Retrieval-based methodology for few-sample logo recognition

1st Dimosthenis Moralis
Dept. of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
dmoralis@csd.auth.gr

2nd Maria Tzelepi
Dept. of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
mtzelepi@csd.auth.gr

3rd Anastasios Tefas
Dept. of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
tefas@csd.auth.gr

***Abstract*—Logo recognition describes the challenging task of detecting and classifying logos in digital images and videos. Former works approach logo recognition as a closed-set problem. However, this approach is accompanied by several shortcomings linked with its incapability of recognising new classes. In this paper, we propose an open-set logo recognition method, named *REtrieval-based methodology For FEw-sample LOgo Recognition* (REFELOR). REFELOR is composed by a generic logo detector and a feature extractor, allowing the generalization on unseen classes, using only a few samples per logo. That is, a single-stage generic logo detector is trained to detect logos in an input image. Then, feature representations for the detected logos are extracted, using the feature extractor, while the feature representations of a database containing only a few samples per class are also extracted. Finally, the detected logo representations are classified to the corresponding class based a similarity search in the representations of the aforementioned database. In addition, a regularization technique is applied to the feature extractor, providing further improvements. The experimental evaluation validates the effectiveness of the proposed method, outperforming current state-of-the-art logo recognition methods.**

***Index Terms*—logo recognition, retrieval, few-sample, logo detection, yolo, regularization**

I. Introduction

Logo recognition refers to the process of identifying a specific logo within an image or video using computer vision techniques. It is linked with many applications, such as brand gathering [1], copyright protection [2] and augmented reality [3]. Logo recognition is a challenging task due to the uniqueness of each logo, which can vary in terms of shape, color, typography, and imagery, introducing inter-class variability. Additionally, there may be different versions of the same logo, introducing intra-class variability, making it difficult to classify them all in the same class.

The task of logo recognition is approached in the literature as a closed set or open set problem. In the former approach, a model is trained on a fixed set of classes and it cannot classify examples belonging to different classes. On the contrary, in open set approach, a model is capable of predicting examples belonging to classes that are not included in the original training set. That is, open set techniques allow generalization on a wide range of unseen classes and even to entirely dissimilar classes, rendering them suitable for real-world scenarios.

In this work, we propose an open set methodology for logo recognition, called *REtrieval-based methodology For FEw-sample LOgo Recognition* (REFELOR). More specifically, REFELOR is inspired by methodologies proposed for addressing the face recognition task [4]. The proposed methodology, as shown in Fig. 1, consists of two main components: a single-stage detector and a feature extractor. The generic logo detector is used to identify regions of logos in input images or videos. Then the feature extractor is fine-tuned on the database consisting of only a few samples per class, and it is used to extract features both from detected logos, and the database of logos of interest. Each class of the database is represented by the mean vector of their feature representations. The logos are then classified in the same class as their most similar class of the database, in terms of a similarity metric. It should be noted that the feature extractor is trained with a classification loss, in contrast to other open-set methods that utilize triplet-based losses, facilitating the fine-tuning with only a few samples per class.

Furthermore, a regularization technique is applied on the model used for feature extraction, in order to increase the discrimination ability of the extracted features. The so-called Discriminant Analysis (DA) regularizer, [5] encourages the feature representations of each class to approach their class centers, improving the generalization ability of the model.

The proposed methodology, which includes training first a generic logo detector and using then a feature extractor and a similarity search process, is appropriate for real-world scenarios. More specifically, considering closed-set approaches, when new classes are introduced, the whole recognizer needs to be retrained for incorporating new classes, requiring also a large number of training samples, rendering them time-consuming and computationally expensive. On the contrary, the proposed methodology is more efficient, since when new classes are introduced, it uses only a few samples per logo to fine-tune the model used for feature extraction. It should be emphasized that the step of fine-tuning is not mandatory in the proposed methodology, however it is highly recommended since it provides remarkable improvements. The proposed methodology, in combination with DA regularization applied to the feature extractor, as it is experimentally validated, provides exceptional performance, outperforming current state-of-

the-art logo recognition methods, being at the same time more flexible and efficient.

