Bayesian Triplet Loss: Uncertainty Quantification in Image Retrieval

Frederik Warburg[†], Martin Jørgensen[‡], Javier Civera[§], and Søren Hauberg[†]
[†]Technical University of Denmark, [‡]University of Oxford, [§]University of Zaragoza
[†]{frwa,sohau}@dtu.dk, [‡]martinj@robots.ox.ac.uk, [§]jcivera@unizar.es

Abstract

Uncertainty quantification in image retrieval is crucial for downstream decisions, yet it remains a challenging and largely unexplored problem. Current methods for estimating uncertainties are poorly calibrated, computationally expensive, or based on heuristics. We present a new method that views image embeddings as stochastic features rather than deterministic features. Our two main contributions are (1) a likelihood that matches the triplet constraint and that evaluates the probability of an anchor being closer to a positive than a negative; and (2) a prior over the feature space that justifies the conventional l_2 normalization. To ensure computational efficiency, we derive a variational approximation of the posterior, called the Bayesian triplet loss, that produces state-of-the-art uncertainty estimates and matches the predictive performance of current state-of-the-art methods.

1. Introduction

Image-based retrieval systems show impressive performance in challenging tasks such as face verification [43, 52, 56], instance retrieval [63], landmark retrieval [38] and place recognition [1, 41]. These systems typically embed images into high-level features and retrieve with a nearest neighbor search. While this is efficient, the retrieval comes with no notion of confidence, which is particularly problematic in safety-critical applications. For instance, a self-driving car relying on visual place recognition should be able to filter out place estimates drawn from uninformative images. In a less critical but still relevant application, quantifying the retrieval uncertainty can significantly improve the user experience in human-computer interfaces by not showing low-confidence results for a query.

Practical retrieval systems do not have a small set of predefined classes as output targets, but rather need high-level features that generalize to unseen classes. For instance, a visual place recognition system may be deployed in a city in which it has not been trained [58]. This is achieved by keeping the encoder fixed and relying on nearest neighbor searches. This pipeline does not easily match current methods for posterior inference, and current uncertainty estimators for retrieval are often impractical and heuristic in nature. To construct a fully

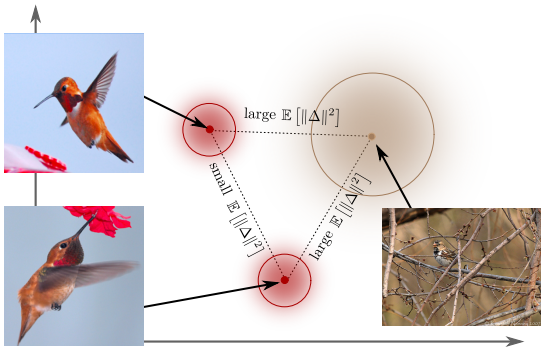

Figure 1: We model embeddings as distributions rather than point estimates, such that data uncertainty is propagated to retrieval. We phrase a Bayesian model that mirrors the triplet loss, which enables us to learn the stochastic features.

Bayesian retrieval system that fits with existing computational pipelines, we first recall the elementary equation

$$\mathbb{E}\left[\|\Delta\|^2\right] = \|\mathbb{E}[\Delta]\|^2 + \text{trace}(\text{cov}[\Delta]), \qquad \Delta \in \mathbb{R}^D, \quad (1)$$

which follows directly from the definition of variance. From this we see that the expected squared distance between two random features, $\mathbb{E}\left[\|\Delta\|^2\right]$, grows with the covariance of such distance, $\text{cov}[\Delta]$, which in turn depends on the uncertainty of the features (Figure 1). This intuition forms the basis of this paper.

In this paper we propose to use stochastic image embeddings instead of the usual deterministic ones. Given an image X, we consider the posterior distribution over possible features $P(F|X)$. From this distribution we get direct uncertainty estimates and can assign probabilities to events such as 'two images belonging to the same place'. To realize this, we derive a likelihood corresponding to the probability that the conventional triplet constraint is satisfied, and a prior over the feature space that mimics conventional l_2 normalization. To build a system that is computationally efficient at both train and test time, we derive a variational approximation to the posterior $P(F|X)$, such that in practice, we encode an image to a distribution in feature space. Across several datasets, we show that the proposed model matches the state-of-the-art in predictive performance, while attaining state-of-the-art uncertainty estimates.

2. Related Work

Image retrieval has been a popular research problem in the last years due to its many applications [46, 63]. Early approaches relied on handcrafted local features aggregated mainly by bag-of-words [22, 39, 45]. More recent models are composed by a deep convolutional backbone followed by an aggregation layer [1, 41] that map images to low dimensional embeddings based on their content similarity [1, 3, 35, 38, 41–43, 52, 54, 56]. The most similar images to a query are found by (approximate) nearest neighbor methods.

Retrieval systems apply either classification losses (e.g., [30, 62]) or metric losses (e.g., [15]). Metric losses operate on the relationships between samples in a batch, while classification losses include a weight matrix that transforms the embeddings into vectors of class logits [34]. Our focus is on metric losses, of which the contrastive loss [19] is the most fundamental one. However, this loss has the limitation that the same margin thresholds apply to all training pairs, even though there may be a large variation in their similarities. The triplet loss [59] accounts for such varying interclass dissimilarities by solely constraining an anchor a to be closer to a positive p than a negative n minus a margin m,

$$\|a-p\|^2 < \|a-n\|^2 - m. \tag{2}$$

Many works have extended the contrastive and triplet losses to incorporate more structural information about the embedding space, for example the quadruplet loss [5], N-pair loss [47], angular loss [55], margin loss [60], signal-to-noise ratio (SNR) contrastive loss [61] and multi-similarity (MS) loss [57]. These methods are often supported by heuristics or empirical experiments, but lack a theoretical grounding. Additionally, in a recent study, Musgrave et al. [34] show that these more complex loss functions offer only marginal improvements over the contrastive and triplet losses. For this reason we focus on the triplet loss, but in principle our approach can be extended to mirror other losses.

