Asymmetric Image Retrieval with Cross Model Compatible Ensembles

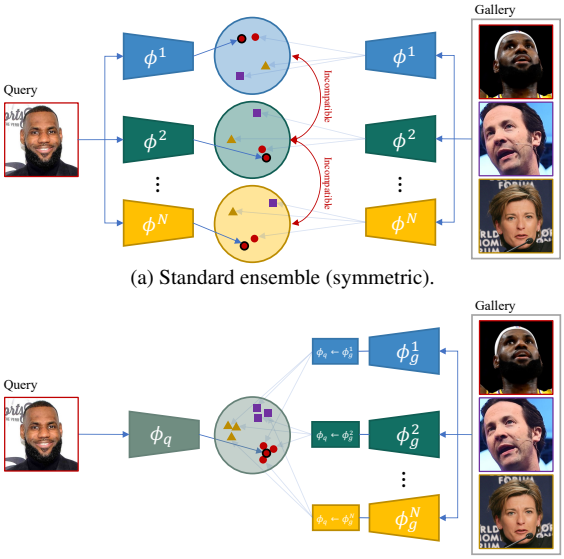
Alon Shoshan[*1] **Ori Linial**[*2] **Nadav Bhonker**[1] **Elad Hirsch**[2]
Lior Zamir[1] **Igor Kviatkovsky**[1] **Gérard Medioni**[1]
[1]Amazon [2]Technion - Israel Institute of Technology

Abstract

The asymmetrical retrieval setting is a well suited solution for resource constrained applications such as face recognition and image retrieval. In this setting, a large model is used for indexing the gallery while a lightweight model is used for querying. The key principle in such systems is ensuring that both models share the same embedding space. Most methods in this domain are based on knowledge distillation. While useful, they suffer from several drawbacks: they are upper-bounded by the performance of the single best model found and cannot be extended to use an ensemble of models in a straightforward manner. In this paper we present an approach that does not rely on knowledge distillation, rather it utilizes embedding transformation models. This allows the use of N independently trained and diverse gallery models (e.g., trained on different datasets or having a different architecture) and a single query model. As a result, we improve the overall accuracy beyond that of any single model while maintaining a low computational budget for querying. Additionally, we propose a gallery image rejection method that utilizes the diversity between multiple transformed embeddings to estimate the uncertainty of gallery images.

1. Introduction

Face recognition and image retrieval at scale are among the most challenging and widely studied topics in computer vision, having many practical applications. Modern systems are expected to provide extremely high retrieval accuracy under real time performance constraints and to support a very large number of classes (identities) in the gallery set. The majority of existing face recognition and image retrieval methods [2, 9, 32, 37, 45, 49] utilize a *symmetric retrieval* approach, *i.e.* the same model is used for extracting feature vectors (embeddings) for the gallery and for the query images. In a symmetric retrieval setting there is a clear trade-off between retrieval accuracy and computing resources. Using a larger model almost always achieves higher accuracy. However, it may prohibit use on user-side-devices due to limited computational and memory resources. On the other hand, using a lightweight model usually results in inferior representation capability and lower overall retrieval accuracy. Instead of compromising accuracy to meet hardware requirements, an *asymmetric retrieval* setting [5, 11, 51] can be incorporated, where a large



(a) Standard ensemble (symmetric).

(b) Embedding transformation based ensemble (asymmetric).

Figure 1. (a) Construction of an ensemble in the standard *symmetric retrieval* setting requires that each model will be used both for computing the gallery and for querying, since the embedding spaces of all models are incompatible. Hence, user-side querying introduces heavy computational costs, requiring to compute embeddings of N models and calculate distances in N embedding spaces. (b) We propose to construct an ensemble in an *asymmetric retrieval* setting, where a single lightweight model is used for querying and an ensemble of diverse independently trained models is used for computing the gallery (a one time process that occurs offline). Using embedding transformation models, $\phi_g^i \rightarrow \phi_q$, all gallery models' embeddings are transformed to the embedding space of the query model. Our approach benefits both from the increased accuracy that is associated with ensembles and from low computational requirements for querying.

[*]Contributed equally.

model is used for indexing the gallery offline, with sufficient computational resources, and a lightweight model is used for processing query images. The former is referred to as the *gallery model* and latter as the *query model*. Asymmetric retrieval is not straightforwardly accomplished by training both gallery and query models independently since their embedding spaces are incompatible. This is also known as the *cross model compatibility* (CMC) problem [47].

