Class Anchor Margin Loss for Content-Based Image Retrieval

Alexandru Ghita and Radu Tudor Ionescu[a]
Department of Computer Science, University of Bucharest, 14 Academiei, Bucharest, Romania

Keywords: Contrastive Learning, Contrastive Loss, Learnable Class Anchors, Content-Based Image Retrieval, Object Retrieval.

Abstract: Loss functions play a major role in influencing the effectiveness of neural networks in content-based image retrieval (CBIR). Existing loss functions can be categorized into metric learning and statistical learning. Metric learning often lacks efficiency due to pair mining, while statistical learning does not yield compact features. To this end, we introduce a novel repeller-attractor loss based on metric learning, which directly optimizes the L_2 metric, without pair generation. Our novel loss comprises three terms: one to ensure features are attracted to class anchors, one that enforces anchor separability, and one that prevents anchor collapse. We evaluate our objective, applied to both convolutional and transformer architectures, on CIFAR-100, Food-101, SVHN, and ImageNet-200, showing that it outperforms existing functions in CBIR.

1 INTRODUCTION

Content-based image retrieval (CBIR) seeks to find images in a database similar to a given query image. Current systems for this task rely on deep neural networks (Cao et al., 2020; Dubey, 2021; Lee et al., 2022; Radenović et al., 2019; Revaud et al., 2019; Wu et al., 2021). While neural models outperform handcrafted CBIR models (Philbin et al., 2007), a key challenge in training neural networks for retrieval is selecting the objective function. Indeed, the employed function should enhance the discriminative power of the learned representations, ensuring small differences for images of the same object and large differences for different objects, thus adapting the embeddings for image retrieval.

Most neural loss functions fall into two main types: statistical learning (e.g., cross-entropy loss (Murphy, 2012), hinge loss (Cortes and Vapnik, 1995)) and metric learning (e.g., contrastive loss (Hadsell et al., 2006), triplet loss (Schroff et al., 2015), quadruplet loss (Chen et al., 2017)). Statistical learning losses are optimized by minimizing a probability distribution, indirectly achieving the properties required for image retrieval. This makes them more suitable for classification, and less for retrieval. Metric learning losses work directly over the embedding space, optimizing the targeted distance metric. However, they often require forming tuples during optimization, which can be time-consuming. To address this, hard example mining schemes have been used (Georgescu and Ionescu, 2021; Georgescu et al., 2022; Harwood et al., 2017; Schroff et al., 2015; Suh et al., 2019; Wu et al., 2017), but statistical learning is still more efficient.

In this research context, we introduce a novel repeller-attractor loss that belongs to metric learning. The proposed loss directly optimizes the L_2 metric, circumventing the costly hard example mining schemes. Our loss consists of three terms tied to learnable class anchors. The primary term attracts image embeddings to designated anchors, the second term enforces anchor separability with a margin, and the third term prevents anchors from collapsing to the origin.

We conduct few-shot CBIR experiments on the CIFAR-100 (Krizhevsky, 2009), Food-101 (Bossard et al., 2014), SVHN (Netzer et al., 2011), and ImageNet-200 (Russakovsky et al., 2015) datasets, comparing our loss with popular statistical and metric learning objectives. Our evaluation includes both convolutional and transformer architectures, e.g. ResNets (He et al., 2016) and Swin transformers (Liu et al., 2021). The results demonstrate that our objective consistently outperforms existing functions across the entire range of datasets and architectures.

In summary, our contribution is twofold:

- We propose a novel repeller-attractor loss that directly optimizes the L_2 metric, alleviating the need

[a] https://orcid.org/0000-0002-9301-1950

Ghita, A. and Ionescu, R.
Class Anchor Margin Loss for Content-Based Image Retrieval.
DOI: 10.5220/0012400500003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 848-853
ISBN: 978-989-758-680-4; ISSN: 2184-433X

of generating pairs via example mining schemes.

