min equals 6.1; Category (2) “coffee beanery”, max/min equals 7.9; Category (3) “luciano soprani”, (the left box) max/min equals 7.7, (the right box) max/min equals 6.5; Category (4) “simple human”, (the left box) max/min equals 5.8, (the right box) max/min equals 4.7. Blue boxes: ground-truth boxes. Histogram (**b**) of is the number of boxes vs the ratio of maximum dimension to minimum dimension of the object on the LogoDet-3K dataset. The value of max/min accounts for 65.1% in the range of (1–2.9), the value of max/min accounts for 23.1% in the range of (3–4.9), the value of max/min value greater than 5 accounts for 11.8%.

2. Related Work

2.1. Object Detection

Object detection is an important task in computer vision research, and the development of deep learning has vastly improved the performance of object detection. A modern detector is usually composed of two parts: the backbone that is pre-trained on ImageNet [23], and a detection head that is used for predicting localization and classification of objects. For those detectors, their backbones include VGG, ResNet, SpineNet, ResNeXt, and DenseNet, etc. As to the detection head, it is generally divided into two kinds, i.e., one-stage detectors and two-stage detectors.

One-stage detectors include YOLO series [18,19,24], RetinaNet [25], SSD [17] and M2Det [16], etc. They are simpler and faster than two-stage detectors but have lags in performance. The classical detectors generally use preset anchor boxes for object detection. However, manually setting the scale and proportion of the anchor boxes lead to inefficiency in detection tasks of different scenes. Recently, anchor-free methods [26–29], and methods of transformers [30,31] for object detection have been proposed. Anchor-free methods drop the preset anchor boxes. They learn key points according to the characteristics of the objects, such as the center point or 4 corners of the object, and then automatically generates the anchor boxes. Methods based on transformers mainly introduce self-attention, which can better extract features.

Two-stage detectors include R-CNN series [32–35], and ThunderNet [36], etc. Faster R-CNN employs the RPN to generate Regions of Interest (RoIs) by modifying preset anchor boxes and this improves the efficiency of detectors. Moreover, many methods have been introduced to enhance Faster R-CNN from different aspects. Cascade R-CNN [37] extended Faster R-CNN to a multi-stage detector through the cascade architecture. Mask R-CNN [38] replaced the RoIPool layer with the RoIAlign layer using bilinear interpolation. Soft NMS [39] was proposed to improve NMS. We on the other hand, apply the object detection method to logo detection. Unlike these methods, we fully consider the characteristics of logo objects. For the issue of small logo objects, we introduce the DSFP to enhance semantic information of low-level feature maps and improve the performance of small logo objects. For the issue of large aspect ratio objects, we adopt the GA, which can generate large aspect ratio anchor boxes to accurately match these logo objects and effectively improve the efficiency of detection.

2.2. Logo Detection

Logo detection is a special case of object detection, and it can be applied to many fields and has great commercial value. Hence, logo detection has attracted extensive attention from researchers. Early logo detection methods were established on hand-crafted visual features (e.g., SIFT and HOG) and conventional classification models (e.g., SVM). Inspired by the recent advances in object detection using deep learning methods, a remarkable progress has been made in logo detection. Some existing detectors often insert some network layers between the backbone and detection head, and these layers are usually used to collect feature maps from different levels and are helpful to improve the detection performance of small logo objects. Normally, it is composed of several bottom-up paths and several top-down paths. Detectors equipped with this mechanism include Feature Pyramid Network (FPN) [15], Path Aggregation Network (PANet) [40], and Balanced Feature Pyramid (BFP) [41]. FPN used lateral connections and a top-down pathway to enhance the semantic information of shallow layers. After that, PANet brought in a bottom-up pathway to further increase the detailed information in deep layers. MFDNet [5] used BFP to integrate balanced semantic features to strengthen original features.

Unlike these feature pyramid networks, our approach relies on integrated rich semantic features to low-level feature maps, which can enrich the discriminative semantic features of these feature maps to detect small logo objects and then bring improvement for logo detection. In addition, whether logo detectors are improved from one-stage or two-stage methods, almost all of them use preset anchor boxes [42,43] to obtain RoIs. However, consider that there are many logos with large aspect ratios, the proposed method generates anchor boxes according to anchor location branch and anchor shape branch by learning features of logo objects instead of using preset anchor boxes. Compared with the existing logo detectors, our proposed DSFP-GA is more effective for small logo objects and large aspect ratio logo objects.

3. Approach

The overall network architecture of DSFP-GA is shown in Figure 3, which is mainly divided into four parts: namely feature extractor, feature pyramid, guided anchoring, and classification and regression. Specifically, the feature maps of input logo images are extracted by the backbone network. Then the DSFP is used to obtain the semantic information of low-level feature maps for small logo objects. In the region proposal stage, the GA is adopted to yield a set of RoIs. And then each RoI is pooled into a fixed-size feature map through RoI Pooling. In the phase of the classification and regression, feature maps are mapped to a feature vector by a fully connected (FC) layer, and then a feature vector is inputted into the classifiers and bounding box regressors. Finally, the model outputs classification and localization of logo images. The code and models can be found at https://github.com/Zhangbaisong/DSFP-GA.

3.1. Discriminative Semantic Feature Pyramid

In order to address the issue of detecting small logo objects accurately, we propose the DSFP to obtain discriminative semantic features via integrating high-level and middle-level features with rich semantic information to low-level feature maps. As shown in the bottom of Figure 3, the whole process mainly includes three steps: lateral connection, multiple up-sampling, and feature fusion.

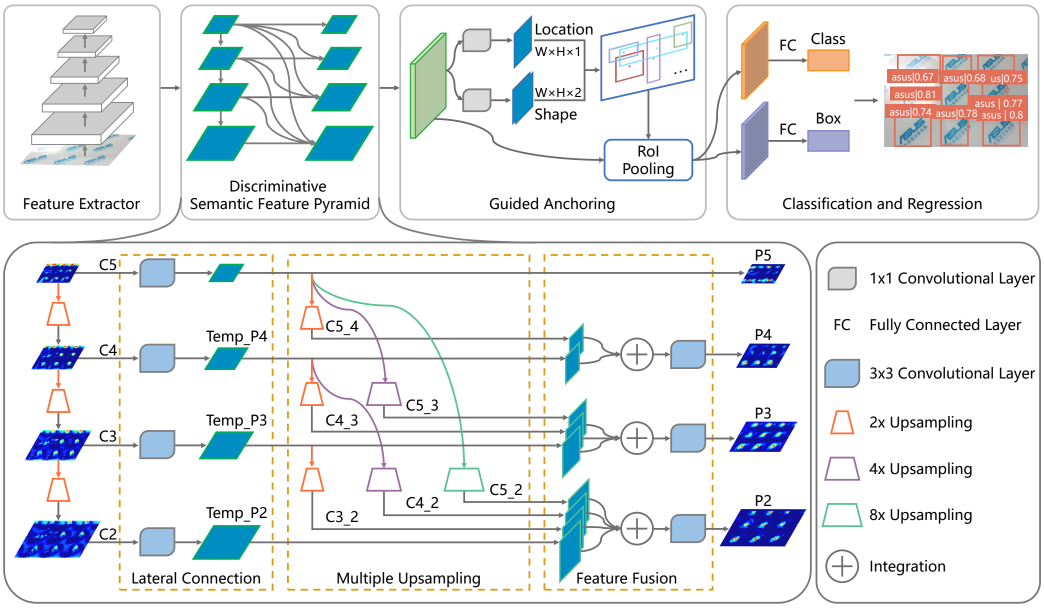


Figure 3. Overview of the proposed Discriminative Semantic Feature Pyramid Network with Guided Anchoring (DSFP-GA). Feature Extractor: we use the ResNet-50 as the backbone to extract feature information. Discriminative Semantic Feature Pyramid: we propose the DSFP to obtain discriminative semantic features. It mainly contains lateral connection, multiple up-sampling, and feature fusion. Here Ci denotes the feature map from stage i of the CNN backbone, and Pi denotes the corresponding feature pyramid level on DSFP. Guided Anchoring: we adopt the GA to generate anchor boxes that can detect a large aspect ratio, and then determine whether it belongs to foreground or background and then apply preliminary bounding box regression. Classification and Regression: output the corresponding category and the final localization.