Previous asymmetric retrieval methods [5, 11, 51] use knowledge distillation [17] based approaches in order to restrict the query model embedding space to be compatible with that of the gallery model. These methods show significant accuracy improvements compared to methods using only the lightweight query model both for indexing and for querying (symmetric retrieval). However, the accuracy of all these methods is also upper-bounded by the gallery model accuracy and is usually lower than when a large model is used for both indexing and querying. In this work we present a new asymmetric retrieval approach that is not based on knowledge distillation. Instead, the approach utilizes embedding space transformations [47] which allows to transform multiple embedding spaces of different gallery models into the one of the query model. In contrast to previous methods, the resulting accuracy is no longer upper-bounded by that of the gallery model. We empirically demonstrate that our approach indeed breaches this "upper-bound". We note that applying an ensemble of models in the case of knowledge distillation is not straightforward. For example, it is not possible to independently train all of the ensemble components as they all have to rely on a common predefined model. This means that models acquired independently from different sources[1], cannot be combined into an ensemble under such a distillation scheme. Additionally, distilling models with the same level of accuracy while optimizing for CMC was shown to reduce the accuracy of the resulting model [36, 47].

The symmetric setting implicitly assumes that using an ensemble of models for computing the gallery will increase the computational cost of user-side querying (since the same ensemble has to be used for querying). It is not straightforward how to reduce the computational cost of the querying part since the model embedding spaces are incompatible, see Fig. 1. Attempts to remediate the incompatibility issue lead to asymmetric knowledge distillation methods, which, as described above, are sub-optimal for creating ensembles.

In addition to improving asymmetric retrieval accuracy, we introduce an uncertainty based gallery image rejection method. By leveraging the diversity between the multiple transformed embeddings of the same gallery image, we can estimate the uncertainty of the final transformed embedding. For both the face recognition and product retrieval domain, we show that excluding embeddings with high uncertainty significantly improves the overall accuracy. For example, by rejecting only 10% of the gallery embeddings we reduce the face recognition error by 17.4%. This rejection approach is only possible because our method allows multiple models to project the same image into a common embedding space, which is not possible in a symmetric setting.

Our approach is also suitable in the *compatible model update* setting [7,19,36,47,57]. In this setting, when a better model becomes available (by improved training data, model architectures, or training regimes), we are tasked to update the retrieval system using the new model without performing *backfilling*. Backfilling is the common naïve process of replacing the embeddings in the gallery set that have been generated by the old model with embeddings from the new model. This method is computationally expensive and in some practical cases not even possible since the gallery images might not be retained by the system.

We summarize our contributions as following:

1. We present a novel embedding transformation based ensemble for asymmetric image retrieval, which significantly improves accuracy without the need for additional computation when performing user-side querying.

2. We introduce a gallery image rejection method that leverages the diversity between multiple transformed embeddings. This method can prevent "hard to match" gallery images from registering, thus further improving accuracy.

3. We demonstrate that our approach can achieve high accuracy even in a challenging compatible model update setting.

2. Related work

2.1. Image retrieval

Modern image retrieval methods [1, 33, 46, 52, 56], rely on deep learning models that encode images to a low dimensional embedding space. Embedding vectors representing similar classes are mapped close to each other, while dissimilar ones are mapped far apart. Such an embedding space can be created by using a classification loss as a proxy [9,42,44,48,49] or by using metric learning techniques such as triplet loss or contrastive loss [4,8,14,18,35,41]. We note that all these approaches focus on the symmetrical retrieval setting.

2.2. Cross model compatibility

In recent years, increased attention has been given to the cross model compatibility (CMC) [47] problem. The

[1]For example, if the models were trained on different datasets that may no longer be available.

aim of this field is to ensure embeddings encoded by different models are compatible, a condition that is usually not possible when models are trained independently. Compatibility between models is critical for asymmetric retrieval settings where the query and gallery models are different. In a resource constrained scenario recent methods [5, 11, 51] propose a large model for indexing and a lightweight model for user-side querying. Compatible model update [7, 31, 36, 47, 54] is another asymmetric retrieval scenario where a new query model replaces an older one and backfilling is to be avoided (due to high computational cost or unavailability of gallery images). The query model is usually better than the gallery model in this scenario since it was trained later with improved architecture, an improved training scheme or more training data. Knowledge distillation based methods [5, 11, 31, 36, 51, 54] use a compatibility loss term during training, enforcing the query model's embedding space to reside in the same space as the gallery model's. Chen *et al*. [7] propose R^3AN that combines reconstruction, representation and regression techniques to transform embeddings from one model to another. Wang *et al*. [47] propose to learn transformations from the embedding spaces of both trained models to a unified embedding space while enforcing compatibility. In this work we demonstrate that our transformation based ensemble approach is applicable for both asymmetric retrieval scenarios.