The rest of the paper is organized as follows. A review of previous work is provided in Section II, followed by a detailed description of the proposed REFELOR method in Section III. Subsequently, the experimental evaluation is provided in Section IV, followed by the conclusions in Section V.

II. Prior Work

Logo recognition constitutes a vivid topic of research for several decades [6]. Former approaches involve the extraction of geometric features, such as lines, vertices, and curves, combined with localization algorithms, such as sliding window algorithms [7]. However, these methods can be sensitive to variations in object geometry and may not be suitable for comparing similar objects. Next, a wide range of methods based on techniques, such as bag-of-words [8] have also been proposed. Despite the potential in non deep learning (DL) techniques for logo recognition, these methods are time-consuming with limited generalization ability and poor performance on large and complex datasets.

Therefore, during recent years, DL methods have been applied for addressing the logo recognition task, providing considerable performance due to their ability to capture complex patterns and features in images [9], [10]. Early DL approaches consider closed set techniques, which rely on region proposal algorithms to identify potential logo regions in an image, followed by a combination of pre-trained Convolutional Neural networks (CNNs) and Support Vector Machine classifiers [9].

More recently, open-set approaches have been developed for logo recognition. These approaches often use triplet-based losses with proxies [11] for training a feature extractor. A more recent work proposes a more complex feature extractor that extracts both visual features using a CNN and text features using an attention head for improved performance [12]. On the contrary, in this work we use a small-sized feature extractor that is pre-trained on a large and diverse set of logos. To train the model, we apply a linear layer on top of the feature extractor and treat it as a classification problem. This design allows for simple and efficient updating of the model when new logo classes are introduced, unlike triplet-based loss techniques that require significant effort in the selection and balancing of appropriate triplets. In addition, we utilize a regularization technique, improving the generalization ability of the model and enhancing its performance on the entire pipeline.

III. Proposed Method

In this work, we propose an adaptive methodology for logo recognition, named REFELOR, drawing inspiration from face recognition techniques. More specifically, the proposed pipeline, illustrated in Fig. 1, consists of a generic logo detector that detects all possible logos in an image and outputs their coordinates. These coordinates are used to crop the detected logos from the original input image. Then features are extracted from these cropped images using a fast and efficient CNN, which is fine-tuned on a database of logos of interest. The database contains only a few sample per class. The mean feature vector is considered for each of the classes. Finally, the process includes a similarity search for classifying each logo to the most similar class representation.

It should be noted that if new classes are introduced to the problem, it is not necessary to retrain the detector nor the feature extractor. The only required action is to add a few samples for the new logo classes in order to calculate their mean feature vector in the database. However, it is recommended to fine-tune the feature extractor on the new classes using these few samples, as this simple action can significantly improve the performance of the feature extractor. This makes this technique easily adaptable to new data without being computationally expensive or time-consuming.

In the following subsections the *Generic Logo Detector* is first presented, namely YOLOv5, which is adapted and trained for logo detection. Next, the *Regularized Feature Extractor* follows, that is the model used for feature extraction along with the utilized regularization technique, which enhances its discrimination ability. Finally, REFELOR, the whole proposed method is presented.

A. Generic Logo Detector

We use the medium-sized version of the successful YOLOv5 [13] detector (abbreviated as YOLOv5m), which is adapted and trained to serve as the generic logo detector. Different detectors could also be used [14]. The input images are of size 640×640, and the network architecture includes a CSP-Darknet53 backbone, which is a Darknet53 [15] network with a Cross Stage Partial (CSP) [16] design, followed by an improved SPP module [17] called SPPF, a PANet module [18], and three convolution heads specialized in small, medium, and large logo detection, respectively. The training of YOLOv5m is conducted using a dataset with binary labels logo/non-logo. It is worth noting that the inclusion of a diverse range of logo classes in the training dataset is beneficial for improving the performance of the detector. It should also be mentioned that the YOLOv5m detector is pre-trained on the well-known Common Objects in Context (COCO [19]) dataset.

