Contrastive Multi-View Textual-Visual Encoding: Towards One Hundred Thousand-Scale One-Shot Logo Identification*

Nakul Sharma
Indian Institute of Technology
Jodhpur, Rajasthan, India
sharma.86@iitj.ac.in

Abhirama S. Penamakuri
Indian Institute of Technology
Jodhpur, Rajasthan, India
penamakuri.1@iitj.ac.in

Anand Mishra
Indian Institute of Technology
Jodhpur, Rajasthan, India
mishra@iitj.ac.in

ABSTRACT

In this paper, we study the problem of identifying logos of business brands in natural scenes in an open-set one-shot setting. This problem setup is significantly more challenging than traditionally-studied 'closed-set' and 'large-scale training samples per category' logo recognition settings. We propose a novel multi-view textual-visual encoding framework that encodes text appearing in the logos as well as the graphical design of the logos to learn robust contrastive representations. These representations are jointly learned for multiple views of logos over a batch and thereby they generalize well to unseen logos. We evaluate our proposed framework for cropped logo verification, cropped logo identification, and end-to-end logo identification in natural scene tasks; and compare it against state-of-the-art methods. Further, the literature lacks a 'very-large-scale' collection of reference logo images that can facilitate the study of one-hundred thousand-scale logo identification. To fill this gap in the literature, we introduce Wikidata Reference Logo Dataset (WiRLD), containing logos for 100K business brands harvested from Wikidata. Our proposed framework that achieves an area under the ROC curve of 91.3% on the QMUL-OpenLogo dataset for the verification task, outperforms state-of-the-art methods by 9.1% and 2.6% on the one-shot logo identification task on the Toplogos-10 and the FlickrLogos32 datasets, respectively. Further, we show that our method is more stable compared to other baselines even when the number of candidate logos is on a 100K scale.

CCS CONCEPTS

• **Computing methodologies** → **Image representations**.

KEYWORDS

supervised contrastive learning, one-shot learning, open-set recognition, logo identification.

ACM Reference Format:
Nakul Sharma, Abhirama S. Penamakuri, and Anand Mishra. 2022. Contrastive Multi-View Textual-Visual Encoding: Towards One Hundred Thousand-Scale One-Shot Logo Identification. In *Proceedings of the Thirteenth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'22), December 8–10, 2022, Gandhinagar, India*, Soma Biswas, Shanmuganathan Raman, and Amit K Roy-Chowdhury (Eds.). ACM, New York, NY, USA, Article 25, 9 pages. https://doi.org/10.1145/3571600.3571625

*Produces the permission block, and copyright information

ICVGIP'22, December 8–10, 2022, Gandhinagar, India

ACM ISBN 978-1-4503-9822-0/22/12…$15.00
https://doi.org/10.1145/3571600.3571625

Figure 1: Our Goal: given a natural scene and a gallery of "one" reference logo each for K "unseen" business brands, our goal is to identify the correct logo. We present a novel contrastive multi-view textual-visual encoding to address this problem. Further, for the first time in the literature, we study the problem of logo identification in an extremely challenging scenario when the number of candidate logos, i.e. K is as large as 100K.

1 INTRODUCTION

We study the problem of logo recognition in a practical setting where "only one" reference logo each for K "unseen" business brands is available during inference, and the task is to detect the logo in a natural scene and identify it as one of the K potential logos. We refer to this problem as *Open-set One-shot Logo Identification in the Wild* and illustrate it in Figure 1. The success of this challenging task can lead to many downstream real-world applications, including comprehensive scene understanding, and image search.

Open-set One-shot Logo Identification in the Wild is a challenging task (especially when K is of one-hundred-thousand scale) and requires a model to learn robust and discriminative encoding of logos that can generalize well even to unseen business brands. Inspired by the seminal works in contrastive multi-view encoding [6, 9, 22, 36],

we present a supervised contrastive learning framework. Our framework encodes textual[1] as well as visual features associated with the graphical design of logos and learns a fused robust representation using our novel supervised contrastive loss formulation. Our framework requires a set of cropped logos during training. During inference, our model, by virtue of these *learned representations*, is able to compare unseen logos reasonably well even with an off-the-shelf method for detecting logos and naïve cosine similarity. Our framework differs from popular contrastive loss-based methods, e.g., pairwise [25] and triplet loss [13] as it jointly optimizes the loss in a batch and learns a discriminative representation.

Furthermore, there does not exist a dataset to study very large-scale logo identification in the literature. To fill this gap, we introduce Wikidata Reference Logo Dataset or WiRLD in short – a very-large-scale logo dataset containing reference logos for 100K business brands. We curate this dataset from an open-source knowledge base, namely Wikidata [40] and use this curated set as a reference dataset in our very-large-scale logo identification experiment. This collection can augment other datasets in the literature for performing large-scale logo identification experiments.

We perform rigorous experiments to evaluate our proposed model in three different settings: (i) cropped logo verification, (ii) cropped logo identification, (iii) end-to-end logo detection and identification, and evaluate the performance of various relevant methods including ours over four public datasets, namely QMUL-OpenLogo [35], FlickrLogos-47 [30], FlickrLogos-32 [21] and TopLogos [34]. Further, in order to perform truly very-large-scale logo identification, we use QMUL-OpenLogo dataset [35] as probe and our newly introduced dataset viz. WiRLD as a reference set. Our method achieves area under the ROC curve of 91.3% on the QMUL-OpenLogo dataset on cropped logo verification task. Further, our proposed framework outperforms state-of-the-art methods by 9.1% and 2.6% on the task of unseen cropped logo identification over TopLogos [34] and Flickr32 [21] datasets, respectively.

