Deep Learning Logo Detection with Data Expansion by Synthesising Context

Hang Su Xiatian Zhu Shaogang Gong
Shchool of EECS, Queen Mary University of London, United Kingdom
hang.su@qmul.ac.uk, xiatian.zhu@qmul.ac.uk, s.gong@qmul.ac.uk

Abstract

Logo detection in unconstrained images is challenging, particularly when only very sparse labelled training images are accessible due to high labelling costs. In this work, we describe a model training image synthesising method capable of improving significantly logo detection performance when only a handful of (e.g., 10) labelled training images captured in realistic context are available, avoiding extensive manual labelling costs. Specifically, we design a novel algorithm for generating Synthetic Context Logo (SCL) training images to increase model robustness against unknown background clutters, resulting in superior logo detection performance. For benchmarking model performance, we introduce a new logo detection dataset TopLogo-10 collected from top 10 most popular clothing/wearable brandname logos captured in rich visual context. Extensive comparisons show the advantages of our proposed SCL model over the state-of-the-art alternatives for logo detection using two real-world logo benchmark datasets: FlickrLogo-32 and our new TopLogo-10[1].

1. Introduction

Logo detection is a challenging task for computer vision, with a wide range of applications in many domains, such as brand logo recognition for commercial research, brand trend research on Internet social community, vehicle logo recognition for intelligent transportation [33, 31, 32, 5, 23, 28]. For generic object detection, deep learning has been a great success [30, 29, 35]. Building a deep object detection model typically requires a large number of labelled training data collected from extensive manual labelling [22, 36]. However, this is not necessarily available in many cases such as logo detection where the publicly available datasets are very small (Table 1). Small training data size is inherently inadequate for learning deep models with millions of parameters. Increasing manual annotation is extremely costly [16] and unaffordable in most cases, not only in monetary but more critically in timescale terms.

Figure 1. Examples of logo exemplar images.

Table 1. Existing logo detection datasets. PA: Public Availability. Apart from labelled logo images, BelgaLogos provides 8,049 images with no bounding boxes and FlickrLogos-32 has 6,000 non-logo images.

Dataset	Logo #	Object #	Image #	PA
BelgaLogos [20]	37	2,695	1,951	Yes
FlickrLogos-27 [21]	27	4,671	1080	Yes
FlickrLogos-32 [33]	32	5,644	2,240	Yes
LOGO-NET [16]	160	130,608	73,414	No

In the current literature, most existing studies on logo detection are limited to small scales, in both the number of logo images and logo classes [18, 23], largely due to the high costs in constructing large scale logo datasets. It is non-trivial to collect automatically large scale logo training data that covers a large number of different logo classes. While web data mining may be a potential solution as shown in other recognition problems [24, 6, 7, 35], it is difficult to acquire accurate logo annotations since no bounding box annotation is available from typical web images and their meta-data.

In this work, we present a novel synthetic training data generation algorithm for improving the learning of a deep

[1] The TopLogo-10 dataset is available at: http://www.eecs.qmul.ac.uk/~hs308/qmul_toplogo10.html

logo detector with only sparsely labelled training images. This approach enlarges significantly the variations of both logo and its context in the training data without increasing manual labelling effort, so that a deep detector can be optimised to recognise the target logos against diverse and complex background clutters not captured by the original sparse training data. The *contributions* of this work are: (1) We formulate a novel Synthetic Context Logo (SCL) training data generation method for learning a logo detector given sparsely labelled images. Unlike typical deep learning models, we do not assume the availability of large quantities of labelled training images but only a handful. Our model is designed specifically to augment the training data by enriching and expanding both logos and their context variations. To our knowledge, it is the first attempt of exploiting large scale synthetic training *data expansion in context* for deep learning a logo detection model. (2) We introduce a large scale automatically synthesised logo dataset, in rich context with labelled logo bounding boxes, consisting of 463 different logos (Figure 1) which is much larger in class number than any existing logo benchmark datasets in the public domain. (3) We further introduce a new logo dataset TopLogo-10, manually collected and labelled from in-the-wild logo images. This TopLogo-10 dataset consists of 10 logo classes of high logo popularity (with high frequency in real-life [11, 1, 8, 17] in rich context, thus more challenging. We evaluated extensively the proposed SCL method for deep learning a logo detector against the state of the art alternatives using two logo benchmark datasets, and provided in-depth analysis and discussion on model generalisation.

2. Related Works

Logo Detection. Most existing approaches to logo detection rely on hand-crafted features, e.g., HOG, SIFT, colour histogram, edge [21, 33, 31, 32, 5, 23, 28]. They are limited in obtaining more expressive representation and model robustness for recognising a large number of different logos. One reason is due to the unavailability of sufficiently large datasets required for exploring deep learning a more discriminative representation. For example, among all publicly available logo datasets, the most common FlickrLogos-32 dataset [33] contains 32 logo classes each with only 70 images and in total 5644 logo objects, whilst BelgaLogos [20] has 2695 logo images from 37 logo classes with bounding box (bbox) location labelled (Table 1).