B. Regularized Feature Extractor

The feature extractor of the proposed pipeline is MobileFaceNet [20], a successful small-scale CNN, based on MobileNetV2. The network is designed to accept truncated logos of size 64×64 as input and generate 128-dim feature vectors. The model provides both compact and efficient logo representations. A linear layer is appended to the network, allowing for training the network using Cross Entropy (CE) loss on the database of logos of interest. Note that we have used MobileFaceNet against MobileNet, since it performed better and faster.

During training, we apply a supervised regularization method, namely DA [5], which has been initially proposed for improving the generalization ability of lightweight CNN

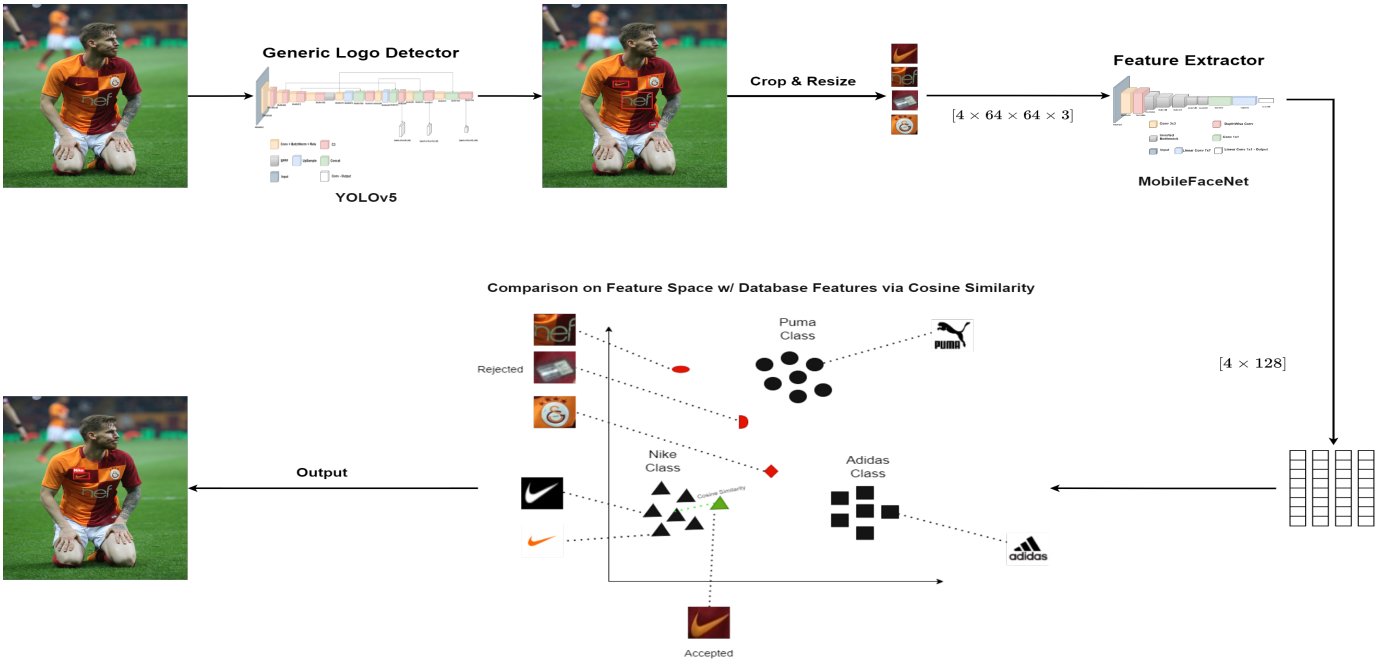

Fig. 1. Pipeline of REFELOR: First the generic logo detector (YOLOv5) is applied to an input image containing logos. The detected logos are cropped and resized to 64×64 and propagated to the feature extractor (MobileFaceNet) to transform them into 128-dim vectors. Then, comparison with the mean vectors of each logo class of the database is performed. The logos are finally classified based on their similarity to the class representation.

