— Supplementary Materials —
Variational Prototyping-Encoder: One-Shot Learning with Prototypical Images

Junsik Kim Tae-Hyun Oh† Seokju Lee Fei Pan In So Kweon
Dept. of Electrical Engineering, KAIST, Daejeon, Korea
†MIT CSAIL, Cambridge, US
mibastro@gmail.com, †taehyun@csail.mit.edu
seokju91@gmail.com, {feipan, iskweon77}@kaist.ac.kr

1. Contents

Here, we present additional details pertaining to the datasets and experiments that could not be included in the main text due to space constraints. All figures and references in this supplementary file are self-contained.

The contents included in these supplementary materials are as follows: 1) The network architecture, 2) Detail descriptions of the datasets used, 3) Embedding space visualization, and 4) Qualitative results of image retrieval.

1.1. Architecture

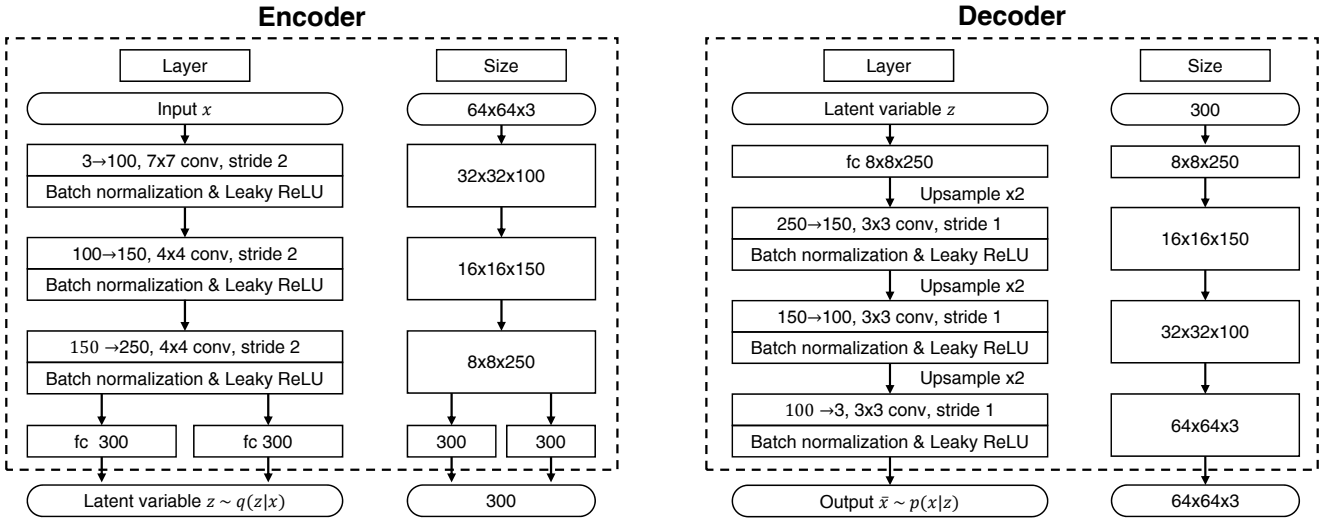

Figure 1. Architectures specifications of encoder and decoder blocks of the proposed variational autoencoder.

The detailed VPE network architecture is shown in Fig. 1.

1.2. Datasets

Dataset	GTSRB	TT100k	BelgaLogos	FlickrLogos-32	TopLogo-10
Instances	51,839	11,988	9,585	3,404	848
Classes	43	36	37	32	11

Table 1. Symbol dataset specifications

In this section, we present the details of each dataset used for the experiments in the main text. Table 1 is provided to summarize the statistics of the datasets.

GTSRB GTSRB [7] is the largest dataset for traffic-sign recognition. It contains 43 classes categorized into three larger categories: prohibitory, danger and mandatory. The dataset contains illumination variations, blur, partial shadings and low-resolution images as well as imbalanced sample distribution. The training set contains 39,209 images and the test set contains 12,630 images.