Nevertheless, a few deep learning based logo detection models have been reported recently. Iandola et al. [18] applied the Fast R-CNN model [13] for logo detection, which inevitably suffers from the training data scarcity challenge. To facilitate deep learning logo detection, Hoi et al. [16] built a large scale dataset called LOGO-Net by exhaustively collecting images from online retailer websites and then manually labelling them. This requires a huge amount of construction effort and moreover, LOGO-Net is inaccessible publicly. In contrast to all existing attempts above, we explore the potentials for learning a deep logo detection model by synthesising a large scale sized training data to address the sparse data annotation problem without additional human labelling cost. Compared to [16], our method is much more cost-effective and scalable for logo variations in diverse visual context, e.g., accurate logo annotation against diverse visual scene context can be rapidly generated without any manual labelling, and potentially also for generalising to a large number of new logo classes with minimal labelling.

Synthesising Data Expansion. Generating synthetic training data allows for expanding plausibly ground-truth annotations without the need for exhaustive manual labelling. This strategy has been shown to be effective for training large CNN models particularly when no sufficient training data are available, e.g., Dosovitskiy et al. [9] used synthetic floating chair images for training optical flow networks; Gupta et al. [14] and Jaderberg et al. [19] generated scene-text images for learning text recognition models; Yildirim et al. [38] exploited deep CNN features optimised on synthetic faces to regress face pose parameters. Eggert et al. [10] applied synthetic data to train SVM models for company logo detection, which shares the spirit of the proposed method but with an essential difference in that we explore synthetic training data *in diverse context variations* in the absence of large scale realistic logo dataset in context.

3. Synthetic Logos in Context

Typically, a large training dataset is required for learning an effective deep network [22]. This however is very expensive to collect manually, particularly when manual annotation of locations and bounding boxes of varying-sized objects are needed, e.g., logos in natural images from the wild street views [16]. Given the small size labelled training data in existing logo datasets (Table 1), it is challenging to learn effectively the huge number of parameters in deep models. To solve this problem, we synthesis additional training data. We consider that logo variations are largely due to change of visual context and/or background plus geometric and illumination transforms. We then develop a *Synthetic Context Logo* (SCL) image generation method for deep learning a logo detection model. By doing so, our model is capable of automatically generating infinite numbers of synthetic labelled logo images in realistic context with little supervision, making deep learning feasible.

In model learning, we exploit a sequential learning strategy by first deploying a large number of synthesised images to pre-train a deep model, followed by fine-tuning the deep model with the sparse manually annotated images. This sequential model training strategy echoes the principle of

Figure 2. Comparing the visual effect of our logo exemplar with transparent background (**left**) and that of [10] with non-transparent background (**right**) in the synthetic logo images. Evidently our logo exemplar allows more diverse context than [10], which imposes a surrounding appearance in the synthetic image. More natural appearance is provided by our synthetic image.

Table 2. The popularity of TopLogo-10 brand logos as ranked by online media reports, among which not all brands were covered in every report. Smaller numbers indicate higher popularity.

Criterion	Clothing [11]	Luxury [1]	Fashion [8]	Streetwear [17]
Adidas Classic	3	-	-	-
Helly Hansen	-	-	1	-
Gucci	5	3	-	-
Nike	1	-	8	-
Lacoste	9	-	-	-
Chanel	-	4	7	-
Puma	11	-	-	-
Micheal Kors	-	9	-	-
Prada	-	7	-	-
Supreme	-	-	-	2

Curriculum Learning (CL), designed for reducing the difficulties in learning tasks by easy-to-hard staged learning [4]. In learning a logo model, we consider learning from a small number of real images against cluttered background captured in the wild is a more challenging learning task than learning from a large number of synthetic images. We evaluated in our experiments the effectiveness of this staged learning strategy against model training based on a fusion of synthetic and manually annotated images (Section 4.5).

3.1. Logo Exemplar and Context Images

To synthesise images for a given logo class, we need an exemplar image for each logo class. This is obtained from Google Image Search with the corresponding logo name as the query keyword. To minimise any context bias in the synthetic images, we utilised those exemplars with pure logo against a homogeneous transparent background (Figure 1). As such, the surrounding pixels around a logo in any synthetic images are completely determined by the background image, rather than the exemplar images. This is very different from [10] where (1) pixel-level logo masks are extracted by tedious manual annotation and (2) the appearance of nearby pixels are biased largely towards the source images thus may inevitably break the diversity of surrounding context (see Figure 2 for example).

Figure 3. Illustration of logo exemplar transformations: (a) scaling; (b) shearing; (c) rotation; (d) colouring. We exploited these transformations jointly for synthesising new training images.

Logo Selection. We selected logos by considering: **(1)** Top popular and luxury brands of clothing/wearable brandname logos based on recent online media reports. Specifically, we particularly explored the clothing brand voting ranking [11] where the brands "Nike", "Adidas classic", "Lacoste" and "Puma" were selected; and the luxury brand ranking [1] where "Gucci", "Chanel", "Prada" and "Michael Kors" were chosen. Also, we picked the top jackets and sports brand "Helly Hansen" [8] and one of the most popular street clothing brand "Supreme" [17]. Table 2 summarises the popularity ranking of these 10 brand logos. **(2)** All 32 logo classes in the popular FlickrLogo-32 dataset. **(3)** Common logos/brands in real-life, such as software, computer, website, food, university. In total, we have 463 logos from a wide range of products, things and places in diverse context. Example logos are shown in Figure 1.

