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

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 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 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 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/logoldent/.

---

*Often business brand names are part of logos, our method leverages this fact while learning representation.*
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 ∩ Brand_test = φ 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 $I = \{I^1, I^2, \ldots, I^n\}$ (where each $I^l \in \mathbb{R}^{3\times H \times W}$) 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, $I_a$ and $I_b$ for each image in a batch are fed to the visual encoder $f_\theta$ and the textual encoder $g$ simultaneously. It should be noted that logos are often composed of graphical design and text, and the encoders $f_\theta$ 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 $V_a$ and $V_b$, with $V_{(a,b)} \in \mathbb{R}^{n \times 2048}$, are obtained from both the views of logo $I_a$ and $I_b$, respectively.

\(^2\)If no text is detected in the logo, $g$ outputs a zero vector.

| Symbol | Meaning |
|--------|---------|
| $f_\theta$ | Visual Encoder |
| $g$ | Textual Encoder |
| $h_\phi$ | Projection MLP |
| $I_{(a,b)}$ | Augmented views of a batch |
| $V_{(a,b)}$ | Visual Features |
| $T_{(a,b)}$ | Textual Features |
| $Z_{(a,b)}$ | 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 $T_a$ and $T_b$ for both the views of logo $I_a$ and $I_b$, respectively. Note that the weights of our textual encoder are frozen.

3.2.3 Contrastive formulation and training objective. Visual features $V_a$ and $V_b$ are then concatenated with textual features $T_a$ and $T_b$ respectively before being projected to a 512-dimensional space using an MLP $h_\phi$. The output embeddings are normalized to obtain final logo representations $Z_a$ and $Z_b$, respectively, with $Z_{(a,b)} \in \mathbb{R}^{512}$, such that $\|v\|_2 = 1$ where $v$ is any row vector in matrix $Z_a$ and $Z_b$. Parameters $\theta$ and $\phi$ are learnable. It should be noted here that each row of matrix $Z_a$ and $Z_b$ 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 $Z_a$ and $Z_b$, we formulate our contrastive loss function based on the intuition that the embeddings of the logos of the same brands across $Z_a$ and $Z_b$ 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:

$$\mathcal{L}_{con}(Z_a, Z_b) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in P(i)} \log \frac{\exp(s_{i,j}^a \cdot s_{i,j}^b)}{\sum_{j=1}^{n} \exp(s_{i,j}^a \cdot s_{i,j}^b)}$$

(2)

Here, $i$ is an anchor in $Z_a$, $P(i)$ is the set of all the positive logo indices corresponding to the anchor in the $Z_a$ matrix. $s_{i,j}^a$ is the $i^{th}$ row in $Z_a$, similarly, $s_{i,j}^b$ is the $i^{th}$ row in $Z_b$. Parameter $r$ is empirically chosen as 0.07 for all our experiments.

Unlike the previously proposed 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”
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 \( f_\theta \) with 256-dimensional textual embedding from \( g \) to obtain 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 \( \{a_1, a_2, \ldots, a_K\} \), with \( a_{1:K} \in \mathbb{R}^{1 \times 2348} \). We rank the \( K \) reference logos based on the cosine similarity between \( a_i \) and b for \( i = \{1, 2, \ldots, K\} \) 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 \( 1 \times 10^{-4} \). 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.
Table 2: Comparison of our newly introduced dataset, logos and 15 textual logos). We randomly pick 30 business brands ages with logo regions spanning across 47 logo classes (32 symbolic FlickrLogos-47 [30]).

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.

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 to train our model with our proposed framework.

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 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 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 spans over 100K logo classes (One logo image for every entity). The URLs of logo images of WiRLD are available for download in our project website4.

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 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].

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.

With the aim of providing a very-large-scale reference set for one-shot logo identification. (An

![ROC curves for cropped logo verification task on the QMUL-OpenLogo dataset [35].](image)

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   | 27            | 1K      |
| FlickrLogos-32   | 32            | 8.2K    |
| BelgaLogos       | 37            | 10K     |
| FlickrLogos-47   | 47            | 8.2K    |
| LOGO-Net         | 160           | 73.4K   |
| TopLogo-10       | 10            | 0.7K    |
| Logo-405         | 405           | 32.2K   |
| Logos in the wild | 871          | 11K     |
| QMUL-OpenLogos   | 300           | 27K     |
| WebLogo-2M       | 194           | 1.8M    |
| PL2K∗ [8]        | 2K            | 295K    |
| Logo-2K+ [42]    | 2.3K          | 167K    |
| LogoDet-3K       | 3K            | 158K    |
| PL8K∗ [27]       | 8K            | 3M      |
| WiRLD (This work)| 100K          | 100K    |

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 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 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 website4.

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 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 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 website4.

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 | 24.7                       | 22.82                       | 30.84 | 30.46 |
| Pretrained ResNet | 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]    | 63.84                      | 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 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
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 %).

| Method                  | QMUL-OpenLogo [35] | FlickrLogos-47 [30] |
|-------------------------|--------------------|---------------------|
|                         | 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 %).

| Method                  | QMUL-OpenLogo [35] | FlickrLogos-47 [30] |
|-------------------------|--------------------|---------------------|
|                         | 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-

Table 6: Logo identification results with our method over vision backbones, on QMUL-OpenLogo dataset [35]. We report Top-

| Method            | Vision backbone | QMUL-OpenLogo [35] |
|-------------------|-----------------|--------------------|
|                   |                 | 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   |

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.

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 datasets.
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.

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 contrastive 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.