TT100K Tsinghua-Tencent 100K (TT100K) [11] is a Chinese traffic sign detection dataset that includes more than 200 classes. We cropped traffic sign instances from scenes to build a classification dataset. We filtered out instances with side lengths of less than 20 pixels because they are either not recognizable or miss annotated. Among more the defined classes, the 36 classes are selected for the evaluation that have available corresponding prototypes and a sufficient number of samples. For more details about the TT100K dataset, please refer to the work of Kim *et al.*[2].

FlickrLogos-32 Dataset FlickrLogos-32 [6] is a collection of images from Flickr containing 32 different logos. Most of the images contain a few and relatively clean, recognizable logo instances located near the center of an image compared to other datasets [1, 9]. The dataset is published to evaluate logo detection and recognition systems with 32 logo classes defined. The dataset has a total of 2,240 logo images, and it is partitioned into 10 training images, 30 validation images and 30 test images per class. It also contains 6,000 no-logo images to evaluate the false alarm rates of recognition systems. We cropped logo instances using bounding box annotations to evaluate our classification systems. In total, 3,372 logo instances were gathered by cropping.

BelgaLogos Dataset BelgaLogos [1, 4] is composed of 10,000 images from various aspects of everyday life with 37 logo classes annotated in a bounding box format. Unlike FlickrLogos-32, logos appear at diverse locations with large-scale variations, blur, saturation and occlusions. The quality levels of the samples are rated as either 'OK' or 'Junk' depending how clearly a sample is recognizable by human annotators. We cropped both 'OK' and 'Junk' logo instances to build a logo classification dataset. In total, 9,475 instances were collected. While FlickrLogos-32 shows an equal sample distribution per class, BelgaLogos shows a severe class imbalance from a small-sized class (2 samples) to a large-sized class (2,242 samples).

TopLogo-10 Dataset TopLogo-10 [8] contains 10 logo classes related to popular cloth, shoes and accessory brands. The images are collected from product images that are relatively clean and recognizable. Each class contains 70 images. We cropped logo instances using bounding box annotations and gathered a total of 853 logo samples. For the experiment, we defined a total of 11 logo classes by separating the 'Adidas' class into the 'Adidas-logo' and the 'Adidas-text' classes.

1.3. Embedding space

We provide t-SNE [5] plots using each method introduced in the main text. We select two representative evaluation scenarios, GTSRB→TT100K and Belga→Flickr32, for visualization. The result shows a clear difference between the feature distribution of VPE and the remaining feature spaces. It should be noted that VPE generates a more discriminative feature distribution compared to those by the competing approaches.