models on generic classification tasks. DA regularizer is inspired by the Linear Discriminant Analysis [21] algorithm, which aims to project samples of each class into a lower-dimensional space while maximizing the separation between classes and minimizing the within-class variability. Thus, the DA regularizer aims to minimize the within-class variability while the main supervised loss function preserves the between class separability.

In this work, we apply the aforementioned regularizer in order to improve the performance of the logo feature extraction model, i.e., MobileFaceNet. Let $\mathcal{X} \subseteq \mathcal{R}^d$ be an input space and $\mathcal{F} \subseteq \mathcal{R}^d$ be an output set, and also let the MobileFaceNet model be defined as $\phi(\cdot; \mathcal{W}) : \mathcal{X} \rightarrow \mathcal{F}$, where $N_L \in \mathbb{N}$ is the number of layers and $\mathcal{W} = \{W_1, ..., W_{N_L}\}$ are the weights, with W_l being the weights of layer l. The database of logo images is defined as $\mathcal{D}_{\mathcal{N}} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, their corresponding labels as $\mathcal{Y}_{\mathcal{N}} = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$, and the representations of these logos on layer l of the MobileFaceNet are given by $\phi(\mathbf{x}_i; \mathcal{W}^l)$. Finally, we consider the set $\mathcal{S}^i = \{\mathbf{x}_k, k = 1, ..., K^i\}$ of K^i logos that belong to the same class with the i-th logo.

Therefore, the objective of the DA regularizer can be formulated as follows:

$$\min_{\mathcal{W}^l} \mathcal{J}_{DA} = \min_{\mathcal{W}^l} \sum_{i=1}^{N} ||\phi(\mathbf{x}_i; \mathcal{W}^l) - \mu_i||_2^2 \tag{1}$$

where $\mu_i = \frac{1}{|\mathcal{S}^i|} \sum_{\mathbf{x}_j \in \mathcal{S}^i} \phi(x_j; \mathcal{W}^l)$.

It should be noted that the DA regularizer can be applied to any layer. In this work, it is applied to the last layer ($l = N_L$) of MobileFaceNet.

As previously mentioned, a linear layer is added on the top of MobileFaceNet, formulating a classification problem, considering M logo classes of interest. The CE loss if used for realizing the training. Considering as $\hat{\mathbf{y}}_i, i = 1, ...N$ the outputs of the linear layer of the network, the classification objective is formulated as:

$$\mathcal{J}_{CE} = -\sum_{i}^{N} \mathbf{y}_i \log \hat{\mathbf{y}}_i. \tag{2}$$

Therefore, the model is trained both with the supervised loss and the regularization objective:

$$\mathcal{J} = \lambda \mathcal{J}_{DA} + \mathcal{J}_{CE}, \tag{3}$$

where parameter λ controls the relative importance of the contributed losses. Through minimization of the objective of eq. (3), the utilized feature extractor produces representations with enhanced discrimination ability.

C. REFELOR

In this Section we conclude the proposed REFELOR methodology for logo recognition. As it is already mentioned, the proposed methodology includes a generic logo detector and a feature extractor. The feature extractor is used to represent the database containing the logos of interest (mean vector representations), and the logos detected using the generic logo detector. Thus, the final step includes a search on the feature space, using as similarity metric the cosine similarity (various similarity metrics could also be used), in order to classify the detected logos to the most similar classes.