Uncertainty in deep networks is hard to quantify due to the large number of parameters [13]. Uncertainty quantification is currently being studied in the context of many computer vision tasks, among others depth completion [12, 18], semantic segmentation [18, 23, 27], object detection [6, 17], object pose estimation [37] and multi-task learning [24]. In practice, Bayesian approximations such as deep ensembles [29], Monte Carlo dropout (MC dropout) [14] and conditional autoencoders [26] have shown most promise. Although scalable [18], these methods do not directly apply to image retrieval, as models typically do not have proper likelihood functions.

Der Kiureghian and Ditlevsen [9] identify the sources of predictive uncertainties as the model (epistemic uncertainty) and the data (aleatoric uncertainty). The latter can be divided into homoscedastic (constant for all input data) and heteroscedastic (variable depending on the particular input). We focus on heteroscedastic uncertainty as this is especially relevant for image retrieval as illustrated in Figure 1.

Uncertainty quantification for image retrieval using deep networks is a challenging and under-addressed topic. Learning stochastic embeddings rather than deterministic ones has been addressed *i.e.* for images [4, 36, 44], human poses [49] and cross-modal data [8, 48]. Most prior work has focused on classification [4] or pairwise losses. Oh et al. [36] use Monte Carlo (MC) sampling from learned Gaussian distributions to evaluate a matching probability between a pair of images. Their approach is successful for low dimension embeddings ($D=3$). The expensive Monte Carlo sampling can be avoided by directly optimizing the likelihood that two Gaussian embeddings belong to the same class [44]. In our work, we extend this intuition to triplets. Triplet losses, contrary to pairwise ones, are known to address varying interclass similarities and dissimilarities [34]. Sun et al. [49] suggest a ratio of two binary cross entropy terms. Taking the log of these gives an expression on the form of the triplet loss. However, their approach relies on MC samples. Taha et al. [51] cast the triplet loss as a regression loss and estimate epistemic uncertainty with MC dropout. In later work, Taha et al. [50] propose to learn the heteroscedastic uncertainty using a noise parameter per image. In contrast to our proposal, no model is provided under which the proposed triplet losses is a proper likelihood. We benchmark our method against a MC sampling method (triplet regression with MC dropout [51]) and a implicit method (triplet regression that implicitly learns a noise parameter [50]), showing significantly better uncertainty estimates in several datasets while matching the predictive performance of the standard triplet loss.

3. Bayesian Triplet Loss

We propose to embed images as distributions rather than point estimates. Given these stochastic embeddings, we ask: What is the probability that an anchor is closer to a positive than a negative, *i.e.*

$$P\big(\|a-p\|^2 < \|a-n\|^2 - m\big), \tag{3}$$

which is the probabilistic equivalent of the triplet constraint (2). To realize this idea, we first derive a likelihood function corresponding to Eq. 3. As we want this to follow the intuition of the triplet loss closely, our likelihood operates on triplets of images. We define our likelihood for all triplets in the dataset,

$$\mathcal{L}(\Omega) = \prod_{X\in\Omega} \prod_{Y\in\Omega} \prod_{Z\in\Omega} P(I(X,Y,Z)|X,Y,Z), \tag{4}$$

where X, Y, Z are images from the dataset and I is the label of a triplet (referred to as triplet label). We note that the triplet labels can take three values with respectively 1, 2 or 3 images from the same class. Since $I(X,Y,Z)$ only takes values on a discrete finite set, we define our *likelihood function* as the multinomial distribution. The traditional triplet loss ignores the situations where all images are from the same or different classes, as these situations are not informative. Making a similar modeling choice,

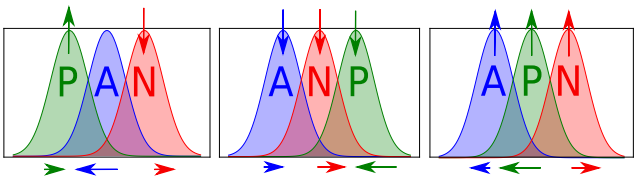

Figure 2: Intuition for the Bayesian triplet loss in three 1D scenarios. The arrows below the figures indicate the gradient direction and magnitude of the means, while the arrows above the distributions indicate the gradients of the variances (downwards indicate more spread, upwards means more peaky).

the likelihood reduces to only consider triplets with one pair

$$P(I(X,Y,Z)|X,Y,Z) = \mathcal{P}^{\mathbb{1}\{I(X,Y,Z)=2\}}. \tag{5}$$

Thus, all probability mass is in triplets where two images are from the same class and one from a different class, just as with the traditional triplet loss. Using the standard triplet notation, we set $\mathcal{P}$ equal to Eq. 3, thus we derived a proper likelihood function that describes the probability of an anchor being closer to a positive than a negative. We adopt this simpler notation throughout the remainder of the paper to make it clear which images are from the same class, but emphasize that the likelihood is defined for *all* triplets via Eq. 5.

Figure 3 shows the proposed negative log-likelihood compared to the traditional triplet loss. Our likelihood is smooth and bounded, making it more robust to outliers. We experience similar training time as with the traditional triplet loss and have not experienced zero-gradients that block learning.

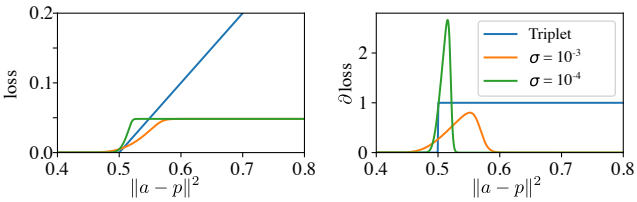

Figure 3: The traditional triplet loss (blue) compared to our negative log likelihood (orange and green). The negative log likelihood is smooth and bounded, yielding better robustness.