Context Images. For context images, we used the $6,000$ non-logo image subset of the FlickrLogo-32 dataset [33], obtained from the Flickr website by query keywords "building", "nature", "people" and "friends". These images (Figure 4) present highly diverse natural scene background in which logo may appear during synthesising training data. Moreover, this ensures that all synthesised logo instances are guaranteed to be labelled in the synthesised images, i.e., fully annotated with no missing logos.

3.2. Logo Exemplar Transformation

In addition to context variations, we consider other sources of logo variation are due to illumination change and geometric/perspective transforms. These factors are approximated by warping logo exemplar images. Formally, two independent types of transforms are deployed as follows [10].

Figure 4. Examples of generated synthetic context images. Given non-logo images from the FlickrLogo-32 dataset [33] as context images, these synthetic training images are fully annotated automatically without any logo instances unlabelled/missing.

(I) Geometric Transform. Suppose a logo image $\boldsymbol{I}$ is on a 3D plane, a general geometric transform of the image is computed as:

$$\boldsymbol{I}^* = \boldsymbol{P}\boldsymbol{I}\boldsymbol{R}_x\boldsymbol{R}_y \tag{1}$$

where the matrix $\boldsymbol{P}$ defines a *scaling* (Figure 3 (a)) and *shearing* (Figure 3 (b)) projections; $\boldsymbol{R}_x$ and $\boldsymbol{R}_y$ represent the *rotation* (Figure 3 (c)) matrices for the corresponding axis respectively, with the angles uniformly sampled from a range of $[0, 360]$. More precisely, we perform exemplar image geometric transforms in three steps as follows: (i) Random scaling of the exemplar images (Figure 3 (a)); (ii) Random shearing of the scaled images (Figure 3 (b)); (iii) Random rotation of the sheared images (Figure 3 (c)).

(II) Colouring Transform. We also modify the colour appearance of logo exemplar images for synthesising illumination variations. Specifically, we vary the pixel value in the RGB colour space as

$$c^* = rc \tag{2}$$

where the scalar c represents the pixel colour value in any channel and r a random number sampled uniformly from $[0, 2]$. The range of c^* is set to $[0, 255]$[2].

3.3. Synthesising Context Logo Images

Given the logo exemplar image transforms described above, we generate a number of variations for each logo, and utilise them to synthesis logo images in context by overlaying a transformed logo exemplar at a random location in non-logo context images. This randomness in logo placement provides a large variety of plausible visual scene context to enrich the synthesised logo training images. For every logo class, we generate 100 synthetic logo images in context through randomly selecting context images and applying geometric plus colouring transforms, resulting in 46,300 synthetic context logo training images. Examples of synthetic context logo images are shown in Figure 4.

[2]One may come across zero-valued (pure black) pixels in clean logo exemplars. In this case, a multiplication operation based transform is invalid. Instead, we simply set the pixel value to 100 before performing colour variation (Figure 3(d)).

Figure 5. Exemplar images of 32 logo classes (**Top**) and test image examples (**Bottom**) from the FlickrLogo-32 dataset [33].

Figure 6. Exemplar images of top 10 logos (**Top**) and test image examples (**Bottom**) from our TopLogo-10 dataset.

4. Experiments

4.1. Datasets

Two logo detection datasets were utilised for evaluation: FlickrLogo-32 [33] and TopLogo-10, newly introduced in this work.

FlickrLogo-32. This is the most popular logo detection dataset containing 32 different logos (Table 1). Examples of logo exemplars and test images are shown in Figure 5.

TopLogo-10. From 463 logos, we selected Top-10 clothing/wearable brandname logos by popularity/luxury, and constructed a logo dataset with manual labelling. We call this new dataset *TopLogo*-10. Specifically, there are ten logo classes: "Nike","Chanel", "Lacoste", "Gucci", "Helly Hansen", "Adidas Classic", "Puma", "Michael Kors", "Prada", "Supreme", with various degrees of composition complexity in the logos. For each logo class, 70 images are included, each with fully manually labelled logo bounding boxes. The exemplars and examples of test images are shown in Figure 6. These logo instances may appear in a variety of context, e.g., shoes, hats, shower gels, wallets, phone covers, lipsticks, eye glasses, spray, jackets, T-shirts, peaked caps, and sign boards. Moreover, logo instances in TopLogo-10 have varying sizes as in natural images, therefore imposes significant detection challenges from small sized logos. TopLogo-10 represents some of the common and natural logo existence scenarios and provides realistic challenges for logo detection.

For each of the two datasets, we divided randomly all the images into two parts: (1) one for training, with 10 images per logo class; (2) one for testing, with the remaining images. Given such a small number of labelled images, it is very challenging to train deep logo models with millions of parameters.