Thus, let $\mathcal{R} = \{r_1, ..., r_M\}$, be the mean vector representation of each of M classes of interest, and $\mathbf{v}$, the representation of a logo detected by the generic logo detector, all extracted using the MobileFaceNet. The vector of the detected logo is then compared with the database $\mathcal{R}$, using cosine similarity:

$$cos\_sim(\mathbf{v}, \mathbf{r}_j) = \frac{\mathbf{v} \cdot \mathbf{r}_j}{||\mathbf{v}|| \cdot ||\mathbf{r}_j||}, j = 1, \cdots M. \quad (4)$$

Then, the detected logo $\mathbf{v}$ is classified to the corresponding class of the most similar mean vector $\mathbf{r}_j$. Finally, a threshold value is incorporated into the technique, which defines the minimum value of cosine similarity between the logo in the input and the closest logo in the database. If this threshold is surpassed, the logo is deemed unknown and rejected.

IV. Experimental Evaluation

In this Section, we present the experimental evaluation of the proposed method. We first describe the datasets utilized for training the generic detector and feature extractor. Next, we present the evaluation metrics used to assess the performance of the proposed method. Subsequently, we provide details on the implementation and experimental setup, and we finally present and discuss the experimental results.

A. Datasets

In this work, we utilize two publicly available and well-known datasets. The first one, Logodet3K [22], which is used for training the generic logo detector, consists of 3,000 distinct logos, 156,652 images that contain a total of nearly 200,000 instances of logos. However, when training the generic detector YOLOv5m, we only use the information about the logo and non-logo classes, rather than the specific class of each logo. To obtain the training and evaluation sets, we divide the dataset into two subsets based on the logo classes: 80% of the classes are used as the training set, while the remaining 20% are used as the validation set. This allows us to evaluate the detector's performance on logos that it has not seen during training, and to determine when its generalization capability begins to degrade.

The second dataset utilized for the recognition task is the Flickr32 [23]. Flickr32 comprises of 32 logo classes. It includes a small training set of 320 images, a validation set of 3,960 images, and a test set of 3,960 images. It is noteworthy that the 3,000 images within the validation set and the test set are background images, which are characterized as *building*, *nature*, *people*, and *friends*. This dataset, and in particular its test set, is utilized in order to evaluate the performance of REFELOR, while its binary version that contains only the logo/no-logo labels, is used to evaluate the generic detector.

From these two datasets, we create two classification datasets. Logodet3K classification dataset with 3,000 classes and 200,000 logo images, and a training set of 176,000 (80%) logo images, and a validation set of 24,000 (20%) images. Note that as explained below, Logodet3k can be used for pre-training the MobileFaceNet on its 3,000 classes, while the evaluation of the pipeline is performed on Flickr32, considering its 32 logo classes. The Flickr32 classification dataset has an extremely small training set of 320 images and evaluation and test sets of 960 images each. This dataset is used for fine-tuning MobileFaceNet and evaluating the impact of DA. The training set of this dataset is also used to create REFELOR's database by extracting features from all samples of each class and calculating their mean vectors.

B. Evaluation Metrics

In this Section the evaluation metrics of the generic logo detector, the feature extractor and the whole pipeline of logo recognition are provided.

In order to evaluate the performance of the generic detector and the entire pipeline, the Mean Average Precision (mAP) metric is used, which is a common metric for object detectors. Specifically, we utilize a IOU threshold of 0.5 for all evaluations (denoted as mAP@0.5). The performance of the classifier and the DA regularizer are evaluated using the accuracy metric. Finally, the performance of REFELOR's inference speed is evaluated using the metric of Frames Per Second (FPS).

C. Implementation Details

The proposed methodology is implemented using the Pytorch framework. The YOLOv5m detector is fine-tuned for 15 epochs on the Logodet3K dataset with a learning rate of $1e^{-2}$, linear decay to $1e^{-4}$, SGD optimizer, and weight decay of $5e^{-4}$. Augmentation techniques like Mosaic and MixUp are applied to improve the training set. The best model weights are selected based on the highest mAP value in the Logodet3K validation set.