Contributions: To summarize, our contributions are three folds, (i) We present a contrastive multi-view encoding of visual-textual features by fusing textual, i.e., text associated with logos and visual, i.e., graphical design of logos and learn more robust and generalizable features. Our proposed contrastive multi-view encoding compels the samples from the same class and their augmented views closer and the samples from different classes and their augmented views farther in the semantic space. (ii) For the first time in the literature, we study the problem of logo identification in an extremely challenging scenario where the number of candidate logos is as large as 100K. In order to facilitate this study, we introduce a very-large-scale logo dataset, Wikipedia Reference Logo Dataset containing 100K reference logos. (iii) Our method achieves state-of-the-art results on the task of one-shot logo identification for unseen logos on four public logo datasets. Further, we also show the robustness of our approach for logo identification in a very-large-scale setting. We make our code and dataset available at our project website: https://vl2g.github.io/projects/logoIdent/.

[1]Often business brand names are part of logos, our method leverages this fact while learning representation.

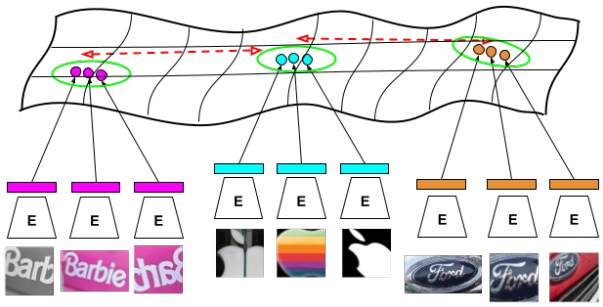


Figure 2: Our proposed contrastive multi-view textual-visual encoding (E) (refer Section 3 for more detail) projects logos in a subspace where multiple views of samples from the same and different business brands become closer and farther, respectively. We achieve this jointly for a batch using (2).

2 RELATED WORKS

2.1 Logo Recognition

The majority of the successful logo recognition approaches, including traditional [20, 21] as well as recent neural methods [2–5, 14, 17, 29] pose the problem as a closed-set recognition problem, where all business brands are seen during training, and a large number of logos per business brand are available. This is not a practical setting for real-world scenarios. Open-set logo recognition methods [8, 27] have been proposed to have a closer to a real-world setting but often relaxing one-shot assumptions. On one-shot logo recognition, recently [38] reported the performance of the Siamese network. We experimentally compare against this approach and outperform it by a large margin. Slightly advanced one-shot learning methods like Variational Prototypical Encoder (VPE) [24] leverage prototype images by learning a mapping from real-world images to prototype images; representations learnt via this mapping aid the one-shot performance of the model. However, the following work, VPE++ [43] has shown that embeddings learnt by VPE suffer from the hubness problem and hence extended the VPE framework by proposing a multi-task loss formulation that reduces hubness. VPE++ method treats contrastive loss and classification loss as isolated losses as part of the multi-task loss. Through this work, we provide a supervised contrastive learning framework that jointly leverages classification and contrastive objectives. Our results show that our framework learns more generalizable representations, which are key for open-set one-shot identification tasks.

2.2 Logo Datasets

Many logo datasets have been proposed for various tasks, including logo detection and classification. The majority of the existing datasets [14–16, 21, 28, 30, 35, 37] have very limited coverage of logo classes, making them unsuitable for large-scale logo identification settings. Few works [8, 27, 33, 41, 42] have proposed logo datasets with more logo classes. However, unfortunately, only some of them are publicly available. Such limitations restrict existing models

and benchmarks from exploring practical settings like very-large-scale logo identification tasks. To overcome such limitations and facilitate models to evaluate over the task of very-large-scale logo identification, we introduce a very-large-scale logo dataset, namely Wikipedia Reference Logo Dataset curated from open-source knowledge base Wikidata [40], containing 100K reference logos.

2.3 Contrastive Learning

Pairwise contrastive learning has been widely leveraged to learn generalizable features using Siamese networks [7, 11]. Triplet loss uses triplets instead of pairs [13], where each triplet consists of an anchor, positive and negative samples, and the goal is to make the anchor closer to the positive sample and farther to the negative sample. However, the performance of these methods depends on the quality of pairs or triplets [38]. Contrastive learning has been widely leveraged in the space of self-supervised representation learning approaches [18]. These methods rely on batch-wise losses [10, 32] and their variants, where they do not sample negatives in isolation; instead, they use other batch samples as negatives. Authors in [22] have extended contrastive learning to leverage class labels in loss formulation. In line with this research space, we present contrastive multi-view textual-visual encoding for robust and generalizable representation of logos.

3 PROPOSED APPROACH

3.1 Task Formulation

In this work, we address *open-set one-shot logo identification* in the following problem setup – during training, images of cropped logos from a set of business brands ($Brand_{train}$) are available. However, during inference, given a natural scene and a set of K business brands ($Brand_{test}$) with one reference logo for each brand, our goal is to localize and identify the logo in the scene. Here, it should be noted that $Brand_{train} \cap Brand_{test} = \phi$ in our setup. In other words, we aim to identify unseen business brands during the inference. Learning discriminative and robust encoding for logos is required to address this task. To this end, we propose a *contrastive multi-view textual-visual encoding* for addressing the problem.

3.2 Contrastive Multi-View Textual-Visual Encoding

3.2.1 Image representation. For a given batch of n logos $\mathcal{I} = \{I^1, I^2, \cdots, I^n\}$ (where each $I^i \in \mathbb{R}^{3\times H\times W}$) sampled from a dataset, we begin by obtaining two distorted views of each image using a set data augmentations $\mathcal{A}$ adopted from [44]. The augmented views thus obtained, $\mathcal{I}_a$ and $\mathcal{I}_b$ for each image in a batch are fed to the visual encoder f_θ and the textual encoder g simultaneously. It should be noted that logos are often composed of graphical design and text, and the encoders f_θ and g are designed to capture and encode these attributes of logos[2]. For encoding the visual features of the logo, any visual encoder can be used in our framework. We use ResNet50 [12] as our visual encoder to obtain 2048-dimensional features representing the graphical design of logos. These features $\mathbf{V}_a$ and $\mathbf{V}_b$, with $\mathbf{V}_{\{a,b\}} \in \mathbb{R}^{n\times 2048}$, are obtained from both the views of logo $\mathcal{I}_a$ and $\mathcal{I}_b$ respectively.