4.2. Baseline Methods

For evaluating the effectiveness of our synthetic context logo images for learning a deep logo detector, we utilised the Faster R-CNN [30] as our logo detector. Other object detectors [29, 25] are also available but this is independent from our proposed method. We trained a Faster R-CNN detector by the following different processes and comparatively evaluated their logo detection performances. **(1)** RealImg: Only labelled real training image are used for model training. **(2)** SynImg-xCls: The synthetic labelled training data from x ($x = 32$ for FlickrLogo-32 and $x = 10$ for TopLogo-10) target logo classes are used for model training; **(3)** SynImg-463Cls: The synthetic labelled training data from all 463 logo classes are used for model training; **(4)** SynImg-xCls+RealImg: We first utilise the synthetic training data from the target logo classes to pre-train a Faster R-CNN, then fine-tune the model using labelled real training data; this is a staged curriculum learning model training scheme [4]; **(5)** SynImg-463Cls+RealImg: Similar to SynImg-xCls+RealImg but with a difference that all synthetic training data are used for pre-training the detector; **(6)** SynImg-463Cls+RealImg (Fusion): Similar to SynImg-463Cls+RealImg but with a difference that the model is trained in a single step using the fusion of synthetic and realistic training data, other than the sequential curriculum learning.

Figure 7. Qualitative evaluation on FlickrLogo-32 [33]. **First row**: detection results by "RealImg"; **Second row**: detection results by "SynImg-463Cls + RealImg". Logo detections are indicated with red boxes. Ground truth is indicated by green boxes.

4.3. Evaluation Metrics

All models are trained on a training set and evaluated on a separate independent testing set. The logo detection accuracy is measured by Average Precision (AP) for each individual logo class and mean Average Precision (mAP) over all logo classes [30].

4.4. Implementation Details

For learning a logo Faster R-CNN detector, we set the learning rate as 0.0001 on either synthetic or real training data. The learning iteration is set to $40,000$. For all the models, we (1) pre-trained a Faster R-CNN with the training data of the ImageNet-1K object classification dataset [34] for parameter initialisation [36], and (2) then further fine-tuned the model on PASCAL VOC non-logo object detection images [12].

4.5. Evaluations

Evaluation on FlickrLogo-32. We performed logo detection on the FlickrLogo-32 logo images [33]. The results of different methods are shown in Table 3. It is evident that logo detection performance by Faster R-CNN can be largely improved by expanding the training data with the synthetic context logo images generated with our proposed method. For example, the combination of full synthetic and realistic training data (i.e., SynImg-463Cls + RealImg) can boost the detection accuracy from 50.4% (by RealImg) to 55.9%, with 5.5% increase in absolute mAP or 10.9% relative improvement. Importantly, such performance gain is achieved *without* any additional exhaustive labelling but by automatically enriching the context variations in the training data.

Specifically, we draw these following observations. *Firstly*, by using merely 10 training images per logo (i.e.,

Table 3. Evaluating different methods on the FlickrLogo-32 dataset [33]. RealImg: Realistic Image; SynImg: Synthetic Image.