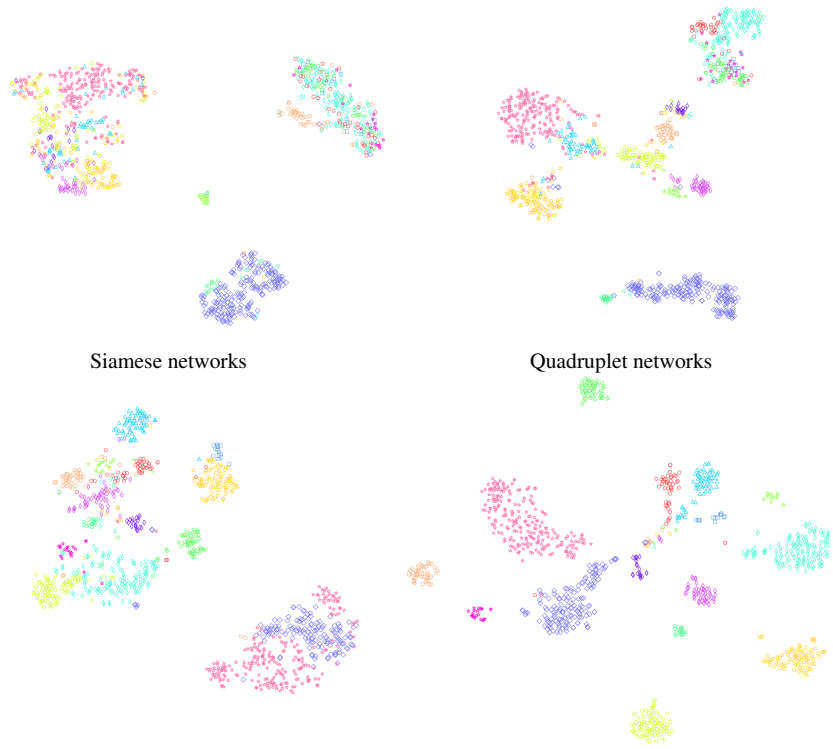

Figure 2. t-SNE visualization of features on embedding space. Features are randomly sampled from 15 different unseen classes under the GTSRB→TT100K scenario for visualization.

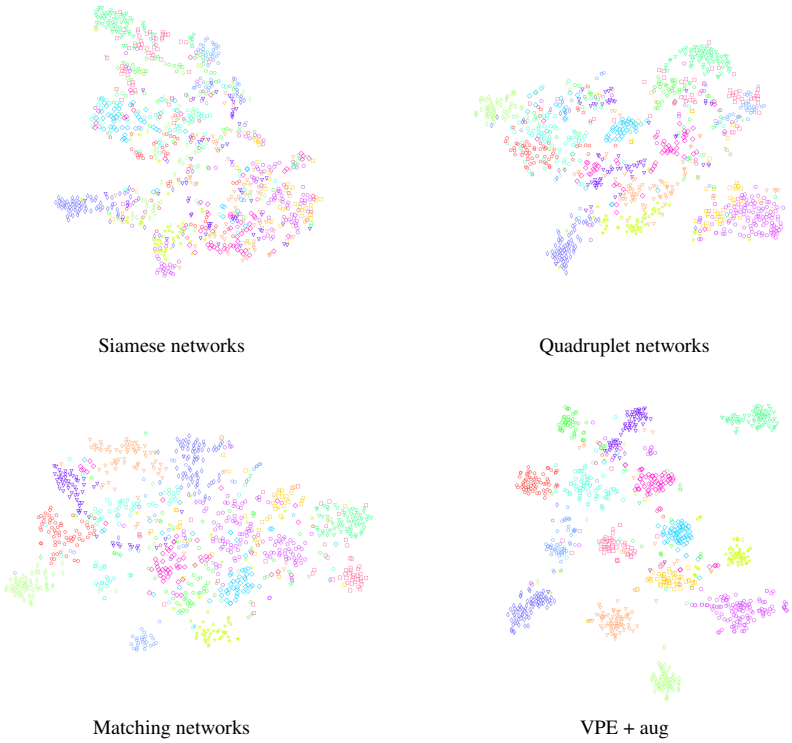

Figure 3. t-SNE visualization of features on embedding space. Features are randomly sampled from 15 different unseen classes under the Belga→Flickr32 scenario for visualization.

1.4. Image retrieval test

We show more image retrieval results that could not be placed in the main text due to space constraints. The average images of the top 100 images retrieved by querying unseen prototypes in each scenario are displayed. The columns from left to right are the average images retrieved using the Siamese networks [3], Quadruplet networks [2], Matching networks [10] and by the proposed method.