- We conduct few-shot retrieval experiments to compare the proposed loss function with popular loss choices on multiple datasets.

2 RELATED WORK

Loss Functions. Generating tightly grouped features for similar images and distant features for distinct images is a challenging task in CBIR. For neural retrieval systems, the objective function plays a crucial role in shaping the geometry of the latent space (Tang et al., 2022). We review related work on diverse loss functions for effective embedding spaces.

Metric learning objectives, often involving pairs or tuples of data samples, directly optimize targeted metrics (Sohn, 2016; Vassileios Balntas and Mikolajczyk, 2016; Yu and Tao, 2019; Chen et al., 2017). The contrastive loss (Hadsell et al., 2006), an early metric learning approach, creates an attractor-repeller system based on positive and negative pairs, which promotes feature cluster separation by enforcing a distance margin m for negative pairs. Triplet loss (Schroff et al., 2015) selects triplets of anchor, positive, and negative examples. For each triplet, it aims to make the distance between the anchor and the positive example lower than the distance between the anchor and the negative example plus a margin m. Other approaches optimize AUC (Gajić et al., 2022), recall (Patel et al., 2022), or AP (Revaud et al., 2019). Common issues of metric learning include slow convergence (Sohn, 2016) and difficulty in pair or tuple generation (Suh et al., 2019). In opposition, our method achieves faster convergence without the need for mining strategies.

Common example mining strategies include hard, semi-hard, and online negative mining (Schroff et al., 2015; Wu et al., 2017). Hard negative mining involves constructing pairs with the farthest positive and closest negative examples for each anchor image, adding a computational step at the start of each training epoch, increasing training time. Semi-hard negative mining can sample negatives closer to positives for better stability, but the mining is still not very efficient. Online negative mining, a more efficient approach, samples negatives during training in each batch, adjusting pair difficulty dynamically. However, because the mini-batch sampling is random, the generated pairs are not always sufficiently hard.

Statistical learning objectives indirectly optimize neural network embeddings. Popular ones, like variations of cross-entropy loss (Deng et al., 2019; Liu et al., 2017; Wang et al., 2018b; Wang et al., 2018a; Liu et al., 2016; Elezi et al., 2022) or cosine loss (Barz and Denzler, 2020), make the model generate embeddings in the vicinity of class centers. For instance, ArcFace (Deng et al., 2019) narrows the optimization space to an n-dimensional hypersphere by normalizing embeddings and corresponding class weights. ArcFace uses an additive penalty term, optimizing the angle between each feature vector and its class center.

Hybrid losses aim for improved embeddings by minimizing both a statistical and a metric learning objective (Min et al., 2020; Khosla et al., 2020). For instance, Center Loss (Wen et al., 2016) minimizes intra-class distances via cross-entropy and an attractor between each example and its class center. Class centers shift to the mean embedding of each class in a batch. A related approach (Zhu et al., 2019) uses predefined class centers. A more elaborate approach (Cao et al., 2020) combines cross-entropy with a cosine classifier and a mean squared error regressor to improve global and local features.

Unlike contrastive and triplet loss objectives that necessitate pair mining, our method eliminates the need for mining strategies. We form positive pairs online for each batch, connecting each embeddings with its class anchor, resulting in a number of positive pairs equal to the number of examples. Negative pairs are based on class centers, removing the need for a negative mining strategy and substantially reducing the number of negative pairs. To our knowledge, existing loss functions do not use self-repelling learnable class anchors as attraction poles for image embeddings.

CBIR. CBIR systems (Dubey, 2021) employ image descriptors to find images similar to a query based on scene or object similarity. These descriptors can be divided into general (Radenović et al., 2019), which represent whole images, and local (Wu et al., 2021), which represent local image regions. They can also be combined into hybrid descriptors (Cao et al., 2020). To enhance the results of global descriptors, a verification step usually re-ranks images using a more precise evaluation (Polley et al., 2022). This step is often carried out by an independent method which can be integrated into the global descriptor (Lee et al., 2022). CBIR can be employed to find visually similar images, or images with similar regions (Philbin et al., 2007) to the query. In this study, we focus on global descriptors. Unlike related methods, we introduce a novel loss function that harnesses class anchors to improve the efficiency and effectiveness of the learning process.