2.3. Using Ensembles

Ensembles of models have been extensively used in machine learning to boost accuracy [10, 24, 26, 28, 34, 39, 55]. Diversity among the ensemble components is important for ensuring a performance gain compared to relying on each component individually [15, 25]. A straightforward way to achieve such diversity is by introducing variations in training data [3] or initialization conditions [22].

An important aspect is designing the fusion scheme used to aggregate the ensemble component outputs [13]. Such fusion may be performed in different stages of the ensemble. For example, in a closed-set scenario the fusion is done by combining the softmax class posteriors, leveraging the common representation between models [16, 38, 43]. Other approaches delay the fusion to an even later stage of combining final model predictions by performing weighted averaging or majority voting, *etc*. [21]. In this work we propose to use feature based fusion for image retrieval, combining embeddings of different models. This approach is non-trivial since embedding spaces of different models are typically incompatible. Despite the fact that ensembles improve accuracy significantly, they are often ignored due to their high resource requirements. In the asymmetric retrieval settings this limitation is entirely avoided, thus making ensembles a well suited approach.

3. Proposed approach

3.1. Image retrieval and cross model compatibility

In a typical image retrieval system a set of *gallery* images, I_g, are associated with C classes (or identities), $Y_g = \{y_i\}_{i=1}^{C}$. A gallery model, ϕ_g, maps each image $i_g \in I_g$ to the gallery embedding space, $\mathcal{G} \subseteq \mathbb{R}^n$. By applying ϕ_g to the entire gallery, we obtain the set of gallery embedding vectors E_g. At test time, we are presented with a query image, i_q, belonging to some class, y_q (not necessarily in Y_g). The query image is then consumed by a query model, ϕ_q, to produce a query embedding, e_q. Assuming a symmetric setting ($\phi_g = \phi_q$), we associate e_q with the class Y_g of the closest gallery embedding vector based on some distance metric $d(\cdot, \cdot)$. If e_q has no sufficiently close match in E_g, the query image is rejected. The CMC problem becomes relevant in an asymmetric setting ($\phi_g \neq \phi_q$), where both models are trained independently. In this case, ϕ_q maps images into a query embedding space, $\mathcal{Q} \subseteq \mathbb{R}^m$, that is incompatible with the gallery embedding space, $\mathcal{G}$.

3.2. Embedding transformation

3.2.1 Unified embedding space

To address the CMC problem, Wang *et al*. [47] suggest to train embedding transformation models, T_g and T_q, to transform both ϕ_g and ϕ_q embedding spaces into a unified embedding space. In the unified space, embedding vectors transformed from ϕ_g and ϕ_q are compatible with one another. To implement the transformation models, the authors propose using four consecutive Residual Bottleneck Transformation (RBT) modules [47], and a compatibility constraining training scheme. The general training scheme is as follows: all training images are encoded using ϕ_g and ϕ_q to produce training embedding sets F_g and F_q. During training, corresponding embeddings from F_g and F_q are transformed by T_g and T_q, respectively, to the unified space. To enforce compatibility in the unified space, a combination of three loss terms are applied: a similarity-, a KL-divergence- and a dual-classification-loss. The first two terms enforce similarity between embeddings, while the third term enforces the embedding spaces to be discriminative by identity. Furthermore, by using a shared classification head, the embedding spaces are constrained to be aligned.

3.2.2 Model-to-model transformation

In this work we modify the unified embedding approach so that only one embedding space is transformed. Specifically, the gallery's embedding space is transformed to the query's. This is achieved by following the same training scheme except that we set T_q to be an identity mapping, *i.e*., we learn a model to model transformation (M2M). Since the embed-

ding space used for querying does not undergo a transformation, we gain the following practical benefits:

1. Multiple transformations from different embedding models' spaces to the same query embedding space can be learned using the same M2M training scheme.

2. No additional parameters are added to the user side querying system, preserving its computational resource efficiency.