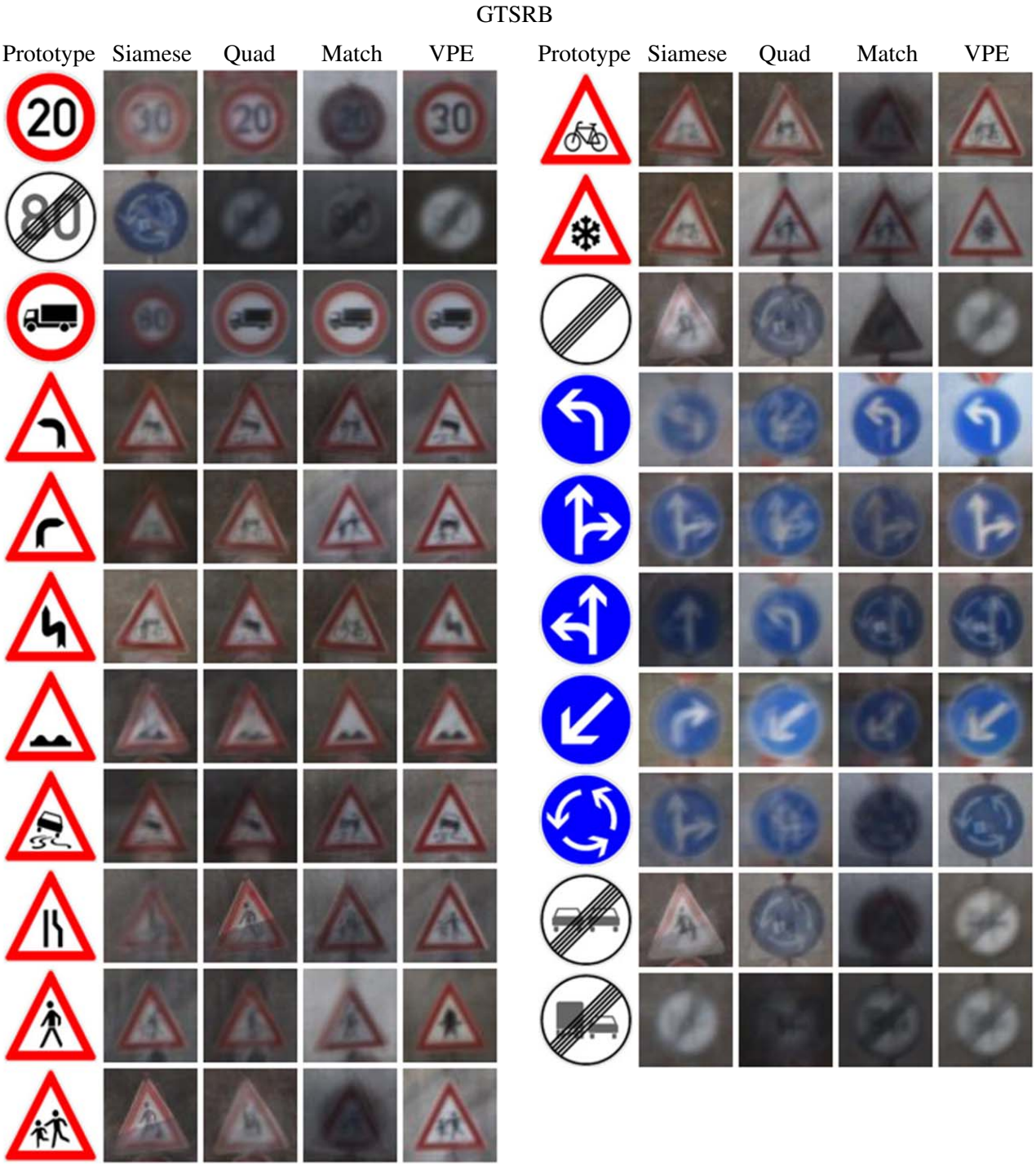

Figure 4. Average images of top 100 retrieved images by querying unseen prototypes in the GTSRB scenario.