3.1. The Triplet Likelihood

Having the likelihood form in place, we proceed to arrive at explicit expressions. We assume that the embeddings are isotropic Gaussians rather than points, such that $x \sim \mathcal{N}(\mu_x, \sigma_x^2 I)$, where $x \in \{a,p,n\}$ (see Figure 1). This will be justified in the next section. Rearranging Eq. 3 gives

$$P(\|a-p\|^2 - \|a-n\|^2 < -m) = P(\tau < -m), \tag{6}$$

$$\text{where } \tau = \sum_{d=1}^{D} (a_d - p_d)^2 - (a_d - n_d)^2, \tag{7}$$

and D is the feature dimension. The squared distance between two normally distributed random variables follows a scaled non-central χ^2-distribution [32]. The likelihood (6) is a linear combination of two such distributed squared distances, which does not have a known density, and we resort to approximations.

By the central limit theorem [31], τ will approximate a Gaussian distribution for large D, *i.e.*

$$\lim_{D\to\infty} P\left(\frac{\tau-\mu}{\sigma} < -m\right) = \Phi(-m), \tag{8}$$

where Φ is the CDF of the standard normal distribution, and μ and σ are the mean and standard deviation of τ. In the supplements, we experimentally demonstrate that this approximation is remarkably accurate even in low dimensions.

We still need to find the leading two moments of τ to apply this approximation. We provide detailed derivations in the supplements, and here only sketch the steps.

The mean is determined as

$$\mathbb{E}[\tau] = \mathbb{E}[p(p-2a)] - \mathbb{E}[n(n-2a)]. \tag{9}$$

We exploit the symmetry and write first

$$\mathbb{E}[p(p-2a)] = \mathbb{E}[p^2] - 2\mathbb{E}[ap] = \mathbb{E}[p^2] - 2\mathbb{E}[a]\mathbb{E}[p], \tag{10}$$

since a and p are independent. Using identical arguments for the second term of $\mathbb{E}[\tau]$ we get

$$\begin{aligned}\mathbb{E}[\tau] &= \mathbb{E}[p^2] - 2\mathbb{E}[a]\mathbb{E}[p] - \mathbb{E}[n^2] + 2\mathbb{E}[a]\mathbb{E}[n] \\ &= \mathbb{E}[p^2] - \mathbb{E}[n^2] - 2\mathbb{E}[a](\mathbb{E}[p] - \mathbb{E}[n]) \\ &= \mu_p^2 + \sigma_p^2 - \mu_n^2 - \sigma_n^2 - 2\mu_a(\mu_p - \mu_n).\end{aligned} \tag{11}$$

The variance requires a longer derivation, so we only present the result here,

$$\begin{aligned}\frac{\text{Var}(\tau)}{2} &= \sigma_p^2(\sigma_p^2 + 2\mu_p^2) + \sigma_n^2(\sigma_n^2 + 2\mu_n^2) - 4\sigma_a^2\mu_p\mu_n \\ &\quad - 2\mu_a\left(\mu_a(\mu_p^2 + \mu_n^2) - 2\mu_p\sigma_p^2 - 2\mu_n\sigma_n^2\right) \\ &\quad + 2(\sigma_a^2 + \mu_a^2)\left((\sigma_p^2 + \mu_p^2) + (\sigma_n^2 + \mu_n^2)\right).\end{aligned} \tag{12}$$

Thus, given a mean and a variance estimate for each image in the triplet, we can analytically compute the mean and variance of τ. Then the likelihood (5) is evaluated by a Gaussian likelihood with these parameters.

The intuition of the likelihood function is shown in Figure 2 for three 1D scenarios. In the left figure, the ordering is correct (anchor is closer to the positive than the negative). The gradients w.r.t the variances are negative, thus reducing the uncertainty of each stochastic embedding (indicated with the arrows above the distributions). In the center figure, the ordering is incorrect (anchor is closer to the negative than the positive) resulting in higher uncertainties. The arrows below the plots indicate the gradient direction and magnitude w.r.t. the means. In all scenarios the mean of the anchor and positive are attracted, while the mean of the negative is repelled.

3.2. Normalization Priors

It is common practice in image retrieval to l_2-normalize the embeddings as this often boosts retrieval performance [1, 40, 41]. A further practical benefit is that for l_2-normalized vectors, the Euclidean distance and the cosine similarity have a monotonic relation, and thus can be interchanged without altering the retrieval order. The cosine similarity is computationally efficient as it reduces to the dot product for normalized vectors [40]. We investigate two priors to imitate this normalization.

In high dimensions, the standard Gaussian distribution concentrates around a sphere of radius D. Hence, we can mimic the l_2 normalization by imposing a Gaussian prior over the embeddings. In particular, the prior $p(x) = \mathcal{N}(x|0, \frac{1}{D}I)$ concentrates around the unit sphere, and can therefore be seen as an *implicit* l_2 normalization. We also consider an *explicit* normalization prior, by choosing a uniform prior over the unit sphere.

3.3. The Approximate Posterior

The posterior embedding is generally intractable, and we resort to variational approximations for computational efficiency [2]. We choose a parametrized approximate posterior q as an isotropic distribution from the same family as the prior. With the Gaussian prior we choose $q(X) = \mathcal{N}(\mu_X, \sigma_X^2 I)$. With the uniform spherical prior we choose the approximate posterior as a von Mises Fisher distribution $q(X) = \text{vMF}(\mu_X, \kappa_X)$ [10]. The distribution parameters are here described by the neural network.