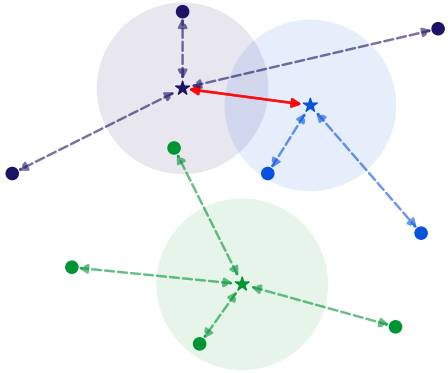

Figure 1: An illustration of our attractor-repeller loss for three classes. Stars depict class anchors C, faded circles represent spheres of radius m around each anchor, and solid circles show embeddings given by the model f_θ. Dashed arrows depict the attraction forces of the attractor $\mathcal{L}_A$, while the solid red arrow between class anchors represents the repelling force of the repeller $\mathcal{L}_R$. Best viewed in color.

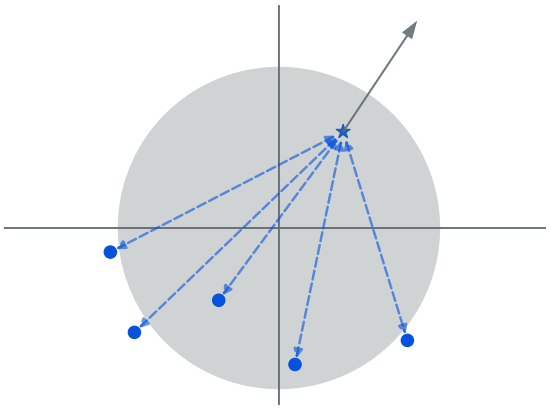

Figure 2: Influence of the minimum norm loss $\mathcal{L}_N$ on a class anchor. The blue star is a class anchor, and solid circles are embeddings returned by f_θ. Dashed arrows show attraction forces of the attractor $\mathcal{L}_A$, while the solid gray line represents the direction in which the anchor is repelled by the origin via $\mathcal{L}_N$. Best viewed in color.

3 METHOD

Overview. Our objective function, formulated around learnable class anchors, has three terms that jointly produce a discriminative embedding space. The first term aims to attract input embeddings to their class anchors, forming clusters of similar images. Making an analogy to how magnets interact in physics, class centers (anchors) can be seen as positive magnets and image embeddings as negative magnets, creating attraction forces within the same class. The second term aims to repel class anchors from each other, treating class centers as magnets with similar charges. Bringing them close makes them repel each other, but at a sufficiently large distance, the repelling force has no effect. The last term introduces an additional fixed magnet (with a positive charge) at the origin, pushing class centers away from it.

Notations. Let $\mathbf{x}_i \in \mathbb{R}^{h\times w\times c}$ be an input image and $y_i \in \mathbb{N}$ its associated class label, $\forall i \in \{1,2,...,l\}$. We aim to optimize a neural encoder f_θ which is parameterized by the learnable weights θ to produce a discriminative embedding space. Let $\mathbf{e}_i \in \mathbb{R}^n$ be the n-dimensional embedding vector of the input image $\mathbf{x}_i$ generated by f_θ, i.e. $\mathbf{e}_i = f_\theta(\mathbf{x}_i)$. In order to employ our novel loss function, we need to introduce a set of learnable class anchors $C = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_t\}$, where $\mathbf{c}_j \in \mathbb{R}^n$ resides in the embedding space of f_θ, and t is the total number of classes.