(1) Lateral Connection. Multi-level feature maps generated by the feature extractor are fed into the DSFP. In Figure 3, $\{C2, C3, C4, C5\}$ are multi-level features from level 2 to 5, and these feature maps are recorded as $\{Temp\_P2, Temp\_P3, Temp\_P4, P5\}$ through lateral connections. Feature maps transform as follows:

$$Ci \xrightarrow{1\times 1\ conv} \begin{cases} Pi & i=5 \\ Temp\_Pi & 2 \le i \le 4 \end{cases} \tag{1}$$

where Ci is the feature map from level i of the CNN backbone. Lateral connections contain a 3×3 convolutional layer on each merged feature map to reduce the aliasing effect of up-sampling and integration.

(2) Multiple Up-sampling. To integrate multi-level features and preserve semantic information, we need to up-sample feature maps $\{P5, Temp\_P4, Temp\_P3\}$ to the corresponding size, and the specific operations are as follows.

$$\begin{aligned} P5 &\xrightarrow{3\ times\ upsample} \begin{cases} C5\_4 \\ C5\_3 \\ C5\_2 \end{cases} \\ Temp\_P4 &\xrightarrow{2\ times\ upsample} \begin{cases} C4\_3 \\ C4\_2 \end{cases} \\ Temp\_P3 &\xrightarrow{1\ times\ upsample} \{\ C3\_2 \end{aligned} \tag{2}$$

where $P5$ and $Temp\_Pi$ are feature maps of level i after lateral connections. Up-sampling $P5$ for three times corresponds to the size of feature maps $\{Temp\_P4, Temp\_P3, Temp\_P2\}$ respectively, and the three obtained feature maps are denoted as $\{C5\_4, C5\_3, C5\_2\}$. Here we use the classical nearest interpolation function for up-sampling. Up-sampling $Temp\_P4$

twice correspond to the size of feature maps $\{Temp\_P3,\ Temp\_P2\}$ respectively, and the two obtained feature maps record as $\{C4\_3,\ C4\_2\}$. Up-sampling $Temp\_P3$ once correspond to the size of the feature map $Temp\_P2$ and one obtained feature map record $C3\_2$. Through this step, we can get the rich semantic information of multi-level feature maps in different resolutions.

(3) Feature Fusion. In this step, we integrate the same size feature maps. The specific operations are as follows.

$$Pi = Temp\_Pi + \sum_{j=i+1}^{j\leq 5} Cj\_i \qquad 2 \leq i \leq 4 \tag{3}$$

where $Temp\_Pi$ denotes the feature map from stage i after lateral connections, and Cj_i denotes the feature map from the up-sampled $Temp\_Pi$, P_i denotes the corresponding feature pyramid level on DSFP. Feature maps $C5\_4$ and $Temp\_P4$ are integrated to get $P4$. Feature maps $C5\_3$, $C4\_3$, and $Temp\_P3$ are integrated to get $P3$. Feature maps $C5\_2$, $C4\_2$, $C3\_2$, and $Temp\_P2$ are integrated to get $P2$. Afterward, we append a 3 × 3 convolutional layer on $\{P2,\ P3,\ P4\}$ to reduce the aliasing effect. $P2$ and $P3$ share the same representation level with the original $C2$ and $C3$ but contains more regional details due to their higher resolution. And the smaller receptive field in $P2$ and $P3$ also helps to better locate small logo objects. Feature maps $\{P2,\ P3,\ P4,\ P5\}$ of final outputs are used for logo detection following the same feature pyramid pipeline.

The proposed DSFP via cross layer fusion from top to bottom, which can ensure that the semantic information of high-level and middle-level feature maps can be directly fused with low-level feature maps. The DSFP achieves the fusion of different levels features through the above three steps, which can obtain discriminative semantic features for detecting small logo objects, and then further improve the performance of the logo detection task.

3.2. Guided Anchoring

GA scheme can predict the aspect ratios of objects at different locations [44], we adopt the GA to adaptively generate the width and height of anchor boxes via learning the features of the logo objects. It mainly consists of two branches; anchor location and anchor shape.

(1) Anchor Location. This branch is used to predict which region could be the center regions of the logo objects. This branch yields a probability map $p(x, y \| F_I)$ of the same size as the input feature map F_I, where x and y are the center coordinates of the anchor boxes. Each entry $p(x, y \| F_I)$ corresponds to the location on the image I as follows,

$$\begin{aligned} x &\to x \cdot s + \frac{s}{2} \\ y &\to y \cdot s + \frac{s}{2} \end{aligned} \tag{4}$$

where s is the stride of the feature map. Through a 1 × 1 convolutional layer, we get the mapping of objectness scores, and we use the sigmoid function to transform it into a probability value. Based on the generated probability map, we determine the possible area of the object by selecting the location with the corresponding probability value higher than the threshold.

(2) Anchor Shape. After determining the possible location of the logo object, our method will predict the shape of the logo objects accurately. The goal of this branch is to predict the width (w_p) and height (h_p) of anchor boxes. Because of great varying range, w_p and h_p are transformed as follows:

$$\begin{aligned} w_p &= s \cdot e^{d_{wp}} \\ h_p &= s \cdot e^{d_{hp}} \end{aligned} \tag{5}$$

This branch is used to predict the shapes of anchor boxes, and it also contains a 1 × 1 convolutional layer and can produce the mapping of two channels, including d_{w_p} and d_{h_p} values, through the formula of conversion to the corresponding w_p and h_p values. And it is difficult to calculate the w and h of objects separately, therefore we directly utilize the Bound IoU Loss [45] as the supervisor of this branch to learn w and h.

Through these methods, we can obtain the location, w and h, so that we can generate the anchor boxes according to w and h at the most appropriate location. In this way, the logo object can be accurately located and the interference from the complex background information can be reduced. The essential difference between the design of guided anchoring and preset anchor boxes is that each position is related to only one anchor box of dynamically predicted shapes instead of a series of preset anchor boxes. Through the two branches of anchor location and anchor shape, our framework can obtain large aspect ratio anchor boxes, and then drastically improve the performance of the logo detection.

3.3. Loss Function

In the training of the DSFP-GA framework, the overall optimization loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{reg} \tag{6}$$

where $\mathcal{L}_{cls}$ and $\mathcal{L}_{reg}$ are losses of classification and localization, respectively. The classification loss is defined as follows:

$$\mathcal{L}_{cls} = \mathcal{L}_{loc} + \mathcal{L}_{ga\_cls} + \mathcal{L}_{head\_cls} \tag{7}$$

where $\mathcal{L}_{ga\_cls}$ and $\mathcal{L}_{head\_cls}$ are classification losses of the GA and the detection head, $\mathcal{L}_{loc}$ is used for anchor location branch. Since the center of the anchor usually accounts for a small portion of the whole feature map, we use the *Focal Loss* [25] to mitigate the imbalance of positive and negative samples.

$$\mathcal{L}_{loc} = \mathcal{L}_{FL} \tag{8}$$

$$\mathcal{L}_{FL} = \begin{cases} -\alpha(1-y')^{\beta}\log y' & , y=1 \\ -(1-\alpha)y'^{\beta}\log(1-y') & , y=0 \end{cases} \tag{9}$$

where $y \in \{\pm 1\}$ is a ground-truth class and $y' \in [0,\ 1]$ is the model's estimated probability by an activation function. Focus loss introduces two factors α and β, where α is used to balance positive and negative samples, while β focuses on the difficult samples. For $\mathcal{L}_{ga\_cls}$ and $\mathcal{L}_{head\_cls}$, we adopt the *Cross Entropy Loss* to calculate the classification loss.