In the supplements we derive the Expected Lower Bound (ELBO) for the marginal likelihood to be the right-hand side here

$$\begin{aligned}\log P(I(X,Y,Z)) \geq \mathbb{E}_{q(X)q(Y)q(Z)}[\log P(I(X,Y,Z)|X,Y,Z)] \\ - \text{KL}(q(X)\|p(X)) - \text{KL}(q(Y)\|p(Y)) \\ - \text{KL}(q(Z)\|p(Z)).\end{aligned} \tag{13}$$

With the chosen distribution families, the KL divergences have closed-form expressions [11, 33].

4. Network Architecture and Training

For each image we learn an isotropic distribution rather than a point embedding. We treat both Gaussian and von Mises Fisher embeddings identically, and here only describe the Gaussian setup. Similar to Taha et al. [50], we use a shared backbone network followed by a mean and a variance head (see Fig. 4). The mean head is a generalized mean (GeM) [41] aggregation layer followed by a fully connected layer that outputs $\mu \in \mathbb{R}^D$. The variance head consists of a GeM layer followed by two fully connected layers with a ReLU activation function. We found it was advantageous to estimate σ^2 with a softplus activation rather than estimating $\log\sigma^2$. We have separate GeM layers for the variance and mean heads, as we found it beneficial to learn different p-norms for the variance head and mean head.

In real world applications this trade-off between predictive performance and uncertainty quantification is important.

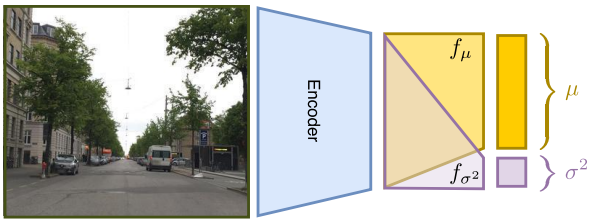

Figure 4: Overview of our network architecture.

Therefore, we ensure that the number of output parameters is the same for probabilistic and non-probabilistic models, such that $D_\mu + D_\sigma = D$. We focus on isotropic distributions and set $D_\sigma = 1$. For the triplet loss, we follow common practice and l_2-normalize the point estimates, that is $x/\|x\|_2 \in \mathbb{R}^D$. For the Bayesian triplet loss, we l_2-normalize the mean embedding $\mu/\|\mu\|_2 \in \mathbb{R}^{D_\mu}$ for the uniform prior, and scale with a single positive trainable parameter for the Gaussian prior.

We use a hard negative mining strategy similar to Arandjelovic et al. [1]. Given a query image, we find the closest negative images in a cache. We only present the model with the triplets that violate the triplet constraint (2). We update the cache with 5000 new images every 1000 iterations. Arandjelovic et al. [1] and Warburg et al. [58] report the importance of updating the cache regularly to avoid overfitting. This speeds up learning by reducing the number of trivial examples presented to the model.

5. Evaluation Metrics

The **Recall at k** (R@k) measures the number of queries that have at least one positive among their closest k neighbors. This is a commonly used metric for image retrieval. However, this metric does not take into account the ratio of positives and negatives among the neighbors. Therefore, we also evaluate the **Mean Average Precision at k** (mAP@k) which measures the precision of the k closest neighbors [1]. These metrics evaluate the predictive performance of our models.

The **Expected Calibration Error** (ECE) describes how well a model's uncertainties correspond with its predicted accuracy, and is a common metric to measure model calibration in classification tasks [16, 18]. The predictions are divided into M equally spaced bins based on their confidences. For each bin B_m, the accuracy is compared to the model confidence, and weighted by the bin size. We reformulate this metric to fit for retrieval problems. Confident queries should have a high mAP and unconfident queries low mAP. We can therefore let ECE@k measure the weighted distance between mAP@k and the M^{th} percentiles of the variance. We set $M = 10$.

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{mAP@k}(B_m) - \text{conf}(B_m)|. \tag{14}$$

6. Experiments and Results

We conduct experiments across three challenging image retrieval datasets. For each of the experiments, we compare

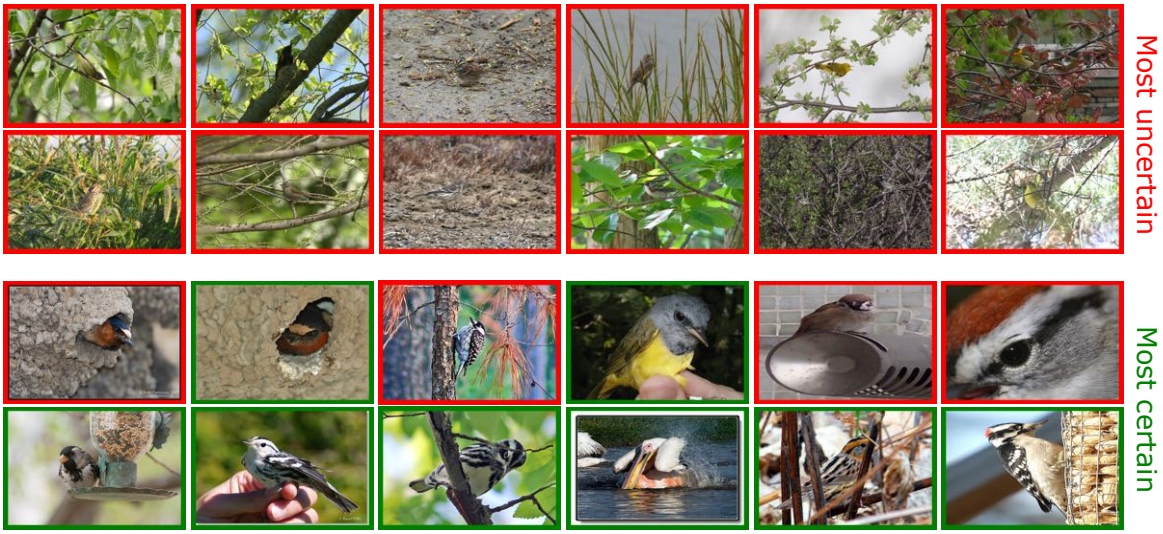

Figure 5: Query images for which our Bayesian embedding gives the highest (two first rows) and lowest uncertainty (two last rows). Scenes associated with high uncertainty mostly correspond to scenes where birds blend in with the background and are hardly discernible. The two most certain ones correspond to Cliff Swallows, easily discernible by the characteristic mud nests that they cement to walls or cliffs. In all images, birds stand out from the background and have unique patterns.