Method	Setting: Training/Test Image Split (per logo class)	Adidas	Aldi	Apple	Becks	BMW	Carls	Chim	Coke	mAP
		Corona	DHL	Erdi	Esso	Fedex	Ferra	Ford	Fost	
		Google	Guin	Hein	HP	Milka	Nvid	Paul	Pepsi	
		Ritt	Shell	Sing	Starb	Stel	Texa	Tsin	Ups	
RealImg		23.7	57.5	63.0	69.6	63.7	50.6	55.2	26.8	
	Training: 10 RealImg	79.0	25.8	61.2	44.2	45.9	80.6	64.3	43.2	
		47.7	58.2	61.8	21.3	19.4	17.4	48.2	17.8	
	Test: 60 RealImg	34.8	**45.8**	71.8	70.2	79.6	56.7	56.9	52.2	50.4
SynImg-32Cls		9.4	47.3	9.6	70.3	39.9	28.3	15.8	21.7	
	Training: 100 SynImg	6.1	11.1	4.1	44.7	22.9	60.9	43.6	28.8	
		23.0	16.7	43.1	9.9	4.6	1.1	38.1	9.7	
	Test: 60 RealImg	22.7	38.3	15.5	65.6	28.7	55.1	27.4	20.1	27.6
SynImg-463Cls		9.2	24.8	4.5	30.5	24.1	15.3	2.4	20.9	
	Training: 100 SynImg	0.4	3.8	4.8	48.7	20.5	45.2	29.0	24.5	
		13.0	1.1	27.8	10.4	1.8	7.2	26.1	10.8	
	Test: 60 RealImg	18.6	38.6	0.7	49.3	28.5	61.1	23.7	17.5	20.5
		26.8	63.7	65.8	72.7	**81.3**	52.7	**63.6**	30.0	
SynImg-32Cls	Training: 100 SynImg	76.0	31.5	63.0	52.2	**54.3**	**90.0**	**84.0**	46.6	
+ RealImg	+10 RealImg	58.0	52.6	65.2	23.2	**24.0**	12.5	54.1	**23.6**	
	Test: 60 RealImg	37.9	45.6	**75.0**	**73.8**	79.0	**64.2**	57.4	**54.4**	54.8
		22.6	**66.6**	**72.0**	**73.2**	78.7	**53.3**	58.0	**31.2**	
SynImg-463Cls	Training: 100 SynImg	**82.7**	**33.7**	**67.2**	**53.5**	50.8	85.6	72.4	**51.3**	
+ RealImg	+ 10 RealImg	**59.6**	**67.7**	**69.6**	**28.1**	21.9	**17.4**	**59.6**	21.8	
	Test: 60 RealImg	**42.7**	45.5	74.0	72.3	**83.1**	63.6	**60.2**	49.3	**55.9**
		10.4	45.2	3.7	63.0	41.6	27.8	9.6	22.8	
SynImg-463Cls	Training: 100 SynImg	10.6	14.0	5.5	56.9	28.4	62.9	48.2	53.6	
+ RealImg	+ 10 RealImg	44.1	21.1	47.1	10.6	6.2	5.2	52.9	15.0	
(Fusion)	Test: 60 RealImg	37.7	36.8	4.5	59.4	25.1	67.4	27.7	22.4	30.9
		68.1	**79.1**	84.5	72.3	**86.4**	**68.0**	**78.0**	**73.3**	
RealImg	Training: 40 RealImg	90.9	**77.4**	**90.9**	88.6	**71.1**	**91.0**	**98.3**	**86.2**	
		98.0	**90.7**	**81.3**	**67.0**	**54.5**	**64.0**	90.9	**59.6**	
	Test: 30 RealImg	**81.0**	57.3	**97.9**	**99.5**	**86.7**	**90.4**	**87.5**	**85.8**	**81.1**
		61.6	67.2	**84.9**	**72.5**	70.0	49.6	71.9	33.0	
Deep Logo [18]	Training: 40 RealImg	**92.9**	53.5	80.1	**88.8**	61.3	90.0	84.2	79.7	
		85.2	89.4	57.8	-	34.6	50.3	**98.6**	34.2	
	Test: 30 RealImg	63.0	**57.4**	94.2	95.9	82.2	87.4	84.3	81.5	74.4
		-	-	-	-	-	-	-	-	
BD-FRCN-M [27]	Training: 40 RealImg	-	-	-	-	-	-	-	-	
		-	-	-	-	-	-	-	-	
	Test: 30 RealImg	-	-	-	-	-	-	-	-	73.5

RealImg), Faster R-CNN is already able to achieve fairly good detection results (50.4% in mAP). This suggests the great capability of deep models partially due to the good parameter initialisation on ImageNet and PASCAL VOC images, confirming the similar findings elsewhere [15, 3, 2]. *Secondly*, when using our synthetic training images alone (i.e., SynImg-32Cls and SynImg-463Cls), the model performance on test data is much inferior than using sparse realistic training data. The potential reasons is that, there may exist great discrepancy between realistic and synthetic training images, also known as the domain drift problem [37, 26], i.e., a trained model may degrade significantly in performance when deployed to a new domain with much disparity involved as compared to the training data. *Thirdly*, it is observed interestingly that synthetic training data from non-target logo classes may even hurt the model generalisation, when comparing the mAP result between SynImg-32Cls (27.6%) and SynImg-463Cls (20.5%). This may be due to the distracting effects introduced during detector optimisation on a large number of (431) non-target logos and thus making the resulting model less effective towards target logos in model deployment. *Fourthly*, it is also found that model pre-training on the full synthetic images turns out to be superior than on those of target logo classes alone. This suggests that more generic model pre-training may produce better initialisation for further incremental model adaptation on sparse real training data.

In addition to comparative evaluations under the same setting, we also compare the reported results from the state-of-the-art methods [18, 27]. It can be seen (last three rows in Table 3) that the best alternative DeepLogo surpasses our method (SynImg-463Cls + RealImg): 74.4% vs. 55.9% in mAP. However, it should also be pointed out that our model was trained on much less (25%) real training images. To provide a more comparable evaluation, we exploited the Faster R-CNN model as an alternative to DeepLogo, since

Table 4. Evaluating different methods on our TopLogo-10 logo dataset.

Method	Adidas	Chanel	Gucci	HH	Lacoste	MK	Nike	Prada	Puma	Supreme	mAP
RealImg	28.6	32.9	32.8	33.9	47.1	40.4	0.5	15.0	9.5	44.4	28.5
SynImg-10Cls	7.1	9.2	3.0	0.0	10.9	13.5	0.1	0.2	9.1	20.1	7.3
SynImg-463Cls	14.1	4.7	9.1	0.4	18.3	22.9	3.0	0.2	4.0	25.1	10.2
SynImg-10Cls + RealImg	51.9	**44.8**	41.1	**38.1**	**53.3**	52.5	11.8	28.9	18.4	63.6	40.4
SynImg-463Cls + RealImg	**52.7**	39.9	**49.7**	36.5	48.4	**62.7**	**14.8**	**29.8**	**18.6**	**64.6**	**41.8**
SynImg-463Cls + RealImg (Fusion)	25.4	11.8	4.7	0.6	17.9	44.2	1.6	0.9	15.5	40.3	16.3