Loss Components. With the above notations, we can now formally define our first term, the attractor loss $\mathcal{L}_A$, as follows:

$$\mathcal{L}_A(\mathbf{x}_i, C) = \frac{1}{2}\|\mathbf{e}_i - \mathbf{c}_{y_i}\|_2^2. \tag{1}$$

This term aims to minimize the distance between each feature vector $\mathbf{e}_i$ and its designated class anchor $\mathbf{c}_{y_i}$, ensuring low intra-class variance by grouping embeddings around their class anchors. However, the attractor loss has little effect on inter-class similarity. To address this, we introduce the following repeller loss $\mathcal{L}_R$:

$$\mathcal{L}_R(C) = \frac{1}{2}\sum_{y,y'\in Y, y\neq y'} \left\{\max\left(0, 2\cdot m - \|\mathbf{c}_y - \mathbf{c}_{y'}\|\right)\right\}^2, \tag{2}$$

where y and y' are distinct labels from the ground-truth set Y, and $m > 0$ is the margin, representing the radius of an n-dimensional sphere around each anchor, preventing other anchors from residing within. This term aims to increase inter-class distances by pushing anchors apart. The margin m restricts the repelling force to an acceptable range, preventing excessive separation that might hinder the encoder from learning features satisfying the attractor loss defined in Eq. (1).

As shown in Figure 1, the attractor-repeller mechanism optimizes the model by pulling same-class samples together (based on the attractor term) and pushing different-class samples apart (based on the repeller term). However, initializing the weights with small random values puts the initial embeddings in random locations near the origin. Thus, right from the start, the samples tend to pull class anchors towards the origin. To prevent anchors from collapsing to the origin, we employ an additional loss to enforce a minimum norm on class anchors, as follows:

$$\mathcal{L}_N(C) = \frac{1}{2}\sum_{y\in Y}\left\{\max\left(0, p - \|\mathbf{c}_y\|\right)\right\}^2, \tag{3}$$