The regression loss is defined as:

$$\mathcal{L}_{reg} = \mathcal{L}_{shape} + \mathcal{L}_{ga\_reg} + \mathcal{L}_{head\_reg} \tag{10}$$

where $\mathcal{L}_{ga\_reg}$, $\mathcal{L}_{shape}$, and $\mathcal{L}_{head\_reg}$ are regression losses of the GA, anchor shape branch, and the detection head, respectively. For $\mathcal{L}_{shape}$ and $\mathcal{L}_{ga\_reg}$, we adopt the *Smooth* L_1 and the *Bounded IoU Loss* respectively.

For $\mathcal{L}_{head\_reg}$, we further incorporate the *CIoU Loss* to obtain more accurate regression results on logo detection. The CIoU loss considers four geometric factors in the process of regression, including the overlap rate, the central point distance, the aspect ratio, and the penalty, and thus can accurately regress the localization of the logo objects and then improve the performance of the logo detection task.

$$\begin{aligned} \mathcal{L}_{head\_reg} &= \mathcal{L}_{CIoU} \\ &= 1 - IoU + R_{CIoU}(B_p, B_g) \\ &= \varphi^2\frac{(b_p, b_g)}{c^2} + \alpha\frac{4}{\pi^2}(arctan\frac{w_g}{h_g} - arctan\frac{w_p}{h_p})^2 \end{aligned} \tag{11}$$

where R_{CIoU} is a penalty term for the predicted box B_p and ground-truth box B_g, b_p and b_g denote the central points of B_p and B_g, $\varphi(\cdot)$ is the Euclidean distance, and c is the diagonal length of the smallest enclosing box covering the two boxes. α is a positive trade-off parameter. w and h are the width and height of the predicted box, respectively.

4. Experiment

In this section, we present a performance evaluation of the proposed method and other trend leading baselines on four logo datasets.

4.1. Experimental Setting

(1) Datasets. We conducted our experiments on four logo datasets with different scales. Most of the experiments were performed on the large-scale LogoDet-3K [22] dataset, which contains 113,710 images for training, 28,432 for validation and 16,510 for testing (Training Set: A set of examples used for learning, which is to fit the parameters [i.e., weights] of the classifier. Validation Set: A set of examples used to tune the parameters [i.e., architecture, not weights] of a classifier, for example to choose the number of hidden units in a neural network. Testing Set: A set of examples used to assess the performance [generalization] of a fully specified classifier [46]. It is noted that validation set which is independent of testing dataset is used for hyperparameter tuning so as to avoid any biasing in choice of hyperparameters. Thus, when the network is completely trained, evaluation is done on completely unseen testing set). To assess the robustness of the DSFP-GA method, experiments were also performed on the LogoDet-3K-1000 [22] dataset, the middle-scale QMUL-OpenLogo [47] dataset, and the small-scale FlickrLogos-32 [48] dataset. The LogoDet-3K-1000 dataset is sampled from the LogoDet-3K dataset, and it consists of 53,049 images for training and 9559 images for testing. The QMUL-OpenLogo dataset contains 27,083 images from 352 logo categories (by aggregating and refining several existing logo datasets). The FlickrLogos-32 dataset consists of 2240 images from 32 logo categories. The detailed statistics of the four datasets are shown in Table 1.

Table 1. Statistics of Four Logo Datasets.

Datasets	Supervision	#Classes	#Images	#Objects	#Trainval	#Test
LogoDet-3K [22]	Object-level	3000	158,652	194,261	142,142	16,510
LogoDet-3K-1000 [22]	Object-level	1000	85,344	101,345	75,785	9559
QMUL-OpenLogo [47]	Object-level	352	27,083	51,207	18,752	8331
FlickrLogos-32 [48]	Object-level	32	2240	3405	1478	762

(2) Implementation Details. The proposed approach is implemented based on the ResNet-50 backbone, which is pre-trained on the ImageNet [23]. For a fair comparison, all baseline detectors are re-implemented based on the publicly available mmdetection toolbox [49] via the same codebase. All models are trained on the training set and validated on the validation set. We adopt the widely used mAP (mean Average Precision) [50] to evaluate the performance of the logo detection. In order to highlight the performance of our method in different sized logos, we also adopt the following evaluation metrics:AP_S is the Average Precision (AP) for small logo objects (area < 32^2), AP_M is the AP for medium logo objects (32^2 < area < 96^2), AP_L is the AP for large logo objects (area > 96^2). The threshold of Intersection over Union (IoU) between the predicted bounding box and ground-truth bounding box is 0.5. We train these detectors with an initial learning rate of 0.002 and the input images are resized to 1000 × 600. All other hyper-parameters follow the settings in the mmdetection toolbox.

4.2. Ablation Study

In this part, we provide empirical analysis for each component in DSFP-GA, DSFP, GA, and CIoU loss. We report the overall ablation studies on the LogoDet-3K dataset shown in Table 2, in which the first row lists the experimental results conducted on Faster R-CNN with ResNet-50-FPN. Furthermore, the results of the ablation studies on the LogoDet-3K-1000, the QMUL-OpenLogo, and the FlickrLogos-32 datasets as are shown in Tables 3–5. These results as illustrated in the tables above, clearly indicate the effectiveness of our method in many aspects on different logo datasets.

Table 2. Evaluating Individual Component on the LogoDet-3K Dataset. Discriminative Semantic Feature Pyramid (DSFP), Guided Anchoring (GA), Complete IoU Loss (CIoU Loss).

DSFP	GA	CIoU Loss	$mAP(\%)$	$AP_S(\%)$	$AP_M(\%)$	$AP_L(\%)$
			83.8	44.7	76.8	87.7
✓			84.5	51.8	78.6	88.0
✓	✓		86.6	54.7	81.8	89.5
✓	✓	✓	87.7	56.2	83.1	90.5

Table 3. Evaluating Individual Component on the LogoDet-3K-1000 Dataset. Discriminative Semantic Feature Pyramid (DSFP), Guided Anchoring (GA), Complete IoU Loss (CIoU Loss).

DSFP	GA	CIoU Loss	$mAP(\%)$	$AP_S(\%)$	$AP_M(\%)$	$AP_L(\%)$
			88.2	40.7	81.3	92.6
✓			88.8	52.0	81.9	92.4
✓	✓		89.4	50.6	81.9	93.7
✓	✓	✓	90.1	56.0	83.9	93.8

Table 4. Evaluating Individual Component on the QMUL-OpenLogo Dataset. Discriminative Semantic Feature Pyramid (DSFP), Guided Anchoring (GA), Complete IoU Loss (CIoU Loss).

DSFP	GA	CIoU Loss	$mAP(\%)$	$AP_S(\%)$	$AP_M(\%)$	$AP_L(\%)$
			51.9	31.3	52.9	65.1
✓			53.5	32.6	55.6	66.9
✓	✓		53.7	32.7	55.9	66.3
✓	✓	✓	54.0	33.2	56.4	66.5

Table 5. Evaluating Individual Component on the FlickrLogos-32 Dataset. Discriminative Semantic Feature Pyramid (DSFP), Guided Anchoring (GA), Complete IoU Loss (CIoU Loss).

DSFP	GA	CIoU Loss	$mAP(\%)$	$AP_S(\%)$	$AP_M(\%)$	$AP_L(\%)$
			85.9	22.8	81.3	91.5
✓			86.6	28.4	83.6	92.0
✓	✓		86.7	28.7	83.7	92.1
✓	✓	✓	87.1	28.5	83.3	92.6

(1) DSFP. We evaluate the effect of the DSFP by comparing it with FPN. The proposed DSFP is mainly used to improve the ability to detect small logo objects, and also enhance the semantic information of feature maps. As shown in Table 6, small logo objects and medium logo objects account for 1.8% and 29.8% on the LogoDet-3K dataset. The DSFP brings 0.7% mAP improvement over the Faster R-CNN on the LogoDet-3K dataset in Table 2. Especially, the AP_S scores increase by 7.1%, validating the effectiveness on small logo objects detection of the DSFP.