the traditional triplet loss [59] with the proposed Bayesian triplet loss with the Gaussian prior (Bayes Triplet) and with the von-Mises Fisher prior (Bayes vMF). We also compare our model's uncertainty estimates with those produced by the triplet regression [50] and MC dropout [51]. We evaluate on two strong backbones, namely Resnet50 [20] ($D = 2048$) and the large Densenet161 [21] ($D = 2208$), to illustrate that the Bayesian triplet loss provides calibrated uncertainty estimates across different architectures. Densenet161 is chosen specifically as it applies dropout in the backbone, which allows us to compare the uncertainty estimates with MC dropout. For comparison with triplet regression [51] we consistently use a dropout rate of 0.2 as reported in [51]. All models are implemented in Pytorch and trained with the Adam optimizer [25] with learning rate 10^{-5}, weight decay 0.001 and an exponential learning rate scheduler decreasing the learning rate by 1% per epoch. We use batches of 25 triplets, each triplet consisting of one anchor, one positive and five negative images. We resize the images to 224×224 in all the experiments. During training, we augment the data with random rotations (up to $10°$), resized cropping ($[0.4; 1]$ of image size), color jitter and horizontal flipping. We use a KL-scale factor 10^{-6} in all experiments.

6.1. CUB 200-2011

We first evaluate the models' retrieval performance and calibration performance. The CUB 200-2011 dataset [53] consists of $11{,}788$ images of 200 bird species. The birds are captured from different perspectives, making this a challenging dataset for image retrieval. We divide the first 100 classes into the training set and the last 100 classes into the test set similarly to Musgrave et al. [34]. Thus, the trained models have not seen any of the bird species in the test set, and the learned features must generalize well across species.

Table 1 shows the retrieval performance for the CUB200 dataset for the models with both a Resnet50 and Densenet161 backbone. We notice that the larger backbone improves the retrieval performance across all models. MC dropout performs worse than the other models in terms of predictive performance and uncertainty quantification. The triplet loss and the triplet regression have slightly better predictive performance than the proposed Bayesian models, however both of the Bayesian models, especially with the Gaussian embeddings, produce notably better uncertainty estimates.

To gain a better understanding of which images have high and low variance estimates, Fig. 5 shows the 12 queries with highest and lowest variance. The (green/red) border indicates if the image's nearest neighbor is correctly retrieved. The network correctly associates high variance to images where the bird blends in with its surroundings, and low variance to birds that are centered or easily distinguishable by their color or patterns.

For applications, the user will often be interested in the covariance of the distance $\mathrm{cov}[\Delta]$ between a query and its nearest neighbor. The covariance depends on both the variance of the query and its nearest neighbor. Figure 6 shows six examples of queries and nearest neighbors which exhibit high or low covariances. The model assigns low confidence when multiple birds are present or when a bird blends into its surroundings. High confidence is associated with birds with unique patterns or colors.

Figure 7 shows the Bayesian model with Gaussian embeddings is better calibrated than other methods which produce uncertainty estimates, especially for uncertain queries. The queries are divided into 10 equi-sized bins. For each bin the

Figure 6: Six query images (top) and their NN (bottom), all true positives. Our Bayesian model assigns low confidence to images with multiple birds or where birds blend in with their surroundings. Birds with distinguishable patterns or colors have high confidence.

		R@1	R@5	R@10	M@1	M@5	M@10	ECE@1	ECE@5	ECE@10
ResNet50	Triplet	**0.648**	**0.864**	**0.917**	**0.648**	**0.505**	**0.440**			
	TripReg	0.643	0.863	0.916	0.643	0.503	0.437	0.196	0.331	0.397
	Bayes vMF	0.635	0.855	0.911	0.635	0.492	0.426	0.138	0.064	0.089
	Bayes Triplet	0.612	0.842	0.902	0.612	0.467	0.397	**0.119**	0.037	0.099
	Bayes Triplet (Exp)	0.632	0.857	0.912	0.630	0.489	0.424	0.137	**0.020**	**0.072**
Densenet161	Triplet	**0.717**	0.894	0.935	**0.717**	**0.598**	**0.537**			
	Triplet (MC=50)	0.349	0.591	0.700	0.349	0.200	0.143	0.290	0.428	0.480
	TripReg	0.711	0.897	**0.939**	0.711	0.587	0.524	0.181	0.302	0.363
	Bayes vMF	0.713	**0.898**	0.938	0.713	0.590	0.528	**0.022**	0.139	0.200
	Bayes Triplet	0.683	0.879	0.926	0.683	0.564	0.502	0.176	0.059	**0.023**
	Bayes Triplet (Exp)	0.617	0.838	0.902	0.617	0.494	0.437	0.101	**0.025**	0.075

Table 1: Recall (R), Mean Average Precision (M) and Expected Calibration Error (ECE) at 1, 5 and 10 on the CUB200 dataset. Bayes Triplet (Exp) refers to the Expected distance rather than the Euclidean distance was used for nearest neighbor search.