[2]If no text is detected in the logo, g outputs a zero vector.

Table 1: Notation used in the paper.

Symbol	Meaning
f_θ	Visual Encoder
g	Textual Encoder
h_ϕ	Projection MLP
$\mathcal{I}_{\{a,b\}}$	Augmented views of a batch
$\mathbf{V}_{\{a,b\}}$	Visual Features
$\mathbf{T}_{\{a,b\}}$	Textual Features
$\mathbf{Z}_{\{a,b\}}$	Projected final representation

3.2.2 Text representation. Any state-of-the-art scene text recognizer can be used to encode the textual features. We use the implementation from [1] based on the CRNN [31] model (referred to as OCR-net in our framework). OCR-net has a traditional convolutional neural network to encode the image, followed by an LSTM module to decode the OCR-text character by character. We use the last hidden-state representation of the LSTM module as textual embedding. We refer to this module as our textual encoder g. We obtain the 256-dimensional textual feature vectors $\mathbf{T}_a$ and $\mathbf{T}_b$ for both the views of logo $\mathcal{I}_a$ and $\mathcal{I}_b$, respectively. Note that the weights of our textual encoder are frozen.

3.2.3 Contrastive formulation and training objective. Visual features $\mathbf{V}_a$ and $\mathbf{V}_b$ are then concatenated with textual features $\mathbf{T}_a$ and $\mathbf{T}_b$ respectively before being projected to a 512-dimensional space using an MLP h_ϕ. The output embeddings are normalized to obtain final logo representations $\mathbf{Z}_a$ and $\mathbf{Z}_b$, respectively, with $\mathbf{Z}_{\{a,b\}} \in \mathbb{R}^{n\times 512}$, such that $||v||_2 = 1$ where v is any row vector in matrix $\mathbf{Z}_a$ and $\mathbf{Z}_b$. Parameters θ and ϕ are learnable. It should be noted here that each row of matrix $\mathbf{Z}_a$ and $\mathbf{Z}_b$ denote normalized feature vector corresponding to one image in a batch. An overview of our proposed framework is illustrated in Figure 3(a). (Notations used in our method are summarized in Table 1.

Once we obtain $\mathbf{Z}_a$ and $\mathbf{Z}_b$, we formulate our contrastive loss function based on the intuition that the embeddings of the logos of the same brands across $\mathbf{Z}_a$ and $\mathbf{Z}_b$ should lie closer in the embedding space, while the embeddings of the logos of different categories should lie farther apart. Our objective is illustrated in Figure 2. Formally, we define our loss function as follows:

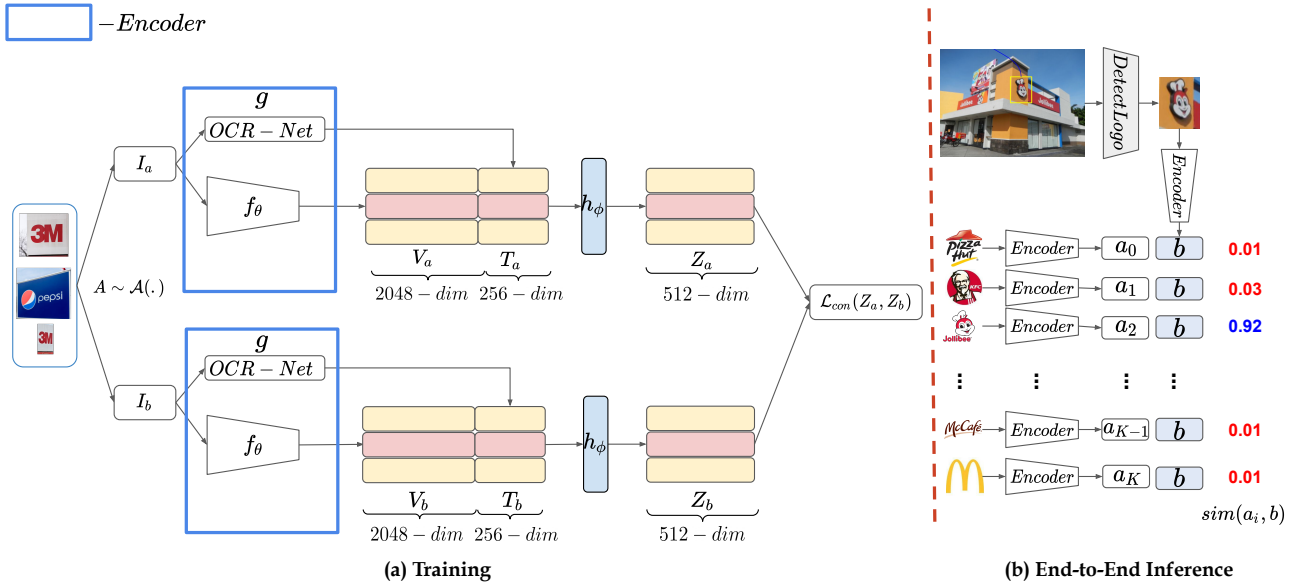
$$\mathcal{L}_{con}(\mathbf{Z}_a, \mathbf{Z}_b) = l(\mathbf{Z}_a, \mathbf{Z}_b) + l(\mathbf{Z}_b, \mathbf{Z}_a) + l(\mathbf{Z}_a, \mathbf{Z}_a) + l(\mathbf{Z}_b, \mathbf{Z}_b), \tag{1}$$

where

$$l(\mathbf{Z}_u, \mathbf{Z}_v) = -\sum_{i=1}^{i=n} \sum_{p\in P(i)} \log \frac{exp(\mathbf{z}_i^u . \mathbf{z}_p^v/\tau)}{\sum_{j=1}^{j=n} \sum_{j\notin P(i)} exp(\mathbf{z}_i^u . \mathbf{z}_j^v/\tau)}. \tag{2}$$

Here, i is an anchor in $\mathbf{Z}_u$, $P(i)$ is the set of all the positive logo indices corresponding to the anchor in the $\mathbf{Z}_v$ matrix. $\mathbf{z}_i^u$ is the i^{th} row in $\mathbf{Z}_\mathbf{u}$, similarly, $\mathbf{z}_i^v$ is the i^{th} row in $\mathbf{Z}_v$. Parameter τ is empirically chosen as 0.07 for all our experiments.