Table 6. The Proportion of Object Size and Aspect Ratio on Four Datasets.

Datasets	Object Size			Aspect Ratio		
	Small	**Medium**	**Large**	**Range (1.0–2.9)**	**Range (3.0–4.9)**	**Range (5.0+)**
LogoDet-3K	1.8%	29.8%	68.4%	65.1%	23.1%	11.8%
LogoDet-3K-1000	1.7%	33.0%	65.3%	63.9%	24.1%	12%
QMUL-OpenLogo	23.1%	44%	32.9%	81.5%	14.2%	4.3%
FlickrLogos-32	5.4%	29.3%	65.3%	94.8%	4.3%	0.9%

Clearly, the DSFP can perform better than the FPN for logo detection tasks. This is because that it can extract more discriminative semantic features. To verify this, we visualize the heatmap in Figure 4, which demonstrates that the DSFP is more effective in extracting discriminative semantic features for logo detection. The results prove that the red areas of $P2$ and $P3$ in DSFP are more accurate and have richer semantic information than those in the FPN. It is noteworthy that feature maps that come from the DSFP are more representative since they have stronger activation values in the foreground and weaker activation values in the background.

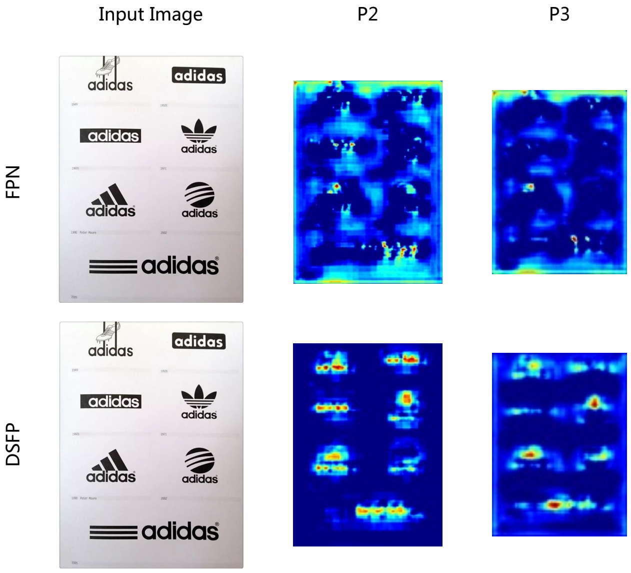
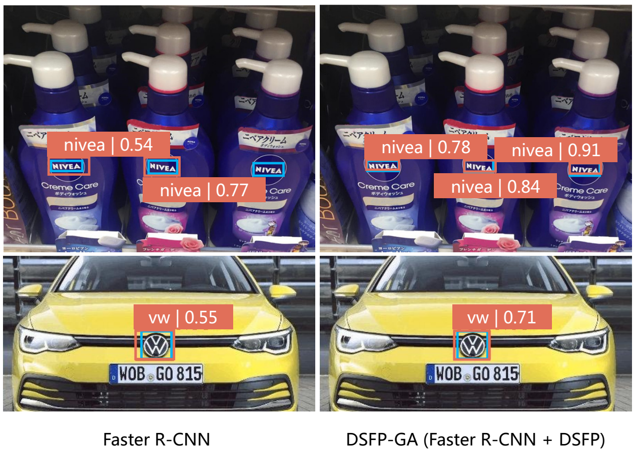


Figure 4. Visualization comparison of the features extracted by FPN and DSFP. P2 and P3: the second level and the third level of the feature pyramid.

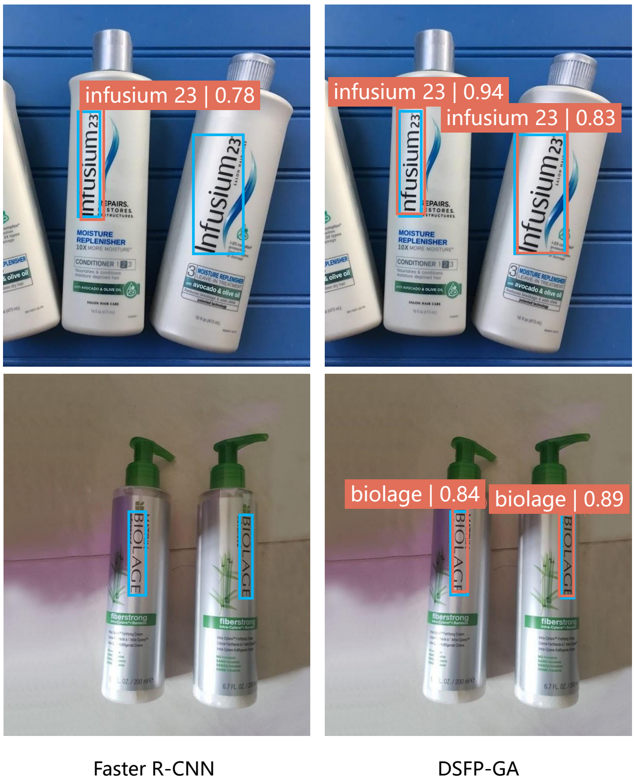
Besides visualizing heatmaps, we also visualize the detection results of two images with small logo objects in Figure 5. Compared with DSFP-GA, Faster R-CNN misses a small logo object in the first image. It further proves the strengths of DSFP-GA in small logo object detection. The second image has small and extremely tall objects in Figure 6. Faster R-CNN lacks a good solution to deal with this kind of logo objects, and the detection result is less satisfactory. In contrast, our method has the strength of detecting the small and extremely tall logo objects.

Moreover, we validate the benefit of the DSFP on the other three datasets. Similar to the LogoDet-3K dataset, small logo objects and medium logo objects account for 1.7% and 33% on the LogoDet-3K-1000 datasets in Table 6. As shown in Table 3, the DSFP increases 0.6% mAP over Faster R-CNN, especially the AP_S scores increase by 11.3%, which shows that our DSFP enriches discriminative semantic information of feature maps. For the QMUL-OpenLogo dataset, more than 23.1% are small logo objects and over 44% are medium logo objects as shown in Table 6. It can be seen that the main challenge is the small logo objects on the QMUL-OpenLogo dataset. The DSFP has an obvious improvement over the baseline in Table 4. It increases by 1.6% mAP, and yields 7.1% AP_S scores improvement, that shows the effectiveness of the DSFP in small logo detection. For the FlickrLogos-32 dataset, we can observe that less than 5.4% are small logo objects and about 29.3% are medium logo objects in Table 6. DSFP still brings 0.7% mAP improvement and 5.6% AP_S scores improvement over Faster R-CNN baseline in Table 5, which can indicate that the DSFP has enhanced discriminative semantic information of feature maps.



Faster R-CNN DSFP-GA (Faster R-CNN + DSFP)

Figure 5. Comparison of small logo detection results between Faster R-CNN and DSFP-GA. Blue boxes: ground-truth boxes. Orange boxes: correct detection boxes.



Faster R-CNN DSFP-GA

Figure 6. Comparison of large aspect ratio (extremely tall) logo detection results between Faster R-CNN and DSFP-GA. Blue boxes: ground-truth boxes. Orange boxes: correct detection boxes.

(2) GA. We evaluate the strength of the GA on the LogoDet-3K dataset. The GA does not implicitly limit the aspect ratio and the size of anchor boxes, thereby addressing the issue of large aspect ratio logo objects well. For the LogoDet3K dataset, more than 35% of logo objects have an aspect ratio greater than 3, and more than 11.8% of logo objects have an aspect ratio greater than 5 as shown in Table 6. There are many large aspect ratio logo

objects in the LogoDet3K dataset. As shown in Table 2, the GA improves the mAP from 84.5% to 86.6% on the LogoDet-3K dataset, which indicates the strength of our method.