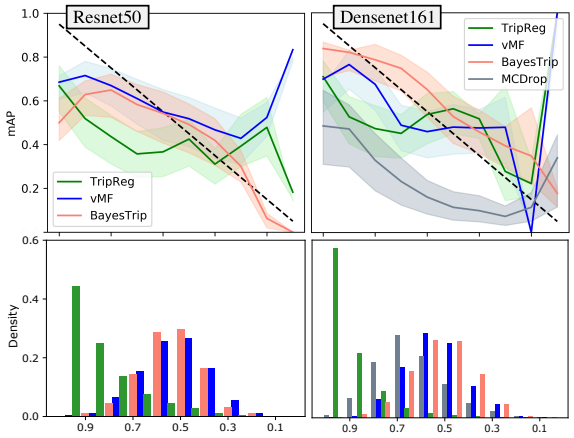

Figure 7: Calibration plots for triplet regression (TripReg), MC dropout (MCDrop), Bayesian triplet loss with Gaussian distribution (BayesTrip) and Bayesian triplet loss with von Mises-Fisher distribution (vMF).The solid line indicates mAP@5 and the shaded area covers from mAP@1 to mAP@10. Note how, for both backbones, the Gaussian distributed embeddings are better calibrated, especially for uncertain queries.

mAP@$\{1, 5, 10\}$ is calculated, indicated with the top part of the shaded area, the solid line, and lower part of shaded area respectively. The black dotted line shows a perfectly calibrated model.

This experiment shows our proposed Bayesian triplet loss produces retrieval performance comparable to the triplet loss across two strong backbones. Further, the experiment shows that our model produces very well-calibrated uncertainty estimates based on the ECE metric, calibration plots, and qualitative visualizations.

6.2. Stanford Car-196

Next, we investigate how well the models perform on out-of-distribution (OOD) examples. This is an important capability for image retrieval systems since it is infeasible to retrain the models continuously, and in many practical applications unseen categories are likely to be added to the database or used as queries over time. To test the OOD capabilities of the models, we use the Stanford Car-196 [28] dataset and the models trained on the CUB200 dataset. The Car-196 dataset is composed of 16,185 images of 196 classes of cars. It is traditionally a classification dataset, but can be cast as a retrieval dataset by using the first 98 categories as a training set and the last 98 ones as a test set [34].

First, we evaluate how well the models generalize to the Car-196 test set. Table 2 shows that both Bayesian models match (and for vMF embeddings with the ResNet50 backbone surpass) the retrieval performance of the triplet loss. The ECE metric shows that the Bayesian models are significantly better calibrated. Among the Bayesian models, the Gaussian

		R@1	R@5	R@10	M@1	M@5	M@10	ECE@1	ECE@5	ECE@10
ResNet50	Triplet	0.451	0.723	0.813	0.451	0.255	0.179			
	TripReg	0.447	0.712	0.813	0.447	0.254	0.178	0.292	0.479	0.555
	Bayes vMF	**0.471**	**0.733**	**0.827**	**0.471**	**0.270**	**0.191**	0.117	**0.162**	**0.225**
	Bayes Triplet	0.431	0.696	0.795	0.431	0.235	0.163	**0.094**	0.271	0.343
Densenet161	Triplet	**0.495**	**0.751**	**0.837**	**0.495**	**0.293**	**0.212**			
	Triplet (MC=50)	0.470	0.717	0.813	0.470	0.269	0.191	0.178	0.367	0.438
	TripReg	0.481	0.744	0.836	0.481	0.285	0.205	0.263	0.456	0.536
	Bayes vMF	0.478	0.740	0.834	0.474	0.284	0.204	0.160	0.288	0.355
	Bayes Triplet	0.467	0.724	0.814	0.467	0.269	0.192	**0.101**	**0.134**	**0.208**

Table 2: Recall (R), Mean Average Precision (M) and Expected Calibration Error (ECE) at 1, 5 and 10 on the CAR196 dataset.

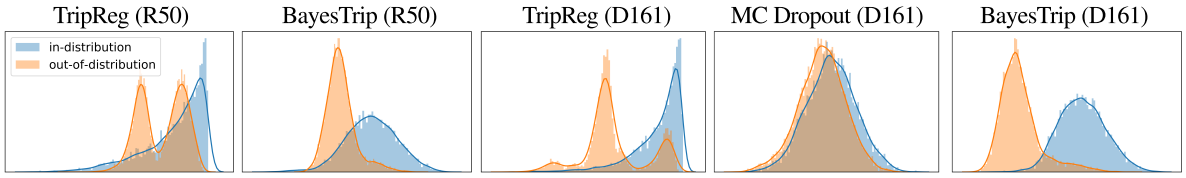

Figure 8: Histogram of the covariance of the distances between queries and their nearest neighbor for triplet regression (TripReg), Bayesian triplet model with Gaussian embeddings (BayesTrip) and MC Dropout for Resnet50 (R50) and Densenet161 (D161). Note that in-distribution (blue) and out-of-distribution (orange) covariances are significantly more separated for the Bayesian model.

Figure 9: Least probable among in and out of distribution for triplet regression (TripReg) and Bayesian triplet Loss with Gaussian embeddings (BayesTrip). Cars are out of distribution queries and birds are in distribution queries.

embeddings give better uncertainty estimates, whereas the vMF embeddings give slightly better predictive performance.

Second, we show that the Bayesian models can detect OOD queries. To do so, we construct a database consisting of bird images. We retrieve the nearest neighbor with bird queries – in distribution (ID) – and car queries – out-of-distribution (OOD). We expect the distance to the nearest neighbor of the OOD queries to have high uncertainty compared to ID queries. Figure 8 shows that the Bayesian model with Gaussian embeddings is significantly better at differentiating between ID and OOD queries.

The six least confident Gaussian query embeddings from our Bayesian model (Fig. 9) reveal OOD queries (cars), whereas for the triplet regression, several ID queries (birds) are seen among the least certain. Even though these queries are challenging birds, we would expect the model to be less confident about images from a completely different domain.