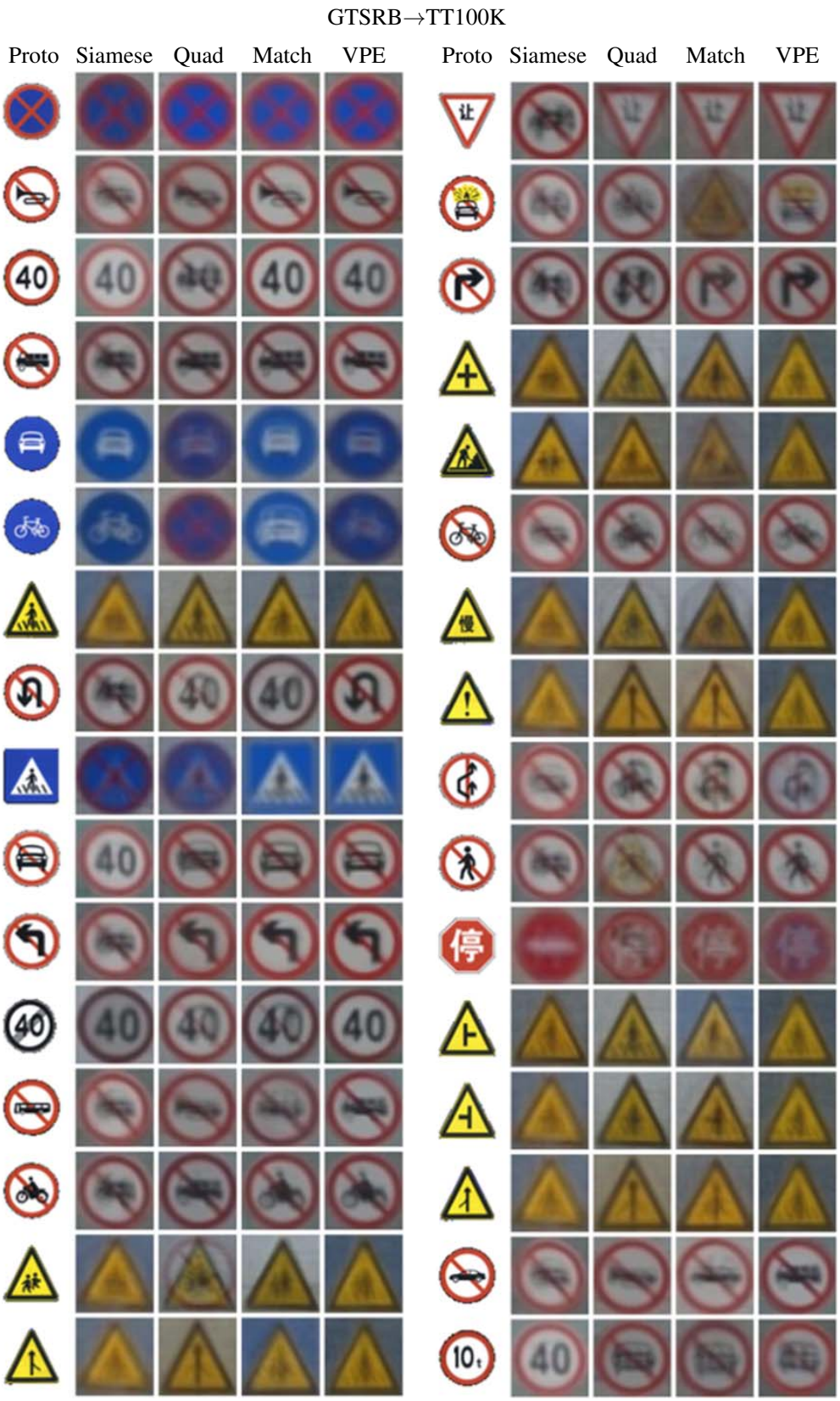

Figure 5. Average images of top 100 retrieved images by querying unseen prototypes in the GTSRB→TT100K scenario.