We visualize the results in Figures 6 and 7 to show that our method is effective in dealing with large aspect ratio logo objects, and the detection results of DSFP-GA are better than that of Faster R-CNN. As shown in Figure 6, for the first image, Faster R-CNN doesn't detect the logo with a tilted angle on the right side of the image at all. On the contrary, DSFP-GA detects the logo object on the right side with high accuracy, which goes to prove that DSFP-GA is robust in detecting difficult logo objects. As for the logo object on the left side, the accuracy of Faster R-CNN is 16% lower than DSFP-GA. In the second image, we can see that the ground-truth boxes are small and extremely tall logo objects. DSFP-GA detects these two logo objects with good accuracy. As shown in Figure 7, for the first two images, DSFP-GA has more accurate detection results than Faster R-CNN. In the third image, Faster R-CNN mistakenly identifies the logo category, and the detection accuracy of the correct box is 26% lower than DSFP-GA. This compellingly proves the superiority of DSFP-GA in detecting large aspect ratio logo objects as well as small logo objects effectively.

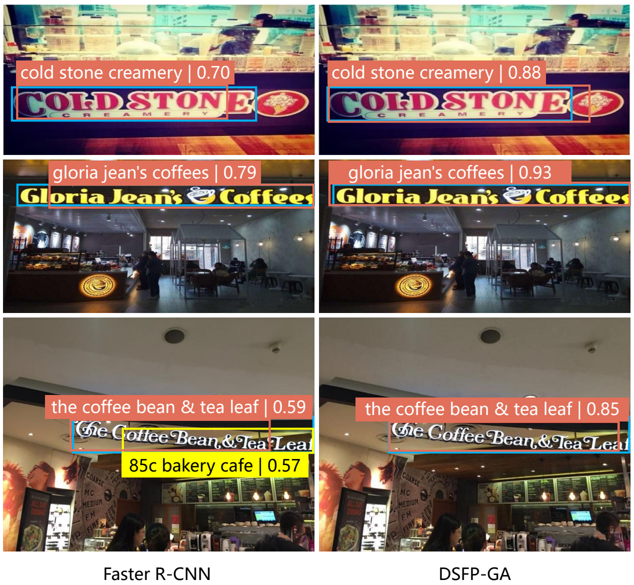

cold stone creamery | 0.70
cold stone creamery | 0.88
gloria jean's coffees | 0.79
gloria jean's coffees | 0.93
the coffee bean & tea leaf | 0.59
85c bakery cafe | 0.57
the coffee bean & tea leaf | 0.85
Faster R-CNN
DSFP-GA

Figure 7. Comparison of large aspect ratio (extremely wide) logo detection results between Faster R-CNN and DSFP-GA. Blue boxes: ground-truth boxes. Orange boxes: correct detection boxes. Yellow boxes: mistaken detection boxes.

The ablation studies on the LogoDet-3K-1000 dataset can further validate the benefit of the GA. In Table 6, logo objects of Range (3+) account for 36.1% and logo objects of Range (5+) account for 11.9% on the LogoDet-3K-1000 dataset. The GA yields 0.6% mAP improvement on the LogoDet-3K-1000 dataset in Table 3, which indicates the effectiveness of GA when addressing the issue of large aspect ratio logo objects.

The ablation studies on the QMUL-OpenLogo dataset and the FlickrLogos-32 dataset shows the performance of the GA. As shown in Table 6, we find that more than 81.5% of the logo objects have an aspect ratio between 1 and 2.9, and about 4.3% have an aspect ratio greater than 5 on the QMUL-OpenLogo dataset. The GA improves the mAP from 53.5% to 53.7% as shown in Table 4. Furthermore, as shown in Table 6, there are approximately 95% of the logo objects that have an aspect ratio between 1 and 2.9, and only about 0.9% have an aspect ratio greater than 5 on the FlickrLogos-32 dataset. Similarly, the GA increases by 0.1% mAP as shown in Table 5. Hence, we can draw a safe conclusion that the GA has a better performance for detecting large aspect ratio logo objects than Faster R-CNN.

(3) CIoU Loss. We also evaluate the benefit of the CIoU loss on these four logo datasets. The CIoU loss can obtain more accurate regression results via solving the problem of inconsistency to improve detection performance. In Table 2, the CIoU loss improves the

mAP from 86.6% to 87.7% on the LogoDet-3K dataset. The CIoU loss also improves the mAP from 89.4% to 90.1% on the LogoDet-3K-1000 dataset in Table 3. As shown in Table 6, the CIoU loss increases the mAP from 53.7% to 54.0% on the QMUL-OpenLogo dataset. The CIoU loss improves the mAP from 86.7% to 87.1% on the FlickrLogos-32 dataset in Table 5. These validate the effectiveness of our method when adopting the CIoU loss. However, the AP_S and AP_M scores decreased slightly on the FlickrLogos-32 dataset. Our observation is that FlickrLogos-32 contains fewer logo images, therefore, CIoU loss does not play significant roles in the AP_S and AP_M scores on this dataset.

In order to evaluate the better performance of DSFP-GA, we selected two images that contain both small size and large aspect ratio logo objects and visualized the detection results. As shown in Figure 8, Faster R-CNN doesn't detect the logo object that is small and wide in the first image. On the same image, DSFP-GA has better results in localization and classification. In the second image, Faster R-CNN mistakenly detects two logo objects and the accuracy of another logo object detected by Faster R-CNN is much lower than DSFP-GA. This amply demonstrates the superior performance of DSFP-GA in both small size and large aspect ratio logo object detections.

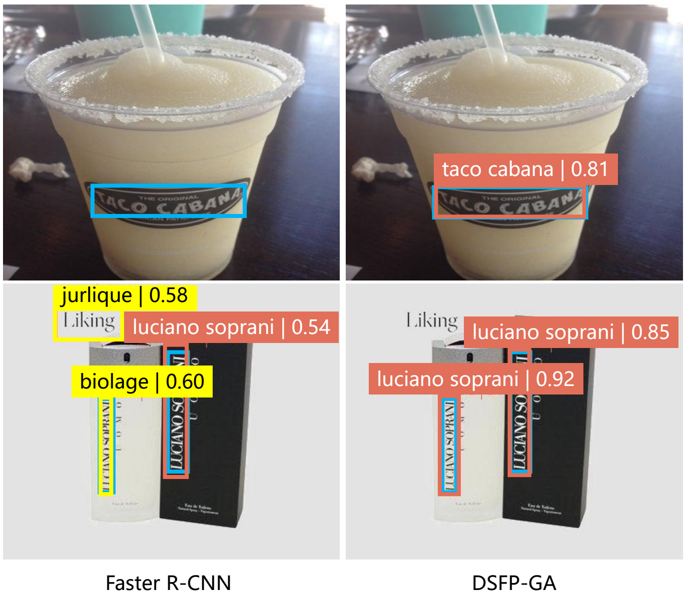


Figure 8. Comparison of both small size and large aspect ratio logo detection results between Faster R-CNN and DSFP-GA. Blue boxes: ground-truth boxes. Orange boxes: correct detection boxes. Yellow boxes: mistaken detection boxes.

4.3. Comparison of State-of-the-Art Frameworks

To further validate the versatility of the proposed DSFP-GA, experiments were performed with multiple top of the trend approaches. We chose several one-stage frameworks that have good performance on general detection datasets in recent years. We also selected a series of standard two-stage frameworks which are improved based on Faster R-CNN and are state-of-the-art.