This experiment shows that the confidences generated by the Bayesian models generalize well to out-of-distribution examples and can be used to discriminate between queries from in- and out-of-distribution. Again, the Bayesian models match the predictive performance to the other methods, and achieve state-of-the-art uncertainty estimates.

6.3. Mapillary Street Level Sequences (MSLS)

Finally, we show that the Bayesian models also have competitive retrieval performance and state-of-the-art uncertainty estimates for large datasets. MSLS [58] is the largest and most diverse place recognition dataset currently available, and comprises $1.6M$ images from 30 cities spanning six continents. The goal in place recognition is to retrieve images from the same place as a query image (where the same place is typically defined within a radius of 25 m). This is challenging due to the large number of unique places and large spectrum of appearance changes for each place, such as weather, dynamic, structural, view-point, seasonal and day/night changes. We use the train/test split recommended in [58], training on 24 cities and testing on six other different cities.

The Bayesian models achieve a performance comparable to the traditional triplet loss, even outperforming it when using the von Mises-Fisher distribution (Table 3). Furthermore,

Figure 10: Low-confidence (columns 1–3) and high-confidence (columns 4 and 5) retrievals from the MSLS dataset. The top row shows queries and the bottom row shows their NNs. Our model gives low confidence to images with harsh sunlight, blur and ambiguous tunnels and vegetation. In contrast, high confidence is given for landmark buildings.

		R@1	R@5	R@10	M@1	M@5	M@10	ECE@1	ECE@5	ECE@10
ResNet50	Triplet	0.350	0.495	0.554	0.350	0.235	0.210			
	TripReg	0.349	**0.499**	0.551	0.349	0.238	0.215	0.482	0.571	0.593
	Bayes vMF	0.349	0.494	**0.559**	0.349	**0.249**	**0.226**	0.482	0.571	0.593
	Bayes Triplet	**0.354**	0.489	0.549	**0.354**	0.241	0.218	**0.208**	**0.319**	**0.341**
Densenet161	Triplet	**0.386**	**0.531**	0.583	**0.386**	**0.270**	**0.246**			
	Triplet (MC=50)	0.282	0.412	0.458	0.282	0.188	0.163	0.540	0.625	0.648
	TripReg	**0.386**	0.529	**0.596**	**0.386**	0.268	0.245	0.424	0.543	0.566
	Bayes vMF	0.383	0.526	0.588	0.383	0.268	0.245	0.227	0.327	0.350
	Bayes Triplet	0.364	0.506	0.571	0.364	0.252	0.228	**0.196**	**0.264**	**0.283**

Table 3: Recall (R), Mean Average Precision (M) and Expected Calibration Error (ECE) at 1, 5 and 10 on the MSLS dataset.

the Bayesian model with Gaussian embeddings provides state-of-the-art uncertainty estimates for both backbones.

Figure 10 illustrates how the Bayesian model with Gaussian embeddings associates low confidence to images that are difficult to retrieve due to harsh sunlight, blur or the ambiguous, repetitive patterns in tunnels. Furthermore, the model is able to assign high confidence to images that have unique structural appearance, as seen in the last two columns. This experiment shows that the Bayesian models have a competitive performance in very large and challenging datasets, matching the predictive performance of the standard triplet loss, and producing state-of-the-art uncertainty estimates.

6.4. Trade-off between predictive performance and uncertainty quantification

In many applications, reliable uncertainties are a requirement. This is common for method that needs to provide guarantees of certifiability. In safety-critical applications uncertainty is important as it can ensure timely interventions from the user e.g. when image retrieval is used in loop closure for robot localization.

Contemporary probabilistic methods often exhibit a slight decrease in performance over non-probabilistic methods, but the trade-off is worth making in the above-mentioned examples. We indeed observe a similar trend, where the Bayesian Triplet Loss is either on par or slightly below state-of-the-art in predictive performance, but is clear state-of-the-art in terms of uncertainty quantification. One reason for the decrease in predictive performance is that we have less free parameters for the mean prediction than other methods: To ensure a fair comparison, we constraint the different models to have the same number of parameters, which imply that we use some of our capacity on predicting σ. This lessens our capacity for mean predictions.

7. Conclusion

We have proposed to model image embeddings as stochastic features rather than point estimates. We derive a new likelihood that follows the intuition of the triplet loss, but works for stochastic features. We introduce a prior over the feature space that together with our likelihood enables us to learn either Gaussian distributed or von Mises-Fisher distributed stochastic features. The proposed method, the Bayesian triplet loss, produces state-of-the-art uncertainty estimates, without sacrificing predictive performance compared to the triplet loss. Quantification of uncertainty in image retrieval is vital for safety-critical applications, while reliable uncertainty estimates also open many other doors, for example related to interpretability or user-friendly retrieval interfaces. We speculate that reliable uncertainty estimates can also be used for hard negative mining and avoidance of outliers in query expansion [7].

Acknowledgments. This work was supported in part by a research grant (15334) from VILLUM FONDEN. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n^{o} 757360), from the Spanish Government (PGC2018-096367-B-I00) and the Aragon Government (DGA T45 17R/FSE). MJ is supported by a research grant from the Carlsberg foundation (CF20-0370).