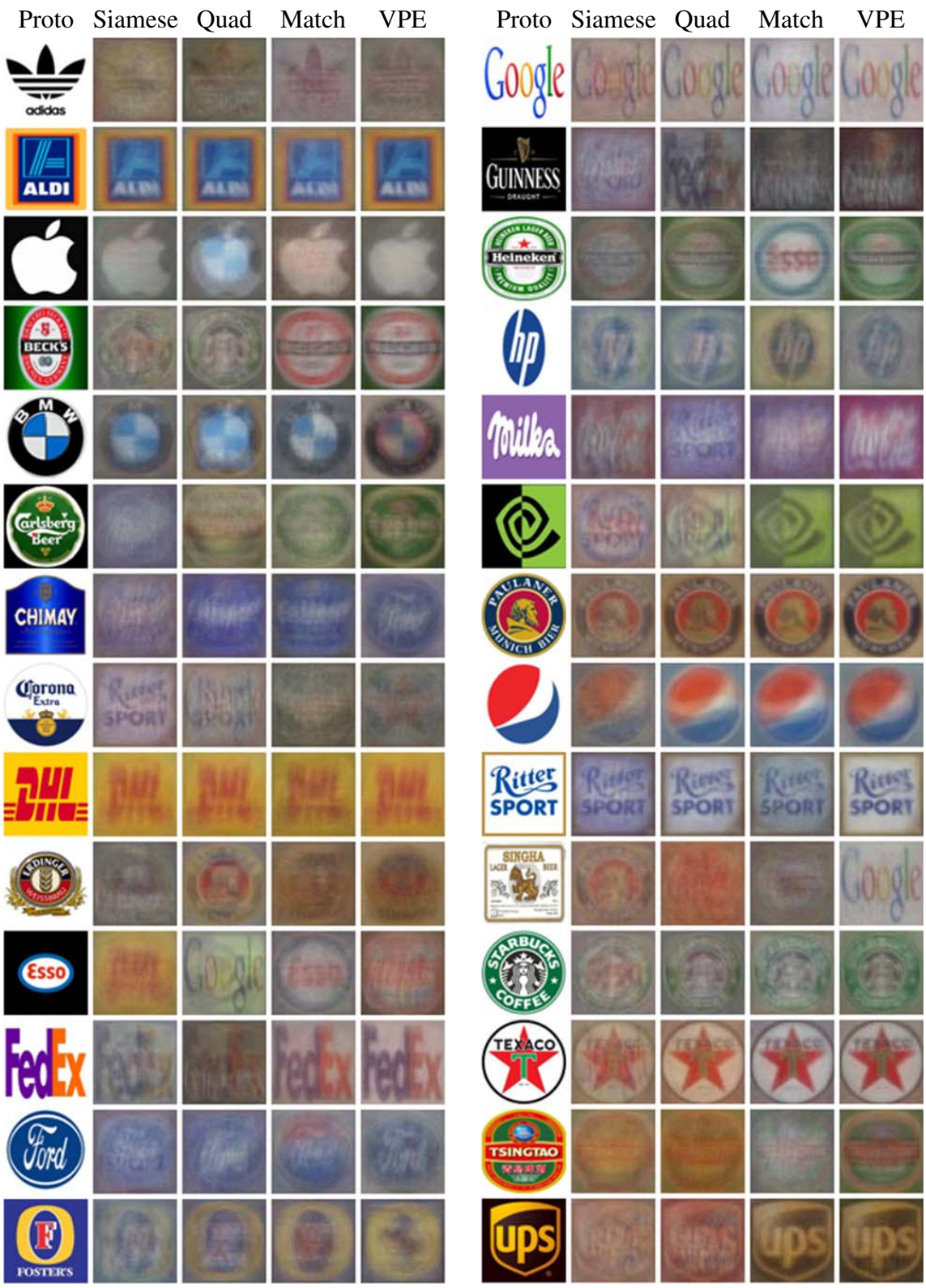

Figure 6. Average images of top 100 retrieved images by querying unseen prototypes in the Belga→Flickr32 scenario.

References

[1] A. Joly and O. Buisson. Logo retrieval with a contrario visual query expansion. In *Proceedings of the 17th ACM international conference on Multimedia*, 2009. 2

[2] J. Kim, S. Lee, T.-H. Oh, and I. S. Kweon. Co-domain embedding using deep quadruplet networks for unseen traffic sign recognition. In *AAAI*, 2018. 2, 4

[3] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, 2015. 4

[4] P. Letessier, O. Buisson, and A. Joly. Scalable mining of small visual objects. In *Proceedings of the 20th ACM international conference on Multimedia*, 2012. 2

[5] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 2

[6] S. Romberg, L. G. Pueyo, R. Lienhart, and R. Van Zwol. Scalable logo recognition in real-world images. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, 2011. 2

[7] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012. 2

[8] H. Su, X. Zhu, and S. Gong. Deep learning logo detection with data expansion by synthesising context. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2017. 2

[9] A. Tüzkö, C. Herrmann, D. Manger, and J. Beyerer. Open set logo detection and retrieval. *arXiv preprint arXiv:1710.10891*, 2017. 2

[10] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, 2016. 4

[11] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu. Traffic-sign detection and classification in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2