Unlike other recently proposed supervised contrastive loss [22], for a given anchor, our loss formulation does not try to maximize the similarity scores for all the positive pairs over "all the possible"



(a) Training (b) End-to-End Inference

Figure 3: (a) Overview of our proposed framework. We obtain two-views $\mathcal{I}_a$ and $\mathcal{I}_b$ of the input logo images for a batch I using a set of data augmentations [44]. For both-view batches I_a and I_b, we obtain: textual embedding (T_a, T_b) obtained from $g(\mathcal{I}_i)$, $\forall\, i \in \{a, b\}$, respectively; g is last hidden-state vector of LSTM module of off-the-shelf OCR-Net [1], and visual embeddings (V_a, V_b) obtained from $f_\theta(I_i)$, $\forall\, i \in \{a, b\}$, respectively; f_θ is visual backbone [12]. We concatenate ($\mathrm{V}_a : \mathrm{T}_a$), ($\mathrm{V}_b : \mathrm{T}_b$) and project using an MLP h_ϕ to obtain Z_a and Z_b respectively. The encoder is trained using the proposed supervised contrastive loss formulation. (b) illustrates the inference setup of our framework. Please refer to Section 3 for more details. [Best viewed in color].

pairs in the batch. Instead, we maximize the cosine similarity of all the positive pairs over all the negative pairs only. This ensures that multiple positive pairs do not compete against each other to achieve a higher similarity score, thereby resulting in robust representations for logos, which is desirable for our task. Further, our proposed method is not only trained to learn the alignment between positive pairs in a batch but also learn to align different views of positive pairs; and similarly learns to push the embeddings of negative pairs as well as different views of negative pairs, farther from the positive pairs in the representation embedding space.

3.3 Inference

For end-to-end inference, given a natural scene, we detect logos using YOLOv5s [19], which is independently fine-tuned on the training set of QMUL-OpenLogo for the task of class-agnostic logo detection. Detected candidate logo bounding boxes are encoded using our "trained" contrastive multi-view textual-visual encoder that concatenates 2048-dimensional visual embedding from f_θ with 256-dimensional textual embedding from g to obtain $\mathbf{b}$ to obtain a 2348-dimensional fused embedding. Reference logos for K business brands (one reference logo per brand) are encoded in a similar fashion to obtain their corresponding fused embeddings $\{\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_K\}$, with $\mathbf{a}_{\{1,\cdots,K\}} \in \mathbb{R}^{1\times 2348}$. We rank the 'K' reference logos based on the cosine similarity between $\mathbf{a}_i$ and $\mathbf{b}$ for $i = \{1, 2, \cdots, K\}$ and take the most similar (= higher cosine similarity) as the identified logo. An overview of our inference setting is illustrated in Figure 3(b).

3.4 Training and Implementation details

We use ResNet50 [12] initialized with ImageNet pre-trained weights and frozen off-the-shelf OCR-Net [1], and LSTM embeddings of the detected OCR-Text as our visual and textual backbones, respectively. We train our encoder with the proposed supervised contrastive loss framework using the SGD algorithm with a momentum of 0.9 and a learning rate of $1e-4$. We train all of our models on Nvidia GTX 1080 Ti GPU. During end-to-end inference, we utilize a class-agnostic YOLOv5s [19] detector fine-tuned on our training split of the QMUL-OpenLogo dataset [35] to detect logos from natural scene images. Additionally, we utilize a synthetic logo from each class in our formulation to have a better intra-class alignment during the experimental setting of [43]. We make implementation of this work available at our project website: https://vl2g.github.io/projects/logoIdent/.

4 EXPERIMENTS AND RESULTS

In this section, we first discuss existing datasets that we use as part of our experimental settings in Section 4.1 and then we present our curated dataset, namely Wikipedia Reference Logo Dataset in Section 4.1.5. We discuss baselines and ablations in Section 4.2 and Section 4.3, respectively. Further, we briefly explain various evaluation settings; and discuss the quantitative and qualitative results in Section 4.4 and Section 4.5, respectively.

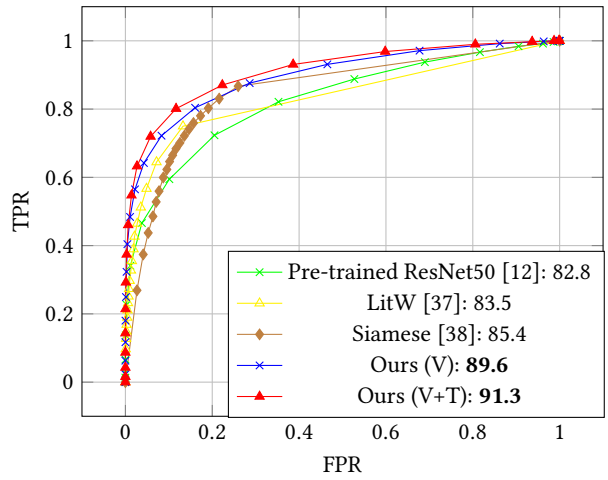


Figure 4: ROC curves for cropped logo verification task on the QMUL-OpenLogo dataset [35]. The legends show the area under the ROC metric corresponding to each method.