(1) Experiment on the LogoDet-3K. Our method DSFP-GA achieves the best performance on the LogoDet-3K datasets. We compared DSFP-GA with the state-of-the-art detection approaches on the large-scale LogoDet-3K dataset in Table 7. Compared with the existing two-stage baselines Faster R-CNN, Libra R-CNN, and Dynamic R-CNN, etc., the DSFP-GA significantly outperforms these state-of-the-art frameworks. Our approach is based on modified on Faster R-CNN and achieves the best mAP of 87.7%, surpassing the Faster RCNN baseline of 3.9% mAP, which indicates the effectiveness of our strategy. Compared with Dynamic R-CNN, which ranks second of mAP the performance of our method is 0.3% mAP better than it. The AP_S, AP_M and AP_L scores of Dynamic R-CNN are 53.7%, 82.0%, and 90.4% respectively, and our method scores are 2.5%, 1.1%, and 0.1%

higher respectively. This clearly demonstrates the effectiveness of our method. In addition, our framework also improves by 4.6% mAP compared with PANet that is equipped with the excellent feature pyramid structure PAFPN. It can be observed that our DSFP has a better effect on fusing the features of logo objects than PAFPN. We also compare DSFP-GA with state-of-the-art one-stage approaches. Our framework brings 7.8% mAP improvement over ATSS [51] and 6.5% mAP improvement over GFL [52]. This superior performance is because, there are many large aspect ratio logo objects in this dataset, and our model is sufficiently equipped for this challenging issue.

Table 7. Detection Results on the LogoDet-3K Dataset.

Methods	Backbone	mAP (%)
One-stage:		
ATSS [51]	ResNet-50-FPN	79.9
FSAF [27]	ResNet-50-FPN	78.3
GFL [52]	ResNet-50-FPN	81.2
Two-stage:		
Faster R-CNN [34]	ResNet-50-FPN	83.8
Soft-NMS [39]	ResNet-50-FPN	82.1
PANet [40]	ResNet-50-PAFPN	83.1
Generalized IoU [53]	ResNet-50-FPN	84.4
Distance IoU [54]	ResNet-50-FPN	83.5
Complete IoU [54]	ResNet-50-FPN	82.7
Libra R-CNN [41]	ResNet-50-BFP	82.4
Cascade R-CNN [37]	ResNet-50-FPN	85.6
Dynamic R-CNN [21]	ResNet-50-FPN	87.4
Sparse R-CNN [55]	ResNet-50-FPN	74.3
DSFP-GA	ResNet-50-DSFP	87.7

Detection results of DSFP-GA given in Figure 9, clearly demonstrate that our model has superior performance in detecting all kinds of logos of various sizes and shapes. We can see from Figure 9 that our model has better detection results on large logo objects (category "cherry 7up", category "waffle house", etc.), medium logo objects (category "cheez whiz", category "swiss miss", etc.), and small logo objects (category "freia", category "nioxin", etc.). It is worth mentioning that there are multiple multi-scale objects in a test image from Figure 9, where our model also can detect all logo objects accurately. These prove our method can well detect logo objects of different sizes and has the capacity to handle multiple logo objects within one image.

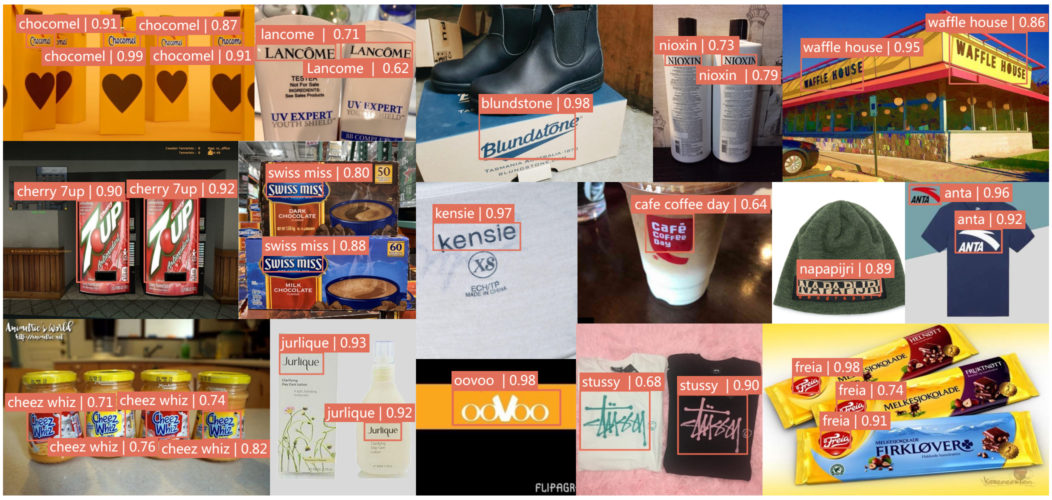


Figure 9. Some examples of detection results of DSFP-GA. The orange box corresponds to the location of the logo objects. On the top of the box is the category name and its accuracy.

(2) Experiment on the LogoDet-3K-1000. Experiment on the LogoDet-3K-1000. DSFP-GA again has the best performance on the LogoDet-3K-1000 dataset. As shown in Table 8, DSFP-GA achieves 90.1% mAP, which is an increase of 1.9% mAP over Faster R-CNN. In addition, our method also improves by 1.0% mAP compared with PANet. Dynamic R-CNN achieves 89.5% mAP, which ranks second to our method, in that, our method yields 0.6% mAP over Dynamic R-CNN. The AP_S, AP_M and AP_L scores of Dynamic R-CNN are 49.7%, 82.2%, and 93.6% respectively, and our method's scores are 6.3%, 1.7%, and 0.2% higher respectively. Compared with the one-stage frameworks, our work yields 2.3% mAP over ATSS and 2.4% mAP over GFL. The LogoDet-3K-1000 dataset contains a huge number of large aspect ratio logo objects, which account for the exceptional performance highlighting the effectiveness of our model in dealing with these kinds of logo objects. The experiments on the LogoDet-3K-1000 dataset further vindicate the superiority of the proposed DSFP-GA method over the evaluated state-of-the-art methods as illustrated in Table 8 below.

Table 8. Detection Results on the LogoDet-3K-1000 Dataset.

Methods	Backbone	mAP (%)
One-stage:		
ATSS [51]	ResNet-50-FPN	87.8
FSAF [27]	ResNet-50-FPN	87.3
GFL [52]	ResNet-50-FPN	87.7
Two-stage:		
Faster R-CNN [34]	ResNet-50-FPN	88.2
Soft-NMS [39]	ResNet-50-FPN	89.1
PANet [40]	ResNet-50-PAFPN	89.1
Generalized IoU [53]	ResNet-50-FPN	88.2
Distance IoU [54]	ResNet-50-FPN	88.7
Complete IoU [54]	ResNet-50-FPN	88.9
Libra R-CNN [41]	ResNet-50-BFP	88.4
Cascade R-CNN [37]	ResNet-50-FPN	89.1
Dynamic R-CNN [21]	ResNet-50-FPN	89.5
Sparse R-CNN [55]	ResNet-50-FPN	86.8
DSFP-GA	ResNet-50-DSFP	90.1

(3) Experiment on the QMUL-OpenLogo. From Table 9, we can see that our method achieves the best performance (by 54.0% mAP) on the middle-scale logo dataset. We also list the experimental results of baselines on the middle scale QMUL-OpenLogo dataset. Compared with Faster R-CNN, DSFP-GA obtains 2.1% mAP improvement. Our method also improves by 1.1% mAP compared with PANet. This further shows that DSFP-GA can handle the QMUL-OpenLogo dataset which contains small logo objects better than Faster RCNN and PANet. Compared with Cascade R-CNN, which ranks second (by 53.1% mAP), the performance of our method is 0.9% mAP improvement over it. The AP_S, AP_M and AP_L scores of Cascade R-CNN are 32.7%, 54.0%, and 67.4% respectively, and our method's scores are 0.5%, 2.4%, and 0.9% higher respectively, indicating the effectiveness of our method. Compared with the best performing one-stage method GFL, our method improves by 4.8% mAP than it (i.e. 54.0% over 49.2%). These results indicate that our model is efficient in dealing with the challenges of small logo objects.