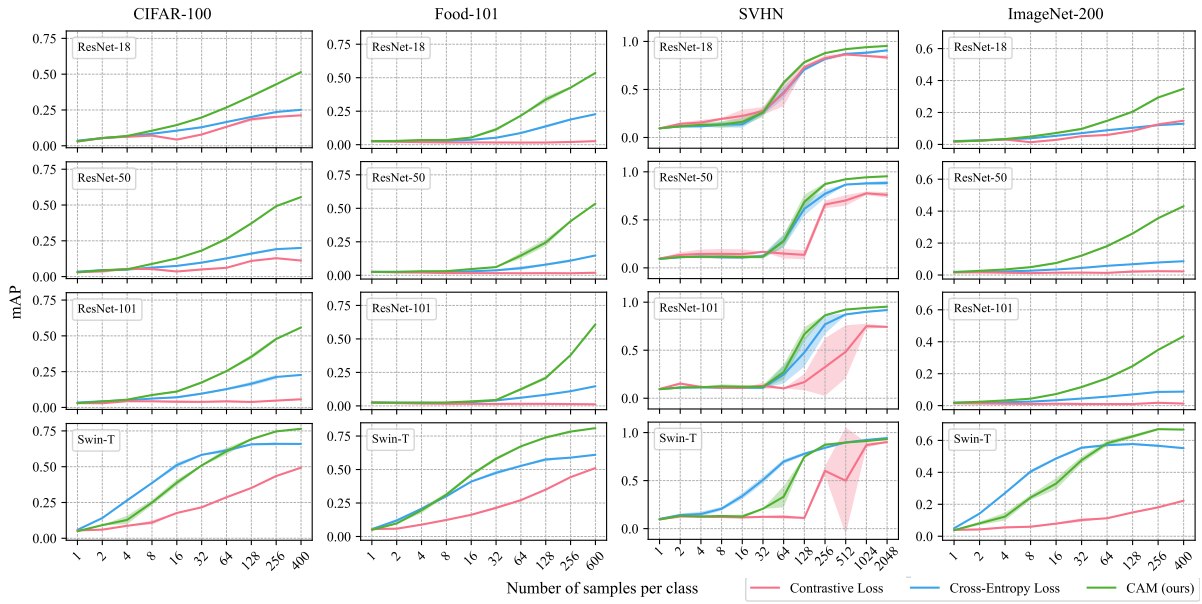

Figure 3: Few-shot retrieval performance (mAP) of four models (ResNet-18, ResNet-50, ResNet-101 and Swin-T) on four datasets (CIFAR-100, Food-101, SVHN and ImageNet-200). On each dataset, the results are shown from one sample per class (one-shot learning) to the maximum number of samples per class, by doubling the number of training samples in each trial.

where p is the minimum norm of a class anchor. This objective contributes to our full loss when at least one class anchor is within a distance p from the origin. Figure 2 shows how the minimum norm loss pushes the class anchor away from the origin, while the attractor loss moves the associated data samples along with their anchor.

We combine the three loss terms introduced so far into a joint objective, which represents the proposed class anchor margin (CAM) loss $\mathcal{L}_{\text{CAM}}$. Our novel loss is formally defined as follows:

$$\mathcal{L}_{\text{CAM}}(\mathbf{x}, C) = \mathcal{L}_A(\mathbf{x}, C) + \mathcal{L}_R(C) + \mathcal{L}_N(C). \quad (4)$$

We hereby underline that only $\mathcal{L}_A$ operates on the training examples, while $\mathcal{L}_R$ and $\mathcal{L}_N$ only attend to the class anchors. Therefore, CAM loss does not require the use of negative mining strategies.

4 EXPERIMENTS

4.1 Datasets

We experiment on four datasets: CIFAR-100 (Krizhevsky, 2009), Food-101 (Bossard et al., 2014), SVHN (Netzer et al., 2011), and ImageNet-200 (Russakovsky et al., 2015). CIFAR-100 contains 50,000 training images and 10,000 test images from 100 classes. Food-101 contains 101,000 images from 101 categories, with 750 training images and 250 test images per category. SVHN contains 73,257 training digits and 26,032 test digits. ImageNet-200 is a subset of ImageNet-1K, which contains 100,000 training images, 25,000 validation images and 25,000 test images from 200 classes.

4.2 Experimental Setup

We consider three ResNet (He et al., 2016) models (ResNet-18, ResNet-50, ResNet-101) and a Swin transformer (Swin-T) (Liu et al., 2021). The models are implemented in PyTorch (Paszke et al., 2019), and testing is performed with Hydra (Yadan, 2019).

We initialize all models randomly, except Swin-T, which starts with weights pretrained on ImageNet (Russakovsky et al., 2015). We optimize all networks using Adam (Kingma and Ba, 2015), setting the learning rate to 10^{-3} for ResNets and 10^{-4} for Swin-T. We train the ResNets from scratch for 100 epochs, and fine-tune Swin-T for 30 epochs. The mini-batch size is 512 for ResNet-18 and ResNet-50, 128 for ResNet-101, and 64 for Swin-T. Residual models use a linear learning rate decay with a factor of 0.5, and a patience of 9 epochs. Input images are normalized such that all pixels belong to $[0, 1]$. For Swin-T, we use ImageNet statistics (Russakovsky et al., 2015) to standardize images.

When employing our loss, we set $m = 2$ and $p = 1$ for all datasets and models. Contrastive models are optimized in the feature space, so we use the nearest

neighbor model based on the learned embeddings for CBIR.

We choose the mean Average Precision (mAP) to evaluate retrieval models. We repeat each experiment for 5 times and compute the average performance and the standard deviation.