Table 2: Comparison of our newly introduced dataset, Wikipedia Reference Logo Dataset with the other related logo datasets. Our introduced dataset provides a very-large-scale reference set for one-shot logo identification. (*-not publicly available)

Dataset	#logo classes	#images
FlickrLogos-27 [21]	27	1K
FlickrLogos-32 [21]	32	8.2K
BelgaLogos [28]	37	10K
FlickrLogos-47 [21]	47	8.2K
LOGO-Net [15]	160	73.4K
TopLogo-10 [34]	10	0.7K
Logo-405 [16]	405	32.2K
Logos in the wild [37]	871	11K
QMUL-OpenLogos [35]	300	27K
WebLogo-2M [33]	194	1.8M
PL2K* [8]	2K	295K
Logo-2K+ [42]	2.3K	167K
LogoDet-3K [41]	3K	158K
PL8K* [27]	8K	3M
WiRLD (This work)	**100K**	**100K**

4.1 Datasets

4.1.1 ***QMUL-OpenLogo Dataset [35].*** This dataset has 27K curated images of 336 business brands. We follow the same split as authors of [38], where logos from 211 business brands are used for training and fine-tuning, and one logo each from 125 business brands is used for testing. Note that train and test classes are disjoint.

4.1.2 ***FlickrLogos-47 [30].*** It contains 2,235 annotated scenic images with logo regions spanning across 47 logo classes (32 symbolic logos and 15 textual logos). We randomly pick 30 business brands out of 47 for training and 17 unseen brands for testing purposes. We leverage the existing bounding box annotation for this dataset and thus obtain 1936 cropped logo images as part of the train set and 4032 as the test set.

Table 3: Comparison of cropped logo identification results on Flickr32 [21] and TopLogos-10 [34] datasets, respectively. We report Top-1 accuracy for both seen and unseen logo classes. Baseline results for methods QuadNet [23], MatchNet [39], VPE [24] and VPE++ [43] are taken directly from [43].

Method	Belga [28] → Flickr-32 [30]		Belga [28] → Toplogos [34]	
	All (Top-1)	Unseen (Top-1)	All (Top-1)	Unseen (Top-1)
VAE	27.17	27.31	23.30	18.59
Siamese Network [25]	24.7	22.82	30.84	30.46
Pretrained ResNet [12]	43.21	44.68	38.35	46.56
LitW [37]	33.96	26.34	57.21	51.10
QuadNet [23]	31.68	28.55	38.89	34.16
MatchNet [39]	38.54	35.28	28.46	27.46
VPE [24]	56.6	53.53	58.65	57.75
VPE++ [43]	65.54	62.56	65.57	70.27
SupCon [22]	65.84	64.84	66.06	70.22
Ours - Vision	**66.42**	**64.92**	**72.05**	**72.49**
Ours - Vision + Text	**66.77**	**65.17**	**72.26**	**79.33**

4.1.3 ***BelgaLogos [28].*** This dataset contains 10K logo images spanning over 26 logo classes. Following the setting in [24], we use this dataset to train our model with our proposed framework.

4.1.4 ***Toplogos [34].*** This dataset consists of 700 logo images over ten logo classes. Following the setting in [24], we use this dataset as a test dataset for our cropped logo verification task for a fair comparison with the baselines.

4.1.5 ***Wikipedia Reference Logo Dataset*** *(WiRLD), (newly introduced in our work).* Many datasets have been proposed in the research space of logo detection, and recognition [8, 14–16, 21, 27, 28, 30, 33, 35, 37, 41, 42]; however, unfortunately, the majority of these datasets have very limited coverage of logo classes or not publicly available; making them unsuitable for the tasks that demand a very-large-scale logo dataset, e.g. large-scale logo identification. (An overview comparing the various logo datasets is shown in Table 2). To overcome shortcomings of existing datasets and to facilitate models to explore the task of very-large-scale logo identification, we curate large-scale logos from an open-source knowledge base, namely Wikidata [40]. We follow a three-stage process to extract logos from Wikidata. In stage-1, we obtain all the entities over Wikidata with a logo with the help of the Wikidata SPARQL[3] query service. Once all entities are obtained, in stage-2, we parse the one-hop neighbourhood for each entity over the Wikidata graph and obtain logo URLs. Finally, in stage-3, we download original logo images from these URLs. We use this curated set of reference logo gallery for our task *viz.* large-scale open-set one-shot logo identification. Our curated dataset has 100K reference logo images spanning over 100K logo classes (One logo image for every entity). The URLs of logo images of WiRLD are available for download in our project website[4].

[3]https://query.wikidata.org/
[4]https://vl2g.github.io/projects/logoIdent/

Figure 5: A selection of logos from our newly introduced Wikipedia Reference Logo Dataset. In total, our dataset has around 100K logo classes, with each class having one reference logo. Note that these logos are noise-free and clean as they are sourced directly from Wikidata. Hence, it has great utility as a reference gallery set, especially for a task like very-large-scale one-shot logo identification.

Table 4: Comparison of cropped logo identification results on both QMUL-OpenLogo [35] and FlickrLogos-47 [30] datasets. We report Top-k (k = 1, 5 and 10) accuracy (in %).

	QMUL-OpenLogo [35]			FlickrLogos-47 [30]		
Method	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
Levenshtein Distance	30.8	34.1	34.1	17.6	17.6	29.4
Siamese Network [38]	23.3	49.2	61.7	41.2	**94.1**	94.1
Pretrained ResNet [12]	30	48.3	59.2	29.4	82.4	88.2
LitW [37]	27.5	54.2	68.3	17.6	76.5	**100**
SupCon [22]	44.2	62.5	70.8	76.5	88.2	**100**
Ours - Vision	48.3	63.3	70	76.5	**94.1**	94.1
Ours - Vision + Text	**55.7**	**68.3**	**73.3**	**82.4**	**94.1**	94.1

Table 5: Comparison of end-to-end logo identification results on both QMUL-OpenLogo [35] and FlickrLogos-47 [30] datasets. We report Top-k (k = 1, 5 and 10) accuracy (in %).