References

[1] Relja Arandjelovic, Petr Gronát, Akihiko Torii, Tomás Pajdla, and Josef Sivic. Netvlad: CNN architecture for weakly supervised place recognition. *CoRR*, abs/1511.07247, 2015. URL http://arxiv.org/abs/1511.07247. 1, 2, 4

[2] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017. 4

[3] Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. Smooth-ap: Smoothing the path towards large-scale image retrieval. *arXiv preprint arXiv:2007.12163*, 2020. 2

[4] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition, 2020. 2

[5] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2017. 2

[6] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 502–511, 2019. 2

[7] Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. pages 1–8, 11 2007. ISBN 978-1-4244-1631-8. doi: 10.1109/ICCV.2007.4408891. 8

[8] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. *arXiv preprint arXiv:2101.05068*, 2021. 2

[9] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009. 2

[10] Inderjit S Dhillon and Suvrit Sra. Modeling data using directional distributions. Technical report, 2003. 4

[11] Tom Diethe. A note on the kullback-leibler divergence for the von mises-fisher distribution. *arXiv preprint arXiv:1502.07104*, 2015. 4

[12] Abdelrahman Eldesokey, Michael Felsberg, Karl Holmquist, and Michael Persson. Uncertainty-aware cnns for depth completion: Uncertainty from beginning to end. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12014–12023, 2020. 2

[13] Yarin Gal. Uncertainty in deep learning. *University of Cambridge*, 1(3), 2016. 2

[14] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016. 2

[15] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, 124(2): 237–254, 2017. 2

[16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *CoRR*, abs/1706.04599, 2017. URL http://arxiv.org/abs/1706.04599. 4

[17] Fredrik K Gustafsson, Martin Danelljan, Goutam Bhat, and Thomas B Schön. Energy-based models for deep probabilistic regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[18] Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 318–319, 2020. 2, 4

[19] Raia Hadsell, Sumit Chopra, and Yann Lecun. Dimensionality reduction by learning an invariant mapping. pages 1735 – 1742, 02 2006. ISBN 0-7695-2597-0. doi: 10.1109/CVPR.2006.100. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385. 5

[21] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. URL http://arxiv.org/abs/1608.06993. 5

[22] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3304–3311. IEEE, 2010. 2

[23] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015. 2

[24] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018. 2

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 5

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[27] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31:6965–6975, 2018. 2

[28] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 6

[29] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413, 2017. 2

[30] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 2

[31] Douglas C Montgomery and George C Runger. *Applied statistics and probability for engineers, (With CD)*. John Wiley & Sons, 2007. 3

[32] Robb J Muirhead. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons, 2009. 3

[33] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 4

[34] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check, 2020. 2, 5, 6

[35] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017. 2

[36] Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew C. Gallagher. Modeling uncertainty with hedged instance embedding. *CoRR*, abs/1810.00319, 2018. URL http://arxiv.org/abs/1810.00319. 2

[37] Brian Okorn, Mengyun Xu, Martial Hebert, and David Held. Learning orientation distributions for object pose estimation. *arXiv preprint arXiv:2007.01418*, 2020. 2

[38] Kohei Ozaki and Shuhei Yokoo. Large-scale landmark retrieval/recognition under a noisy and diverse dataset. *CoRR*, abs/1906.04087, 2019. URL http://arxiv.org/abs/1906.04087. 1, 2

[39] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 2

[40] Filip Radenovic. *Visual Retrieval with Compact Image Representations*. PhD Thesis CTU–CMP–2019–01, Center for Machine Perception, K13133 FEE Czech Technical University, Prague, Czech Republic, May 2019. 4

[41] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. Fine-tuning CNN image retrieval with no human annotation. *CoRR*, abs/1711.02512, 2017. URL http://arxiv.org/abs/1711.02512. 1, 2, 4

[42] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5107–5116, 2019.

[43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015. URL http://arxiv.org/abs/1503.03832. 1, 2

[44] Yichun Shi, Anil K. Jain, and Nathan D. Kalka. Probabilistic face embeddings. *CoRR*, abs/1904.09658, 2019. URL http://arxiv.org/abs/1904.09658. 2

[45] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV2003*. IEEE, 2003. 2

[46] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12):1349–1380, 2000. 2

[47] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in neural information processing systems*, pages 1857–1865, 2016. 2

[48] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. *CoRR*, abs/1906.04402, 2019. URL http://arxiv.org/abs/1906.04402. 2

[49] Jennifer J. Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. *CoRR*, abs/1912.01001, 2019. URL http://arxiv.org/abs/1912.01001. 2

[50] Ahmed Taha, Yi Ting Chen, Teruhisa Misu, Abhinav Shrivastava, and Larry Davis. Unsupervised data uncertainty learning in visual retrieval systems. *CoRR*, abs/1902.02586, 2019. URL http://arxiv.org/abs/1902.02586. 2, 4, 5

[51] Ahmed Taha, Yi-Ting Chen, Xitong Yang, Teruhisa Misu, and Larry Davis. Exploring uncertainty in conditional multi-modal retrieval systems. *CoRR*, abs/1901.07702, 2019. URL http://arxiv.org/abs/1901.07702. 2, 5

[52] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014. 1, 2

[53] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 5

[54] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 2

[55] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2593–2601, 2017. 2

[56] Mei Wang and Weihong Deng. Deep face recognition: A survey. *CoRR*, abs/1804.06655, 2018. URL http://arxiv.org/abs/1804.06655. 1, 2

[57] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019. 2

[58] Frederik Warburg, Søren Hauberg, Manuel López-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 4, 7

[59] Kilian Weinberger, J. Blitzer, and L. Saul. *Distance Metric Learning for Large Margin Nearest Neighbor Classification*, volume 10. 01 2006. 2, 5

[60] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2840–2848, 2017. 2

[61] Tongtong Yuan, Weihong Deng, Jian Tang, Yinan Tang, and Binghui Chen. Signal-to-noise ratio: A robust distance metric for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4815–4824, 2019. 2

[62] Andrew Zhai and Hao-Yu Wu. Classification is a strong baseline for deep metric learning. *arXiv preprint arXiv:1811.12649*, 2018. 2

[63] Liang Zheng, Yi Yang, and Qi Tian. Sift meets cnn: A decade survey of instance retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1224–1244, 2017. 1, 2