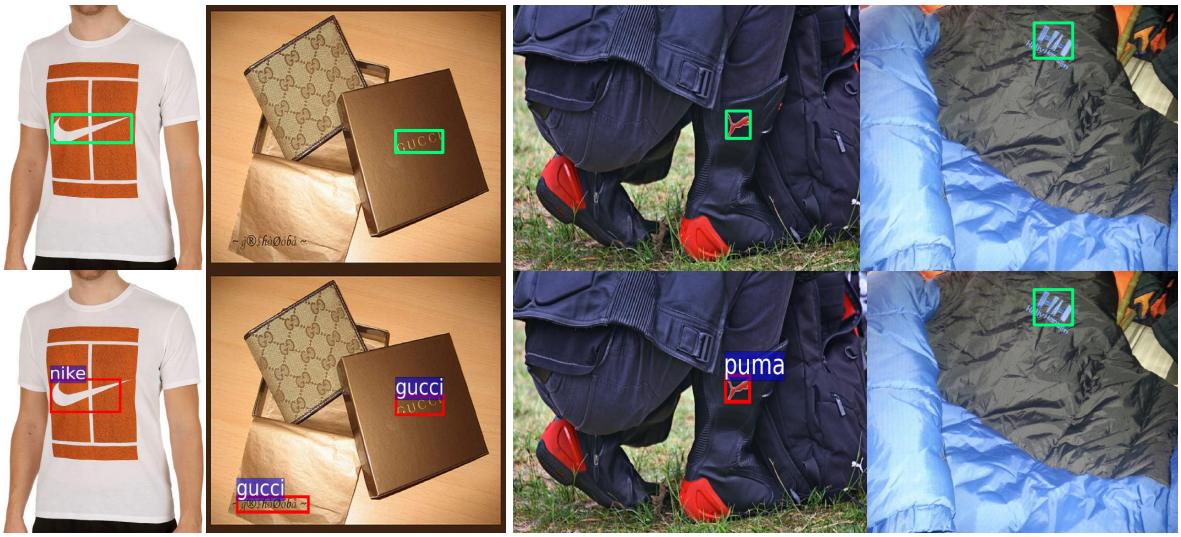

Figure 8. Qualitative evaluations on our TopLogo-10 logo dataset. **First row**: detection results by "RealImg"; **Second row**: detection results by "SynImg-463Cls + RealImg". Logo detections are indicated with red boxes. Ground truth is indicated by green boxes.

DeepLogo code is not released publicly, and trained a new model using same training data split as in DeepLogo. Our results on the FlickrLogo-32 dataset under the DeepLogo $40/30$ image split by our Faster RCNN based logo detector is 81.1% (the third last row in Table 3), significantly better than DeepLogos 74.4%. Moreover, under our $10/60$ split, a Faster R-CNN based logo detector (RealImg in Table 3) yields 50.4% in mAP vs. 55.9% by our method SynImg-463Cls + RealImg. Taken both into account, it suggests that the proposed method is more likely to outperform the DeepLogo model if compared under the same $10/60$ split setting. Given that this work focuses on exploring synthetic training data for deep learning logo detection regardless the detection model, any other detector [25] can be readily integrated into our framework.

We further evaluated the effect of the curriculum learning strategy in detection performance. To that end, we can compare SynImg-463Cls + RealImg and SynImg-463Cls + RealImg (Fusion) in Table 3. It is evident that by blindly learning a Faster R-CNN logo detector on the fusion (combined) of real and synthetic training images, the model generalisation is significantly degraded, likely due to severe bias towards the synthetic data. This is because that the synthetic images dominate in numbers in combined training data, resulting in that the trained model suffers from the domain shift problem. This result validates the efficacy of the proposed curriculum learning method in learning a more robust logo detection model when given heterogeneous training data sources.

Lastly, we carried out qualitative evaluations by examining the effect of synthetic context logo images on the detector performance. Figure 7 shows that context variations/diversity is effective for improving deep model learning resulting in more discriminative logo features therefore less missing logo detections – less false negatives (see 1^{st} and 2^{nd} columns). It is also evident that both models have difficulties in detecting small logos (3^{rd} column) and logos under extreme lighting conditions (4^{th} column). Also, the model trained by a single stage with full synthetic and real training data is likely to generate more false positive detections (3^{rd} column), likely due to over-diversity (aka noise) introduced by the synthetic context.

Evaluation on TopLogo-10. We evaluated Faster R-CNN detectors trained with different methods on our TopLogo-10 dataset. We present the detection results in Table 4.

By exploiting our synthetic context logo training images, the logo detection performance can be improved more significantly than on FlickrLogo-32, e.g., from 28.5% (by RealImg) to 41.8 (by SynImg-463Cls + RealImg) with

Table 5. Evaluating the effect of different synthetic context. Dataset: TopLogo-10.