	QMUL-OpenLogo [35]			FlickrLogos-47 [30]		
Method	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
Levenshtein Distance	16.6	19.2	22.5	0	5.9	17.6
Siamese Network [38]	12.9	25.9	39.7	43.8	81.2	87.5
Pretrained ResNet [12]	16.4	28.4	39.7	43.8	**87.5**	**93.8**
LitW [37]	17.2	33.6	43.1	43.8	81.2	87.5
SupCon [22]	23.3	30.2	37.9	**62.5**	81.2	**93.8**
Ours - Vision	24.1	32.8	41.4	56.2	**87.5**	**93.8**
Ours - Vision + Text	**26.7**	**39.7**	**48.3**	56.2	81.2	**93.8**

4.2 Baselines

We choose various state-of-the-art methods as baselines that are closely related to our problem setup. We group baselines into two categories, namely (i) single-stream methods and (ii) contrastive-loss based approaches. Under single stream networks, we use a pretrained ResNet [12] model and a method mentioned in LitW [37]. Under contrastive-loss based approaches, we use the two approaches Siamese network-based approach [38] and the recently proposed supervised contrastive loss-based approach [22]. Additionally, we consider recent works, namely VPE++ [43], VPE [24], matching network [39], quadruplet networks [23] and variational autoencoder as our baselines. For fair comparison against these additional baselines, we follow a similar experimental setup as [43].

4.3 Ablations

We perform the following ablations, (i) our method's performance on seen classes: to benchmark and contrast the performance of our proposed framework over seen vs unseen logo classes, (ii) our method (without Text): to estimate the importance of textual pipeline, (iii) our method using different visual backbones: to estimate the role and importance of visual backbone. Further, to illustrate the performance of a method that only ranks the logos based on the recognized text and does not use visual cues, we also show results using Levenshtein distance between text detected from the logo and the reference logo crops.

4.4 Quantitative Results

We quantitatively evaluate our proposed framework in four experimental settings and compare it with various related approaches.

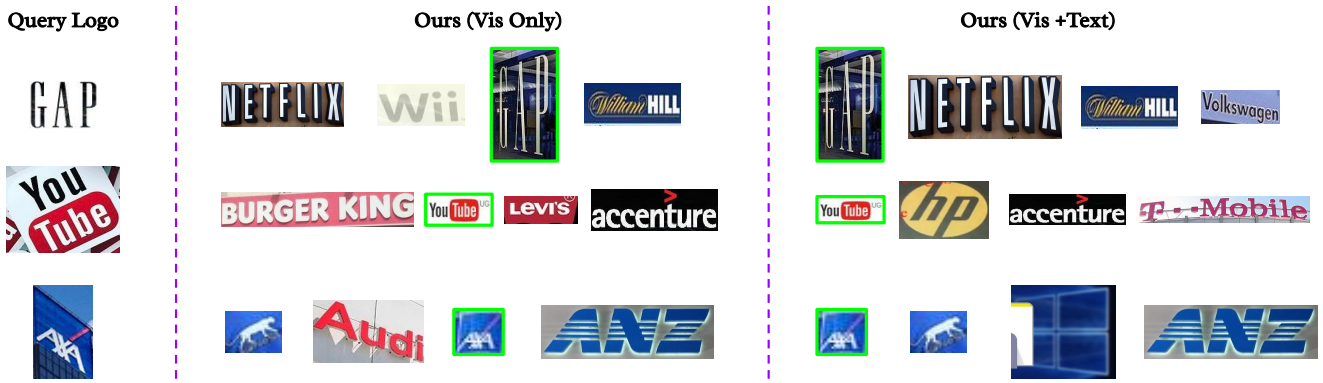


Figure 6: A selection of test logos detected from the natural scene as queries. Each row has a query (on the left), and top-4 most similar logos obtained using Ours (Vision only) and Ours (Vision+Text) models on the cropped logo identification on the QMUL-OpenLogo. Logos with a green bounding box represent the correct match. These results show that our framework is able to learn robust representations leveraging both textual and visual cues from logos. [Best viewed in color].

Note that the test set's classes (business brands) in all evaluation settings are unseen during training.

4.4.1 ***Cropped logo verification.*** In this setting, a pair of cropped logos (from 20,000 logo image pairs [35]) are compared against each other for a match. We present the ROC curve comparison of our framework with the baselines in Figure 4 on the QMUL-OpenLogo dataset. Our framework outperforms the previous state-of-the-art model by achieving an area under the ROC curve of 91.2% on the QMUL-OpenLogo dataset.

4.4.2 ***Cropped logo identification.*** In this task, we follow two settings: (i) Similar to [24, 43] where a noise-free clean logo is matched over a set of cropped logos from natural scene images. We follow the same training and evaluation protocols, and we train our proposed framework on Belgalogo [28] dataset and evaluate over Flickr32 [21] and TopLogos-10 [34] datasets, respectively, and baseline results are taken directly from [24, 43] for this setting; We present accuracy of seen vs unseen classes in Table 3. Our framework outperforms the baselines on both seen and unseen categories. We have not included these baselines in further evaluation settings due to different training paradigms. (ii) Challenging setting where a noisy cropped logo is compared against 'one' reference logo of K business brands (where K can be potentially large, and reference logos can be noisy as well). The reference logos are ranked based on similarity with the cropped logo. We compare Top-k (k = 1, 5 and 10) accuracy of our framework with the baselines in Table 4 on both QMUL-OpenLogo and FlickrLogos-47 datasets. On the QMUL-OpenLogo dataset, our vision-only encoder trained with the proposed loss framework has outperformed the baselines, indicating the robustness of the proposed loss formulation.