(4) Experiment on the FlickrLogos-32. Our framework also has good performance on the small-scale FlickrLogos-32 dataset. The experimental results of baseline and our framework on the small-scale FlickrLogos-32 dataset are summarized in Table 10. Our method achieves 87.1% mAP that is the same as Cascade R-CNN in Table 10. Cascade R-CNN achieves 87.0% mAP by cascading multiple detection heads. However, our framework only uses one detection head, which achieves great performance 87.1% mAP. The AP_S, AP_M and AP_L scores of Cascade R-CNN are 17.3%, 80.5%, and 93.4% respectively, Especially the AP_S and AP_M scores of our method are 11.2% and 2.8% better than it, This can show the

effectiveness of our method. For the one-stage framework, our method increases 1.0% mAP than ATSS and 0.9% mAP than GFL. Two-stage detectors have one more region proposal network, which can improve the detection result. Although the FlickrLogos-32 dataset is small and simple, DSFP-GA still has better detection result than one-stage frameworks.

Table 9. Detection Results on the QMUL-OpenLogo Dataset.

Methods	Backbone	mAP (%)
One-stage:		
FoveaBox [56]	ResNet-50-FPN	35.6
SSD [17]	VGG-16	41.6
ATSS [51]	ResNet-50-FPN	48.6
FSAF [27]	ResNet-50-FPN	44.7
GFL [52]	ResNet-50-FPN	49.2
Two-stage:		
Faster R-CNN [34]	ResNet-50-FPN	51.9
Soft-NMS [39]	ResNet-50-FPN	52.3
PANet [40]	ResNet-50-PAFPN	52.9
Libra R-CNN [41]	ResNet-50-BFP	52.7
Dynamic R-CNN [21]	ResNet-50-FPN	51.8
Cascade R-CNN [37]	ResNet-50-FPN	53.1
Sparse R-CNN [55]	ResNet-50-FPN	46.9
DSFP-GA	ResNet-50-DSFP	54.0

Table 10. Detection Results on the FlickrLogos-32 Dataset.

Methods	Backbone	mAP (%)
One-stage:		
FoveaBox [56]	ResNet-50-FPN	84.1
SSD [17]	VGG-16	80.2
RetinaNet [25]	ResNet-50-FPN	78.4
ATSS [51]	ResNet-50-FPN	86.1
FSAF [27]	ResNet-50-FPN	82.5
GFL [52]	ResNet-50-FPN	86.2
Two-stage:		
BD-FRCN-M [57]	VGG-16	73.5
Deep Logo [58]	VGG-16	74.4
Faster R-CNN [34]	ResNet-50-FPN	85.9
Soft-NMS [39]	ResNet-50-FPN	86.5
PANet [40]	ResNet-50-PAFPN	86.2
Libra R-CNN [41]	ResNet-50-BFP	84.6
Dynamic R-CNN [21]	ResNet-50-FPN	85.8
Cascade R-CNN [37]	ResNet-50-FPN	87.0
Sparse R-CNN [55]	ResNet-50-FPN	73.7
DSFP-GA	ResNet-50-DSFP	87.1

4.4. Result Analysis

To further evaluate the performance of DSFP-GA in detecting small logo objects and large aspect ratio logo objects, we selected two categories that small logo objects account for a large proportion and the other two categories that large aspect ratio logo objects occupy a substantial part in the LogoDet-3K dataset as shown in Figure 10. The Average Precision (AP, evaluation indicators for a single category) values of four categories are shown in Figure 11. We analyze the characteristics and AP values of these four categories below.

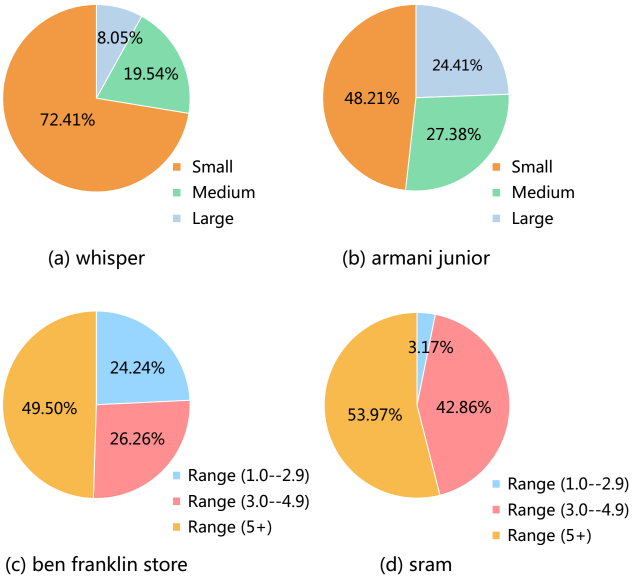


Figure 10. The proportion of small, medium, and large logo objects in categories "whisper" and "armani junior", the proportion of different aspect ratios in categories "ben franklin store" and "sram".

The small logo objects in the "whisper" category have 72.41% of the proportion in Figure 10a. As shown in Figure 11, the AP value is 15.2% in this category on Faster R-CNN. DSFP-GA improves by 10% AP over Faster RCNN. Similarly, in Figure 10b, small logo objects account for nearly half of the proportion in the "armani junior" category. DSFP-GA again improves by 17% AP over Faster R-CNN. It is observed that our method has a vast performance improvement in small logo detection. As for the categories of "ben franklin store" and "sram", we can see that large aspect ratio logo objects account for a large proportion (49.50% and 53.97%) in Figure 10c,d. As shown in Figure 11, DSFP-GA increases by 12.2% and 26.4% AP over Faster R-CNN respectively. Faster R-CNN cannot deal well with large aspect ratio logo objects through the preset anchor boxes. However, DSFP-GA performs better in these two categories, indicating that DSFP-GA has a superiority in detecting large aspect ratio logo objects over the state-of-the-art Faster R-CNN.

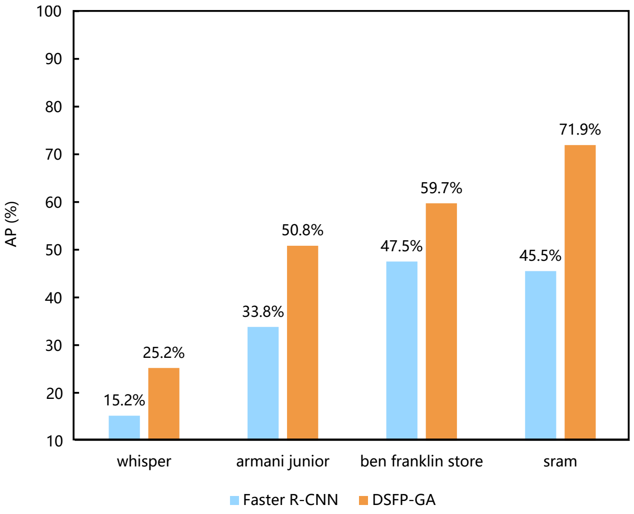


Figure 11. The AP value of four categories in Faster R-CNN and DSFP-GA on LogoDet-3K dataset.

5. Conclusions

In this work, we propose a novel logo detection method, namely DSFP-GA, for detecting both small logo objects and large aspect ratio logo objects which is a rarely explored problem in logo detection. This work is proposed to overcome the logo detection challenges. We have designed the DSFP component of the method to enhance discriminative semantic features, which can improve the performance of small logo detections. The GA component on the other hand, can generate the adaptive widths and heights of anchor boxes accordingly and effectively deal with large aspect ratio logo objects. To the best of our knowledge, our framework is the first work to focus on the issue of large aspect ratio logo objects detections. We further adopt the CIoU loss for regression to enhance the performance of the framework. Extensive evaluations were conducted on four standard logo benchmarks to validate the strengths of the proposed DSFP-GA method over selected state-of-the-art methods. In the future, we aim to design an enhanced feature pyramid and region proposal network to further improve the performance of logo detection systems.