Logo Name	Adidas	Chanel	Gucci	HH	Lacoste	MK	Nike	Prada	Puma	Supreme	mAP
Scene Context	**7.1**	**9.2**	**3.0**	0.0	**10.9**	**13.5**	**0.1**	**0.2**	**9.1**	**20.1**	**7.3**
Clean Context	2.8	2.0	0.0	**4.5**	0.5	6.0	**0.1**	0.0	**9.1**	10.2	3.5
Scene Context + RealImg	**51.9**	**44.8**	**41.1**	**38.1**	**53.3**	52.5	**11.8**	28.9	**18.4**	63.6	**40.4**
Clean Context + RealImg	48.7	37.3	38.1	33.9	46.4	**61.7**	2.5	**31.0**	12.4	**67.4**	37.9
No Context + RealImg	28.6	32.9	32.8	33.9	47.1	40.4	0.5	15.0	9.5	44.4	28.5

Table 6. Evaluating the effect of individual logo transformations. Dataset: TopLogo-10.

Logo Name	Adidas	Chanel	Gucci	HH	Lacoste	MK	Nike	Prada	Puma	Supreme	mAP
No Colouring	**15.4**	4.6	7.6	0.0	9.1	0.5	**9.1**	4.5	**9.1**	9.1	6.9
+ RealImg	52.9	40.5	**46.4**	27.7	53.2	48.0	**12.6**	25.9	17.4	**66.6**	39.1
No Rotation	13.6	**9.2**	**11.5**	**3.0**	13.6	10.8	0.3	**9.1**	**9.1**	**23.6**	**10.4**
+ RealImg	56.7	37.5	44.7	30.7	**55.1**	55.4	4.2	**35.3**	**24.5**	61.4	**40.5**
No Scaling	4.6	4.6	3.0	2.3	9.1	1.2	0.0	0.0	**9.1**	12.0	4.6
+ RealImg	55.7	35.3	46.3	31.2	54.7	49.0	10.6	30.6	13.2	61.6	38.8
No Shearing	3.9	1.7	9.1	0.0	9.1	3.5	0.0	0.0	**9.1**	12.1	4.9
+ RealImg	**58.8**	34.7	44.2	35.3	47.9	**55.7**	6.0	24.0	16.2	60.9	38.4
Full	7.1	**9.2**	3.0	0.0	**10.9**	**13.5**	0.1	0.2	**9.1**	20.1	7.3
+ RealImg	51.9	**44.8**	41.1	**38.1**	53.3	52.5	11.8	28.9	18.4	63.6	40.4

$13.3\%/46.7\%$ absolute/relative boost in mAP, as compared to $5.5\%/10.9\%$ on FlickrLogo-32. This further suggests the effectiveness of our synthetic context driven training data expansion method for logo detection, particularly for the practical and more challenging clothing brand logos. Mostly, we observed similar phenomenons as those on FlickrLogo-32 but one difference that SynImg-463Cls outperforms considerably SynImg-10Cls. The possible reason is that, the generic semantic context learned from a large number of logo classes becomes more indicative and useful since the background of clothing logos tends to be more clutter than those from FlickrLogo-32. Similarly, we show some qualitative detection examples in Figure 8.

Further Analysis. To give more insight, we performed further experiments with SynImg-10Cls on the TopLogo-10 dataset.

The Effect of Different Synthetic Context. We specifically evaluated the impact of context on model learning. To that end, we introduce a new type of context - clean black context ("Clean Context") as an alternative to the natural scene context ("Scene Context") used early in our SCL model. Table 5 shows when only synthetic training images were used in model training, the "Scene Context" is much more superior than the "Clean Context" for model training. This advantage remains when real training images were exploited for model adaptation. Interestingly, we also found that the "Clean Context" is able to improve logo detection performance, as revealed by comparison to the results without using any synthetic context. This further suggests that synthetic context based data expansion is an effective strategy for addressing the training data scarcity challenge.

The Effect of Individual Logo Transform. We evaluated the impact of different logo image transforms. To that end, we eliminated selectively each specific transform in synthesising the training images and then evaluated any change in model performance. Table 6 shows that: (1) With synthetic training images alone, all geometric and colour transforms except rotation bring some benefits, especially shearing and scaling. The little difference made by rotation makes sense considering that logo objects appear mostly without any rotation in real scenes. (2) The benefit of each transform remains after the detector is fine-tuned on real training data.

5. Conclusion

In this work, we described a new Synthetic Context Logo (SCL) training image generation algorithm capable of improving model generalisation capability in deep learning a logo detector when only sparse manually labelled data is available. We demonstrated the effectiveness and superiority of the proposed SCL on performing logo detection on unconstrained images, e.g., boosting relatively the detection accuracy of state-of-the-art Faster R-CNN network model by $>10\%$ on FlickrLogo-32 and $>40\%$ on the more challenging TopLogo-10 benchmark datasets. Importantly, this performance boost is obtained without the need for additional manual annotation. It shows the effectiveness of expanding training data through synthesising pseudo data especially *with rich logo context variations*. As such, deep detection model can be learned more reliably with better robustness against complex background clutters during model deployment. We carried out detailed evaluation and analysis on different strategies for model training. We further introduced a new logo dataset TopLogo-10, consisting of top 10 most popular clothing/wearable logos in challenging visual scene context, designed for more realistic testing of logo detections in real-world applications.

References

[1] D. Aroche. What the 2015 BrandZ™ Top 100 Means for Luxury. http://luxurysociety.com/articles/2015/06/what-the-2015-brandztm-top-100-means-for-luxury, 2015.