4.4.3 ***End-to-end logo detection and identification.*** This is the practical setting where we do not assume that cropped logos are provided to us. Instead, we first detect the logo and then compare it against reference logos. We compare Top-k (k = 1, 5 and 10) accuracy of our framework with the baselines in this setting as shown in Table 5 on both QMUL-OpenLogo and FlickrLogos-47 datasets. On FlickrLogs-47, our method Top-1 accuracy is slightly inferior to one of the recent approaches. However, our Top-5 and Top-10 accuracy on this dataset are comparable.

Table 6: Logo identification results with our method over vision backbones, on QMUL-OpenLogo dataset [35]. We report Top-k (k = 1, 5 and 10) accuracy (in %).

		QMUL-OpenLogo [35]		
Method	Vision backbone	Top-1	Top-5	Top-10
Ours - Vision	AlexNet [26]	33.3	54.2	64.2
Ours - Vision + Text	AlexNet [26]	35.0	52.5	66.7
Ours - Vision	ResNet [12]	48.3	63.3	70.0
Ours - Vision + Text	ResNet [12]	**55.8**	**68.3**	**73.3**

4.4.4 ***Cropped logo identification against large-scale reference logos.*** This setting enables us to evaluate the performance of our framework in real-world scenarios where a cropped logo is compared against a very large set of logo images with the scale ranging from 1K to 100K. We evaluate our proposed framework on the task of logo identification over the QMUL-OpenLogo dataset as a probe set along with our curated large-scale open-set one-shot WiRLD as a reference set. Similar to the previous evaluation setting, we present Top-1 accuracy of our framework with the baselines over various scales of images in the gallery in a line chart in Figure 8. In a large-scale logo identification setting, a performance drop is expected with an increase in scale. However, our results reported in Figure 8 suggest that the representations learnt by our framework remain robust when compared against the previous best-performing baseline SupCon [22]. Our vision-only method slightly outperforms the vision-text method at higher scales, owing to the training constraints of OCR-Net, e.g. indifference in the image sizes used during training of OCR-Net vs size of the cropped logo images, original model being trained on english text.

We present the results of Levenshtein distance-based approach along with a vision-only encoder in Figure 4, Table 4, Table 5. In

Figure 7: Logo identification from natural scene images. Each row has a natural scene query image (on the left), and top-4 most similar logos obtained using our proposed method over vision only and vision+text variants on the end-to-end logo identification setting on the QMUL-OpenLogo dataset [35]. Logos with a green bounding box represent the correct match.

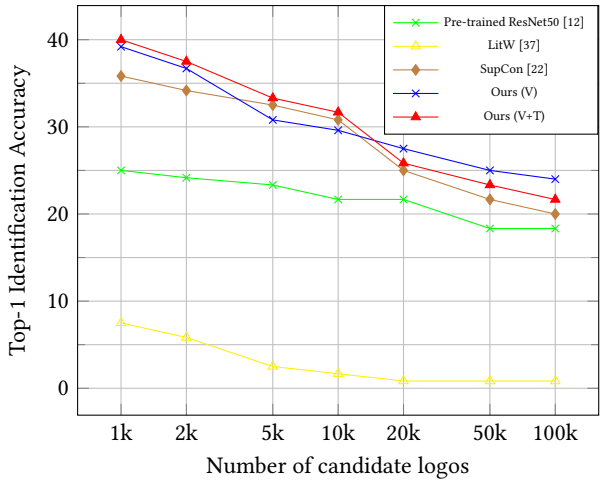


Figure 8: Large-scale logo identification. We present Top-1 accuracy of our framework with the baselines over varying scales. Performance drop is expected with an increase of scale; however, our framework retains its performance over baselines owing to the robustness of learnt representations.

Table 6, we present the comparison of Top-k (k = 1, 5 and 10) accuracy of our proposed encoder by varying visual encoders [12, 26] as backbones on the task of cropped logo identification on the QMUL-OpenLogo dataset. An encoder with our proposed fusion of both text and visual embeddings trained with the proposed loss formulation brings in the best from both modalities and induces better representative capabilities of the model, thereby resulting in noticeably superior performance over the baselines on unseen logo identification tasks at scale.

4.5 Qualitative Results

We perform an extensive qualitative analysis of our framework on both cropped logo identification as well as end-to-end logo identification from natural scene images. A selection of visual results on cropped logo identification is shown in Figure 6; similarly, a selection of visual results on end-to-end logo identification on natural scene images is shown in Figure 7.

4.6 Limitations and Future scope

We observe the following limitations of our work: (i) our proposed constrastive formulation of textual-visual features of logos is not tailored for time efficiency, (ii) we have used an off-the-shelf OCR-Net model to extract text from logos, which is trained and tested over English texts; hence, our model might suffer when logo images contain text from languages other than English, and (iii) the problem is far from solved when the scale is 100K in the task of large-scale open-set one-shot logo identification. We leave addressing these limitations as a future work.

5 CONCLUSION

Text within the logo has been underexplored for the task of *Open-set One-shot Logo Identification*. Towards this end, we have presented a framework that fuses textual as well as visual features associated with the graphical design of logos and learns robust representation using a novel formulation of supervised contrastive learning. Our proposed method outperformed previous state-of-the-art methods under one-shot constraints. We have also introduced a large-scale logo dataset, Wikipedia Reference Logo Dataset, which has a potentially huge scope in benchmarking and evaluating large-scale open-set one-shot logo identification techniques. Furthermore, our exhaustive experiments have demonstrated that the representations learned by our framework are fairly robust compared to competent baselines on the task of large-scale open-set one-shot logo identification. We made our data and implementation publicly available for enabling future research.

ACKNOWLEDGMENTS

Abhirama S. Penamakuri is supported by Prime Minister Research Fellowship (PMRF), Ministry of Education, Government of India.

REFERENCES

[1] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. 2019. What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis. In *ICCV*.

[2] Muhammet Bastan, Hao-Yu Wu, Tian Cao, Bhargava Kota, and Mehmet Tek. 2019. Large scale open-set deep logo detection. *arXiv preprint arXiv:1911.07440* (2019).