4.3 Results

For each dataset, we sample a certain number of training images from each class, starting from one example per class. We gradually increase the number of training samples, doubling the number of training examples per class in each experiment, until we reach the maximum amount of available images. In Figure 3, we present the corresponding results on CIFAR-100 (Krizhevsky, 2009), Food-101 (Bossard et al., 2014), SVHN (Netzer et al., 2011), and ImageNet-200 (Russakovsky et al., 2015).

For all three ResNet models, our CAM loss outperforms contrastive loss when the training samples are four or more per class. Contrastive loss consistently attains the worst performance, trailing both cross-entropy and class anchor margin losses. Cross-entropy is the best choice for Swin-T as long as the number of samples per class is below 64. However, as the number of samples increases, our loss exhibits a lower performance gap and even surpasses cross-entropy, typically beyond 128 samples per class. Generally, all models improve with more training data. With very few exceptions, CAM loss consistently achieves the best performance, which confirms that it is suitable for CBIR. Our loss achieves optimal results with more than four samples per class.

5 CONCLUSION

In this paper, we proposed a novel loss function based on class anchors to optimize convolutional networks and transformers for object retrieval in images. We performed few-shot CBIR experiments employing four neural models on four image datasets, showing that our loss function outperforms conventional losses based on statistical learning and contrastive learning.

In future work, we aim to extend the applicability of our approach to other data types, beyond images. We also aim to explore new tasks and find out when our loss is likely to outperform the commonly used cross-entropy.

REFERENCES

Barz, B. and Denzler, J. (2020). Deep learning on small datasets without pre-training using cosine loss. In *Proceedings of WACV*, pages 1360–1369. IEEE.

Bossard, L., Guillaumin, M., and Van Gool, L. (2014). Food-101 – Mining Discriminative Components with Random Forests. In *Proceedings of ECCV*, pages 446–461. Springer.

Cao, B., Araujo, A., and Sim, J. (2020). Unifying deep local and global features for image search. In *Proceedings of ECCV*, pages 726–743. Springer.

Chen, W., Chen, X., Zhang, J., and Huang, K. (2017). Beyond Triplet Loss: A Deep Quadruplet Network for Person Re-identification. In *Proceedings of CVPR*, pages 1320–1329. IEEE.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of CVPR*, pages 4685–4694. IEEE.

Dubey, S. R. (2021). A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2687–2704.

Elezi, I., Seidenschwarz, J., Wagner, L., Vascon, S., Torcinovich, A., Pelillo, M., and Leal-Taixe, L. (2022). The Group Loss++: A deeper look into group loss for deep metric learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2505–2518.

Gajić, B., Amato, A., Baldrich, R., van de Weijer, J., and Gatta, C. (2022). Area under the ROC curve maximization for metric learning. In *Proceedings of CVPR*, pages 2807–2816. IEEE.

Georgescu, M.-I., Duţă, G.-E., and Ionescu, R. T. (2022). Teacher-student training and triplet loss to reduce the effect of drastic face occlusion: Application to emotion recognition, gender identification and age estimation. *Machine Vision and Applications*, 33(1):12.

Georgescu, M.-I. and Ionescu, R. T. (2021). Teacher-student training and triplet loss for facial expression recognition under occlusion. In *Proceedings of ICPR*, pages 2288–2295. IEEE.

Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *Proceedings of CVPR*, volume 2, pages 1735–1742. IEEE.

Harwood, B., Kumar, V. B., Carneiro, G., Reid, I., and Drummond, T. (2017). Smart mining for deep metric learning. In *Proceedings of ICCV*, pages 2821–2829. IEEE.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of CVPR*, pages 770–778. IEEE.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. In *Proceedings of NeurIPS*, volume 33, pages 18661–18673. Curran Associates, Inc.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *Proceedings of ICLR*.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Lee, S., Seong, H., Lee, S., and Kim, E. (2022). Correlation verification for image retrieval. In *Proceedings of CVPR*, pages 5374–5384. IEEE.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L. (2017). SphereFace: Deep Hypersphere Embedding for Face Recognition. In *Proceedings of CVPR*, pages 6738–6746, Los Alamitos, CA, USA. IEEE.