[2] Y. Bengio, F. Bastien, A. Bergeron, N. Boulanger-Lewandowski, T. M. Breuel, Y. Chherawala, M. Cisse, M. Côté, D. Erhan, J. Eustache, et al. Deep learners benefit more from out-of-distribution examples. In *International Conference on Artificial Intelligence and Statistics*, pages 164–172, 2011.

[3] Y. Bengio et al. Deep learning of representations for unsupervised and transfer learning. *International Conference on Machine learning*, 27:17–36, 2012.

[4] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *International Conference on Machine learning*, pages 41–48, 2009.

[5] R. Boia, A. Bandrabur, and C. Florea. Local description using multi-scale complete rank transform for improved logo recognition. In *IEEE International Conference on Communications*, pages 1–4, 2014.

[6] X. Chen and A. Gupta. Webly supervised learning of convolutional networks. In *IEEE International Conference on Computer Vision*, pages 1431–1439, 2015.

[7] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1409–1416, 2013.

[8] I. Design. Top 10 Fashion Logos. http://inkbotdesign.com/top-10-fashion-logos/, 2015.

[9] A. Dosovitskiy, P. Fischery, E. Ilg, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, T. Brox, et al. Flownet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision*, pages 2758–2766, 2015.

[10] C. Eggert, A. Winschel, and R. Lienhart. On the benefit of synthetic data for company logo detection. In *ACM Conference on Multimedia Conference*, pages 1283–1286, 2015.

[11] EmStacks. Top 10 Best Clothing Brands. http://www.thetoptens.com/best-clothing-brands/.

[12] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[13] R. Girshick. Fast r-cnn. In *IEEE International Conference on Computer Vision*, 2015.

[14] A. Gupta, A. Vedaldi, and A. Zisserman. Synthetic data for text localisation in natural images. *arXiv e-prints*, 2016.

[15] J. Hoffman, E. Tzeng, J. Donahue, Y. Jia, K. Saenko, and T. Darrell. One-shot adaptation of supervised deep convolutional models. *arXiv e-prints*, 2013.

[16] S. C. Hoi, X. Wu, H. Liu, Y. Wu, H. Wang, H. Xue, and Q. Wu. Logo-net: Large-scale deep logo detection and brand recognition with deep region-based convolutional networks. *arXiv preprint arXiv:1511.02462*, 2015.

[17] B. HUNDREDS. The 50 Greatest Streetwear Brands. http://uk.complex.com/style/2011/06/the-50-greatest-streetwear-brands/, 2011.

[18] F. N. Iandola, A. Shen, P. Gao, and K. Keutzer. Deeplogo: Hitting logo recognition with the deep neural network hammer. *arXiv preprint arXiv:1510.02131*, 2015.

[19] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1):1–20, 2016.

[20] A. Joly and O. Buisson. Logo retrieval with a contrario visual query expansion. In *ACM International Conference on Multimedia*, pages 581–584, 2009.

[21] Y. Kalantidis, L. G. Pueyo, M. Trevisiol, R. van Zwol, and Y. Avrithis. Scalable triangulation-based logo recognition. In *ACM International Conference on Multimedia Retrieval*, page 20, 2011.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[23] K.-W. Li, S.-Y. Chen, S. Su, D.-J. Duh, H. Zhang, and S. Li. Logo detection with extendibility and discrimination. *Multimedia tools and applications*, 72(2):1285–1310, 2014.

[24] X. Liang, S. Liu, Y. Wei, L. Liu, L. Lin, and S. Yan. Towards computational baby learning: A weakly-supervised approach for object detection. In *IEEE International Conference on Computer Vision*, December 2015.

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, 2016.

[26] H. V. Nguyen, H. T. Ho, V. M. Patel, and R. Chellappa. Joint hierarchical domain adaptation and feature learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):5479–5491, 2015.

[27] G. Oliveira, X. Frazão, A. Pimentel, and B. Ribeiro. Automatic graphic logo detection via fast region-based convolutional networks. *arXiv preprint arXiv:1604.06083*, 2016.

[28] C. Pan, Z. Yan, X. Xu, M. Sun, J. Shao, and D. Wu. Vehicle logo recognition based on deep learning architecture in video surveillance for intelligent traffic system. In *IET International Conference on Smart and Sustainable City*, pages 123–126, 2013.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[30] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[31] J. Revaud, M. Douze, and C. Schmid. Correlation-based burstiness for logo retrieval. In *ACM International Conference on Multimedia*, pages 965–968, 2012.

[32] S. Romberg and R. Lienhart. Bundle min-hashing for logo recognition. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 113–120. ACM, 2013.

[33] 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*, page 25. ACM, 2011.

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, pages 1–42, 2014.

[35] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. *arXiv preprint arXiv:1604.03540*, 2016.

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[37] A. Torralba, A. Efros, et al. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011.

[38] I. Yildirim, T. D. Kulkarni, W. A. Freiwald, and J. B. Tenenbaum. Efficient and robust analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations. In *Annual Conference of the Cognitive Science Society*, 2015.