The MobileFaceNet is pre-trained on the Logodet3K train classification set using CE Loss, with 300 epochs, a batch size of 512, a learning rate of $1e^{-1}$ with linear decay to $1e^{-3}$, and simple augmentations. For the scenario of fine-tuning with few samples, MobileFaceNet is then fine-tuned on the Flickr32 train classification set for 120 epochs with a learning rate of $1e^{-3}$, SGD optimizer, and batch size of 64. The best weights are selected based on the highest accuracy on the Flickr32 validation set. Random Rotation and Random Perspective augmentations are applied during this stage. The DA regularizer with a weight of $\lambda = 5$ is incorporated to boost the model's performance and generalization ability.

D. Experimental Setup

In this Section, we present the experiments conducted in order to evaluate the proposed method for logo recognition. Five sets of experiments were conducted.

In the first set of experiments we evaluate the performance of the general logo detector. The evaluation is done on Logodet3K validation set and Flickr32 test set, using the metric mAP@0.5. It is also important to say that, in terms of evaluation, logo detections were allowed to overlap by 20%, and only logos with a detection confidence greater than 0.1% were displayed.

In the second set of experiments, we evaluate the performance of the feature extractor after fine-tuning using the accuracy metric on the Flickr32 validation and test classification sets. Additionally, we investigate the impact of the DA

regularizer on the model's accuracy by comparing the results of training with and without it on the aforementioned Flickr32 classification sets, using 10 repetitions to ensure a reliable outcome.

The third set of experiments evaluates the performance of the REFELOR method on the Flickr32 test set using the mAP@0.5 metric. In the first experiment, REFELOR's performance is compared in three scenarios: without fine-tuning MobileFaceNet, with fine-tuning on the few-sample Flickr32 training set, and with fine-tuning while using the DA regularizer. The database and the optimal threshold for similarity search used in each scenario vary as they are based on features extracted from MobileFacenet, which is affected by fine-tuning and the use of the DA regularizer. The second experiment compares the best-performing scenario from the first experiment to state-of-the-art logo recognition methods, which use the same evaluation set and open-set architectures.

In the fourth set of experiments, the efficiency of the proposed method is evaluated, considering the inference speed in term of FPS for different GPUs. Finally, in the fifth set of experiments a qualitative analysis of the pipeline's results on the Flickr32 test set is presented, providing a comprehensive understanding of the pipeline's performance.

E. Experimental Results

The results of the first set of experiments are shown in Table I. The detector performs better on the Logodet3K set, however its performance on the Flickr32 set is still satisfactory.

TABLE I
GENERIC LOGO DETECTION EVALUATION

Dataset	mAP@0.5
Logodet3K	69.6

The results of the second set of experiments for fine-tuning and implementing the DA regularizer are shown in Table II. We can observe roughly 1% improvement in performance on both sets, which confirms the effectiveness of DA regularizer.

TABLE II
EVALUATION OF THE CLASSIFIER

Model	Val Accuracy	Test Accuracy
MobileFaceNet	91.6 ± 0.08	92.42 ± 0.26
MobileFaceNet + DA	$\mathbf{92.61 \pm 0.07}$	$\mathbf{93.35 \pm 0.14}$

In the third set of experiments, the performance of the entire pipeline is evaluated. First the three scenario results of the first experiment are shown in the Table III. The first scenario achieves a mAP@0.5 of 57.9 which is a satisfactory result for our approach without any fine-tuning with the samples of the dataset of interest. The second scenario achieves a mAP@0.5 of 60.4 which is a 2.5 improve in the performance of REFELOR by fine-tuning the model with few samples. Finally, in the third scenario, we achieve an increase of 1.6%, resulting in a mAP score of 62, through the use of the DA regularizer, demonstrating that the feature extractor exhibits improved generalization capabilities. It is important to note the significant improvement in mAP compared to the generic detector stage. This occurs because the detector suggests many regions as candidate logo regions, generating many False Positives. These False Positives are filtered out with the introduction of MobileFaceNet, as they do not resemble any logo in our database. This significantly improves the Precision of the pipeline, also leading to a significant increase in the overall mAP.