Author Contributions: Conceptualization, S.H. and B.Z.; methodology, S.H., B.Z. and J.W.; software, B.Z.; validation, B.Z.; data curation, B.Z.; writing—original draft preparation, S.H. and B.Z.; writing—review and editing, S.H., A.K., J.W. and W.J.; visualization, B.Z.; supervision, S.H., W.J. and Y.Z.; project administration, S.H.; funding acquisition, S.H. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (No.62072289, No.61702313).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Eggert, C.; Zecha, D.; Brehm, S.; Lienhart, R. Improving small object proposals for company logo detection. In Proceedings of the ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 167–174.
2. Marshita, M. PDF Redesign and Analysis of Logo in Campus Publishing Business as Corporate Identity. *J. Appl. Multimed. Netw.* **2021**, *5*, 1–12. [CrossRef]
3. Jain, R.K.; Watasue, T.; Nakagawa, T.; Takahiro, S.; Iwamoto, Y.; Xiang, R.; Yen-Wei, C. LogoNet: Layer-Aggregated Attention CenterNet for Logo Detection. In Proceedings of the IEEE International Conference on Consumer Electronics (ICCE), Taiwan, China, 10–12 January 2021; pp. 1–6.
4. Wang, J.; Min, W.; Hou, S.; Ma, S.; Zheng, Y.; Wang, H.; Jiang, S. Logo-2K+: A large-scale logo dataset for scalable logo classification. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Hilton New York Midtown, New York, NY , USA, 7–12 February 2020; Volume 34, pp. 6194–6201.
5. Hou, Q.; Min, W.; Wang, J.; Hou, S.; Zheng, Y.; Jiang, S. 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, Chengdu, China, 20–24 October 2021; pp. 4670–4679.
6. Gao, Y.; Wang, F.; Luan, H.; Chua, T.S. Brand data gathering from live social media streams. In Proceedings of the International Conference on Multimedia Retrieval (ICMR), Glasgow, UK, 1–4 April 2014; pp. 169–176.
7. Yang, S.; Bo, C.; Zhang, J.; Gao, P.; Li, Y.; Serikawa, S. VLD-45: A Big Dataset for Vehicle Logo Recognition and Detection. *IEEE Trans. Intell. Transp. Syst.* **2021**, *99*, 1–7. [CrossRef]
8. Chen, R.; Jalal, M.A.; Mihaylova, L.; Moore, R.K. Learning capsules for vehicle logo recognition. In Proceedings of the IIEEE International Conference on Information Fusion (FUSION), Salamanca, Spain, 7–10 July 2018; pp. 565–572.
9. Li, Y.; Shi, Q.; Deng, J.; Su, F. Graphic logo detection with deep region-based convolutional networks. In Proceedings of the IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; pp. 1–4.
10. Su, H.; Gong, S.; Zhu, X. Scalable logo detection by self co-learning. *Pattern Recognit.* **2020**, *97*, 107003. [CrossRef]
11. Tüzkö, A.; Herrmann, C.; Manger, D.; Beyerer, J. Open set logo detection and retrieval. *arXiv* **2017**, arXiv:1710.10891.
12. Sahel, S.; Alsahafi, M.; Alghamdi, M.; Alsubait, T. Logo Detection Using Deep Learning with Pretrained CNN Models. *Eng. Technol. Appl. Sci. Res. (ETASR)* **2021**,*11*, 6724–6729. [CrossRef]
13. Su, H.; Gong, S.; Zhu, X. WebLogo-2M: Scalable logo detection by deep learning from the web. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 270–279.
14. Sharma, N.; Mandal, R.; Sharma, R.; Pal, U.; Blumenstein, M. Signature and logo detection using deep CNN for document image retrieval. In Proceedings of the International Conference on Frontiers in Handwriting Recognition (ICFHR), Niagara Falls, NY, USA, 5–8 August 2018; pp. 416–422.

15. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
16. Zhao, Q.; Sheng, T.; Wang, Y.; Tang, Z.; Chen, Y.; Cai, L.; Ling, H. M2Det: A single-shot object detector based on multi-level feature pyramid network. In Proceedings of the AAAI Conference on Artificial Intelligence, Hilton Hawaiian Village, Honolulu, HI, USA, 27 January–1 February, 2019; Volume 33, pp. 9259–9266.
17. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 10–16 October 2016; pp. 21–37.
18. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
19. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
20. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
21. Zhang, H.; Chang, H.; Ma, B.; Wang, N.; Chen, X. Dynamic R-CNN: Towards high quality object detection via dynamic training. In Proceedings of the European Conference on Computer Vision (ECCV), Virtual Platform, 23–28 August 2020; pp. 260–275.
22. Wang, J.; Min, W.; Hou, S.; Ma, S.; Zheng, Y.; Jiang, S. LogoDet-3K: A Large-Scale Image Dataset for Logo Detection. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2022**, *18*, 1–19. [CrossRef]
23. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255.
24. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
25. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
26. Law, H.; Deng, J. CornerNet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
27. Zhu, C.; He, Y.; Savvides, M. Feature selective anchor-free module for single-shot object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 840–849.
28. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
29. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully convolutional one-stage object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 9627–9636.
30. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Virtual Platform, 23–28 August 2020; pp. 213–229.
31. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
32. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
33. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 11–18 December 2015; pp. 1440–1448.
34. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]
35. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object detection via region-based fully convolutional networks. *arXiv* **2016**, arXiv:1605.06409
36. Qin, Z.; Li, Z.; Zhang, Z.; Bao, Y.; Yu, G.; Peng, Y.; Sun, J. ThunderNet: Towards real-time generic object detection on mobile devices. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6718–6727.
37. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
38. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
39. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS–improving object detection with one line of code. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5561–5569.
40. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
41. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra R-CNN: Towards balanced learning for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 821–830.
42. Leng, F. A Gradient Balancing Approach for Robust Logo Detection. In Proceedings of the ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 4765–4769.
43. Xu, W.; Liu, Y.; Lin, D. A Simple and Effective Baseline for Robust Logo Detection. In Proceedings of the ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 4784–4788.

44. Wang, J.; Chen, K.; Yang, S.; Loy, C.C.; Lin, D. Region proposal by guided anchoring. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 2965–2974.
45. Tychsen-Smith, L.; Petersson, L. Improving object localization with fitness nms and bounded iou loss. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6877–6885.
46. Ripley, B.D. *Pattern Recognition and Neural Networks*; Cambridge University Press: Cambridge, UK, 2007.
47. Su, H.; Zhu, X.; Gong, S. Open logo detection challenge. *arXiv* **2018**, arXiv:1807.01964.
48. Romberg, S.; Pueyo, L.G.; Lienhart, R.; Van Zwol, R. Scalable logo recognition in real-world images. In Proceedings of the ACM International Conference on Multimedia, Scottsdale, AZ, USA, 28 November–1 December 2011; pp. 1–8.
49. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv* **2019**, arXiv:1906.07155.
50. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
51. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. 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 (CVPR), Virtual Platform, 14–19 June 2020; pp. 9759–9768.
52. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized Focal Loss: Learning qualified and distributed bounding boxes for dense object detection. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21002–21012.
53. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. 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 (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 658–666.
54. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, Hilton New York Midtown, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
55. Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. Sparse R-CNN: End-to-end object detection with learnable proposals. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 14454–14463.
56. Kong, T.; Sun, F.; Liu, H.; Jiang, Y.; Li, L.; Shi, J. FoveaBox: Beyound anchor-based object detection. *IEEE Trans. Image Process.* **2020**, *29*, 7389–7398. [CrossRef]
57. Oliveira, G.; Frazão, X.; Pimentel, A.; Ribeiro, B. Automatic graphic logo detection via fast region-based convolutional networks. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 985–991.
58. Iandola, F.N.; Shen, A.; Gao, P.; Keutzer, K. DeepLogo: Hitting logo recognition with the deep neural network hammer. *arXiv* **2015**, arXiv:1510.02131.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.