Liu, W., Wen, Y., Yu, Z., and Yang, M. (2016). Large-margin softmax loss for convolutional neural networks. In *Proceedings of ICML*, pages 507–516. JMLR.org.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of ICCV*, pages 10012–10022. IEEE.

Min, W., Mei, S., Li, Z., and Jiang, S. (2020). A two-stage triplet network training framework for image retrieval. *IEEE Transactions on Multimedia*, 22(12):3128–3138.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. (2011). Reading digits in natural images with unsupervised feature learning. In *Proceedings of NIPS 2011 Workshop on Deep Learning and Unsupervised Feature Learning*. Curran Associates, Inc.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of NeurIPS*, pages 8024–8035. Curran Associates, Inc.

Patel, Y., Tolias, G., and Matas, J. (2022). Recall@k surrogate loss with large batches and similarity mixup. In *Proceedings of CVPR*, pages 7502–7511. IEEE.

Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of CVPR*, pages 1–8. IEEE.

Polley, S., Mondal, S., Mannam, V. S., Kumar, K., Patra, S., and Nürnberger, A. (2022). X-vision: Explainable image retrieval by re-ranking in semantic space. In *Proceedings of CIKM*, pages 4955–4959, New York, NY, USA. Association for Computing Machinery.

Radenović, F., Tolias, G., and Chum, O. (2019). Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668.

Revaud, J., Almazán, J., Rezende, R. S., and Souza, C. R. d. (2019). Learning with Average Precision: Training Image Retrieval with a Listwise Loss. In *Proceedings of ICCV*, pages 5107–5116. IEEE.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115:211–252.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of CVPR*, pages 815–823. IEEE.

Sohn, K. (2016). Improved Deep Metric Learning with Multi-class N-pair Loss Objective. In *Proceedings of NIPS*, volume 29. Curran Associates, Inc.

Suh, Y., Han, B., Kim, W., and Lee, K. M. (2019). Stochastic Class-Based Hard Example Mining for Deep Metric Learning. In *Proceedings of CVPR*, pages 7244–7252. IEEE.

Tang, Y., Bai, W., Li, G., Liu, X., and Zhang, Y. (2022). CROLoss: Towards a Customizable Loss for Retrieval Models in Recommender Systems. In *Proceedings of CIKM*, pages 1916–1924, New York, NY, USA. Association for Computing Machinery.

Vassileios Balntas, Edgar Riba, D. P. and Mikolajczyk, K. (2016). Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Proceedings of BMVC*, pages 119.1–119.11. BMVA Press.

Wang, F., Cheng, J., Liu, W., and Liu, H. (2018a). Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Li, Z., Gong, D., Zhou, J., and Liu, W. (2018b). CosFace: Large Margin Cosine Loss for Deep Face Recognition. In *Proceedings of CVPR*, pages 5265–5274. IEEE.

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *Proceedings of ECCV*, pages 499–515. Springer.

Wu, C.-Y., Manmatha, R., Smola, A. J., and Krähenbühl, P. (2017). Sampling matters in deep embedding learning. In *Proceedings of ICCV*, pages 2859–2867. IEEE.

Wu, H., Wang, M., Zhou, W., and Li, H. (2021). Learning deep local features with multiple dynamic attentions for large-scale image retrieval. In *Proceedings of ICCV*, pages 11416–11425. IEEE.

Yadan, O. (2019). Hydra - a framework for elegantly configuring complex applications. Github.

Yu, B. and Tao, D. (2019). Deep metric learning with tuplet margin loss. In *Proceedings of ICCV*, pages 6489–6498. IEEE.

Zhu, Q., Zhang, P., Wang, Z., and Ye, X. (2019). A New Loss Function for CNN Classifier Based on Predefined Evenly-Distributed Class Centroids. *IEEE Access*, 8:10888–10895.