TABLE III
LOGO RECOGNITION EVALUATION

Model	mAP@0.5
YOLOv5 + MobileFaceNet	57.9
YOLOv5 + MobileFaceNet (Finetuned)	60.4
YOLOv5 + MobileFaceNet (Finetuned + DA)	**62**

In the second experiment of the same set of experiments, the proposed method is compared to two recent works, as shown in Table IV, which also use the same evaluation set and an open set architecture. The first work uses Faster R-CNN for the generic detector and SE-Resnet50 [24] for feature extraction trained with a combination of Proxy-Triplet Loss and a Spatial Transformer Network (STN) layer. Both the detector and extractor were trained on a custom logo dataset named PL2K. The second work uses a Receptive Field Block Network (RFBNet) [25] as the generic detector and DenseNet169 [26] as the feature extractor, which is trained by their Simple Deep Metric Learning (SDML) technique. Both sets are trained on a custom dataset named OSLD. It achieves its performance by fine-tuning the generic RBFNet detector on the Flickr32 detection training set. The comparison results indicate that the proposed REFELOR method demonstrates superior performance as compared to state-of-the-art works. This is attributed to the efficiency of our feature extractor and the effectiveness of the DA regularizer. Furthermore, by performing a small 10 epoch finetune on the generic detector on the Flickr32 training set with binary logo/no-logo tags, as done in the work [27], we can achieve a mAP@0.5 of 53.3% for the generic detector and an overall mAP of 76.5% for the whole pipeline. Despite this improvement, it is recommended to maintain a fixed generic detector and utilizing a larger training set would further enhance the pipeline's generic performance.

TABLE IV
PIPELINE COMPARISON WITH STATE OF THE ART PAPERS

Method	mAP@0.5
Faster R-CNN + SE-Resnet50 [11]	44.42
RBFNet + DenseNet169 [27]	60.6
REFELOR	**62**

Next, we report the results of the fourth set of experiments, considering the inference speed, in terms of FPS. The experiments concern the Flickr32 set, where the detector detects, on average, 35 bounding boxes per image. REFELOR runs at 45

FPS on an RTX 3080, and at 35 FPS on an RTX 2080, that is real-time.

The results of the fifth set of experiments are illustrated in Fig. 2, where several examples of the REFELOR method validate its effectiveness.

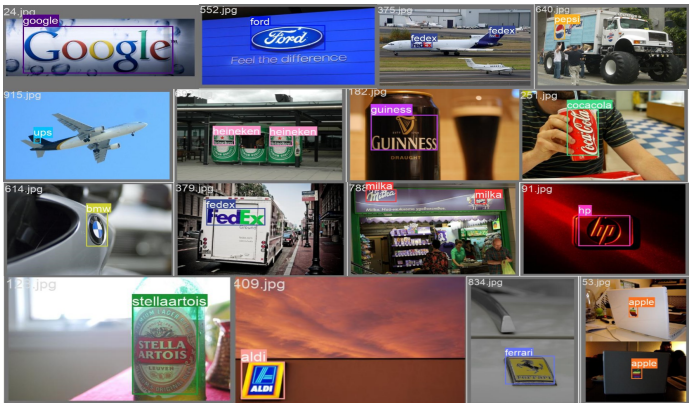

Fig. 2. Qualitative Results on Flickr32 test Set

V. Conclusions

In this paper, we proposed REFELOR, an open-set method for logo recognition that is able to generalize to unseen classes with just a few samples per logo, without the need for retraining. Furthermore, REFELOR showed an increased performance with fine-tuning on only a few-sample per logo dataset and additional improvement through the use of the DA regularizer. The proposed method achieved superior results against state-of-the-art logo recognition methods, validating its effectiveness.