[3] Ayan Kumar Bhunia, Ankan Kumar Bhunia, Shuvozit Ghose, Abhirup Das, Partha Pratim Roy, and Umapada Pal. 2019. A deep one-shot network for query-based logo retrieval. *Pattern Recognition* 96 (2019), 106965.

[4] Simone Bianco, Marco Buzzelli, Davide Mazzini, and Raimondo Schettini. 2015. Logo recognition using cnn features. In *International Conference on Image Analysis and Processing*.

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

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*.

[7] S. Chopra, R. Hadsell, and Y. LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*.

[8] István Fehérvári and Srikar Appalaraju. 2019. Scalable logo recognition using proxies. In *WACV*.

[9] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *NeurIPS*.

[10] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 297–304.

[11] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

[13] Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*. Springer, 84–92.

[14] Steven CH Hoi, Xiongwei Wu, Hantang Liu, Yue Wu, Huiqiong Wang, Hui Xue, and Qiang Wu. 2015. Logo-net: Large-scale deep logo detection and brand recognition with deep region-based convolutional networks. *arXiv preprint arXiv:1511.02462* (2015).

[15] Steven CH Hoi, Xiongwei Wu, Hantang Liu, Yue Wu, Huiqiong Wang, Hui Xue, and Qiang Wu. 2015. Logo-net: Large-scale deep logo detection and brand recognition with deep region-based convolutional networks. *arXiv preprint arXiv:1511.02462* (2015).

[16] Sujuan Hou, Jianwei Lin, Shangbo Zhou, Maoling Qin, Weikuan Jia, and Yuanjie Zheng. 2017. Deep hierarchical representation from classifying logo-405. *Complexity* 2017 (2017).

[17] Forrest N Iandola, Anting Shen, Peter Gao, and Kurt Keutzer. 2015. Deeplogo: Hitting logo recognition with the deep neural network hammer. *arXiv preprint arXiv:1510.02131* (2015).

[18] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2021. A survey on contrastive self-supervised learning. *Technologies* 9, 1 (2021), 2.

[19] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Jiacong Fang, imyhxy, Kalen Michael, Lorna, Abhiram V, Diego Montes, Jebastin Nadar, Laughing, tkianai, yxNONG, Piotr Skalski, Zhiqiang Wang, Adam Hogan, Cristi Fati, Lorenzo Mammana, AlexWang1900, Deep Patel, Ding Yiwei, Felix You, Jan Hajek, Laurentiu Diaconu, and Mai Thanh Minh. 2022. *ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference*. https://doi.org/10.5281/zenodo.6222936

[20] Alexis Joly and Olivier Buisson. 2009. Logo retrieval with a contrario visual query expansion. In *ACM-MM*.

[21] Y. Kalantidis, LG. Pueyo, M. Trevisiol, R. van Zwol, and Y. Avrithis. 2011. Scalable Triangulation-based Logo Recognition. In *ICMR*.

[22] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *NeurIPS* (2020).

[23] Junsik Kim, Seokju Lee, Tae-Hyun Oh, and In So Kweon. 2018. Co-domain embedding using deep quadruplet networks for unseen traffic sign recognition. In *AAAI*.

[24] Junsik Kim, Tae-Hyun Oh, Seokju Lee, Fei Pan, and In So Kweon. 2019. Variational prototyping-encoder: One-shot learning with prototypical images. In *CVPR*.

[25] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *NeurIPS* (2012).

[27] Chenge Li, István Fehérvári, Xiaonan Zhao, Ives Macedo, and Srikar Appalaraju. 2022. SeeTek: Very Large-Scale Open-set Logo Recognition with Text-Aware Metric Learning. In *WACV*.

[28] Jan Neumann, Hanan Samet, and Aya Soffer. 2002. Integration of local and global shape analysis for logo classification. *Pattern recognition letters* 23, 12 (2002), 1449–1457.

[29] Stefan Romberg and Rainer Lienhart. 2013. Bundle min-hashing for logo recognition. In *ICMR*.

[30] Stefan Romberg, Lluis Garcia Pueyo, Rainer Lienhart, and Roelof Van Zwol. 2011. Scalable logo recognition in real-world images. In *ICMR*.

[31] Baoguang Shi, Xiang Bai, and Cong Yao. 2017. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. *IEEE TPAMI* 39 (2017), 2298–2304.

[32] Kihyuk Sohn. 2016. Improved Deep Metric Learning with Multi-class N-pair Loss Objective. In *NeurIPS*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.).

[33] Hang Su, Shaogang Gong, and Xiatian Zhu. 2017. Weblogo-2m: Scalable logo detection by deep learning from the web. In *CVPRW*.

[34] Hang Su, Xiatian Zhu, and Shaogang Gong. 2017. Deep learning logo detection with data expansion by synthesising context. In *WACV*.

[35] Hang Su, Xiatian Zhu, and Shaogang Gong. 2018. Open Logo Detection Challenge. In *BMVC*.

[36] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive Multiview Coding. In *ECCV*.

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

[38] Camilo Vargas, Qianni Zhang, and Ebroul Izquierdo. 2020. One shot logo recognition based on siamese neural networks. In *ICMR*.

[39] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *NeurIPS* (2016).

[40] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[41] Jing Wang, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, and Shuqiang Jiang. 2022. LogoDet-3K: A Large-Scale Image Dataset for Logo Detection. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 1 (2022), 1–19.

[42] Jing Wang, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, Haishuai Wang, and Shuqiang Jiang. 2020. Logo-2K+: A large-scale logo dataset for scalable logo classification. In *AAAI*.

[43] Chenxi Xiao, Naveen Madapana, and Juan Wachs. 2021. One-Shot Image Recognition Using Prototypical Encoders with Reduced Hubness. In *WACV*.

[44] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. 2021. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In *ICML*.