Acknowledgment

This research was funded by the project "SEMANTIC ANNOTATION AND METADATA ENRICHMENT OF OPEN VIDEO STREAMS USING DEEP LEARNING" (Project code: KMP6-0079092) that was implemented under the framework of the Action "Investment Plans of Innovation" of the Operational Program "Central Macedonia 2014 2020", that is co-funded by the European Regional Development Fund and Greece.

References

[1] Y. Gao, F. Wang, H. Luan, and T.-S. Chua, "Brand data gathering from live social media streams," in *Proceedings of International Conference on Multimedia Retrieval*, ser. ICMR '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 169–176. [Online]. Available: https://doi.org/10.1145/2578726.2578748

[2] Z. Li, "The study of security application of logo recognition technology in sports video," *EURASIP Journal on Image and Video Processing*, vol. 2019, pp. 1–10, 2019.

[3] N. Hagbi, O. Bergig, J. El-Sana, and M. Billinghurst, "Shape recognition and pose estimation for mobile augmented reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, pp. 1369–1379, 2009.

[4] Q. Wang, P. Zhang, H. Xiong, and J. Zhao, "Face. evolve: A high-performance face recognition library," *arXiv preprint arXiv:2107.08621*, 2021.

[5] M. Tzelepi and A. Tefas, "Graph embedded convolutional neural networks in human crowd detection for drone flight safety," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 2, pp. 191–204, 2019.

[6] D. S. Doermann, E. Rivlin, and I. Weiss, "Logo recognition using geometric invariants," in *Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR'93)*. IEEE, 1993, pp. 894–897.

[7] Y. Lamdan, J. Schwartz, and H. Wolfson, "Object recognition by affine invariant matching." in *Proc CVPR 88 Comput Soc Conf on Comput Vision and Pattern Recognit*, ser. Proc CVPR 88 Comput Soc Conf on Comput Vision and Pattern Recognit. Publ by IEEE, 1988, pp. 335–344.

[8] S. Yu, S. Zheng, H. Yang, and L. Liang, "Vehicle logo recognition based on bag-of-words," in *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2013, pp. 353–358.

[9] S. Bianco, M. Buzzelli, D. Mazzini, and R. Schettini, "Logo recognition using cnn features," in *International Conference on Image Analysis and Processing*. Springer, 2015, pp. 438–448.

[10] ——, "Deep learning for logo recognition," *Neurocomputing*, vol. 245, pp. 23–30, 2017.

[11] I. Fehérvári and S. Appalaraju, "Scalable logo recognition using proxies," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 715–725.

[12] C. Li, I. Fehérvári, X. Zhao, I. Macedo, and S. Appalaraju, "Seetek: Very large-scale open-set logo recognition with text-aware metric learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2544–2553.

[13] G. J. et al., "ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation," Nov. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7347926

[14] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, 2023.

[15] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[16] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 390–391.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[18] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[20] S. Chen, Y. Liu, X. Gao, and Z. Han, "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices," in *Chinese Conference on Biometric Recognition*. Springer, 2018, pp. 428–438.

[21] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[22] J. Wang, W. Min, S. Hou, S. Ma, Y. Zheng, and S. Jiang, "Logodet-3k: A large-scale image dataset for logo detection," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 1, pp. 1–19, 2022.

[23] S. Romberg, L. G. Pueyo, R. Lienhart, and R. Van Zwol, "Scalable logo recognition in real-world images," in *Proceedings of the 1st ACM international conference on multimedia retrieval*, 2011, pp. 1–8.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[25] S. Liu, D. Huang *et al.*, "Receptive field block net for accurate and fast object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 385–400.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[27] M. Bastan, H.-Y. Wu, T. Cao, B. Kota, and M. Tek, "Large scale open-set deep logo detection," *arXiv preprint arXiv:1911.07440*, 2019.