3.3. Cross model compatible ensembles

Leveraging the above benefits and taking into account that gallery indexing is performed offline, we propose to register an image with multiple gallery models (Fig. 1b). We train a set of N gallery models, $\{\phi_g^i\}_{i=1}^N$, and a corresponding set of transformation models $\{T_g^i\}_{i=1}^N$. During indexing, each gallery image is processed by all gallery models $\{\phi_g^i\}_{i=1}^N$, producing a set of embeddings $\{e_g^i\}_{i=1}^N$ for each image[2]. We then apply $\{T_g^i\}_{i=1}^N$ to transform all embeddings to the embedding space of ϕ_q. Subsequently, we produce a single gallery embedding in ϕ_q's embedding space by averaging the transformed embeddings of each image:

$$\hat{e}_q = \frac{1}{N}\sum_{i=1}^{N} T_i(e_g^i). \tag{1}$$

where $\hat{e}_q$ is the final embedding of a single gallery image. We emphasize that this proposition does not increase the test-time latency, query model size and the number of comparisons during the querying process. Alternative approaches for combining multiple embedding spaces were considered. In this section we presented the best performing approach. Experiments of alternative approaches are presented in Section 4.4.

4. Experiments

We report experiments in two domains of image retrieval, *i.e.*, face recognition and product retrieval. We use face recognition as a particular case of image retrieval to evaluate and compare our method in various scenarios. We then report results of the most interesting experiments on the product retrieval task.

4.1. Face recognition experimental setup

We use three common network architectures used in the face recognition domain: ResNet18, ResNet100 [16] and MobileFaceNet (MBF) [7]. Each architecture was trained ten times on the VggFace2 dataset [6], creating a total of 30 models. For evaluation, we follow Shen *et al*. [36] and utilize the widely used IJB-C benchmark [29]. We adopt the two standard testing protocols for face recognition, namely, 1:1 verification and 1:N open-set search. In 1:1 face verification the algorithm decides whether a pair of templates belongs to the same person, where a template contains one or multiple images of a single person. In 1:N open-set search, each query template is compared to a gallery of templates. The algorithm then decides if and which gallery template matches the query template. For both tasks, the evaluation metric is true acceptance rate (TAR) at a specific false acceptance rate (FAR). For 1:1 verification we present TAR@FAR=0.01% results, and for 1:N search we present TAR@FAR=1%.

[2]In the compatible model update scenario, the gallery images can be discarded at this point.

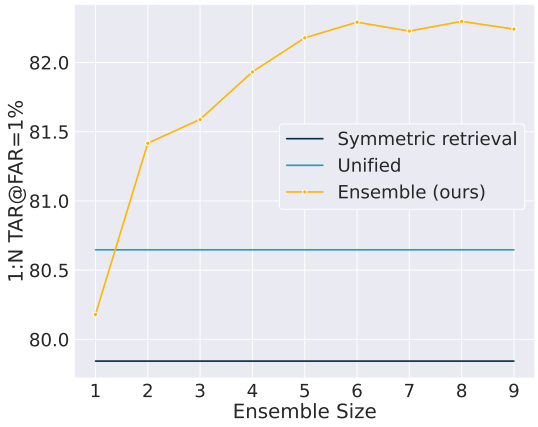
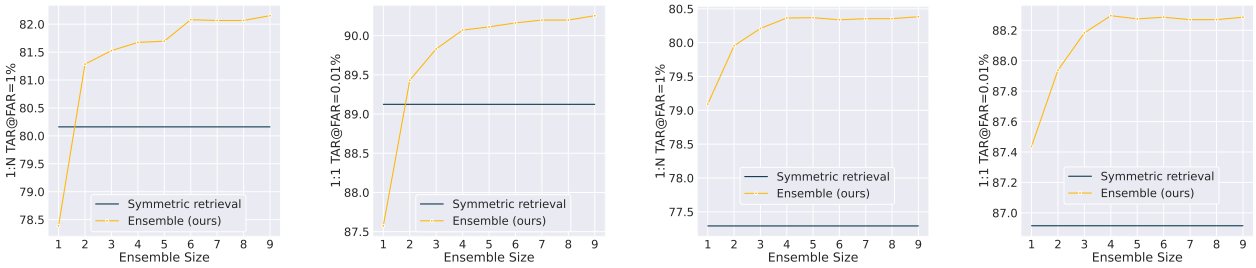


Figure 2. 1:N search for increasing ensemble sizes. Since the symmetric retrieval setting and Unified use only one gallery model they are visualized as lines.

	1:N TAR @FAR=1%	1:1 TAR @FAR=0.01%
Symmetric retrieval	$79.843_{\pm 0.50}$	$88.064_{\pm 0.27}$
Unified [47]	$80.647_{\pm 0.32}$	$88.439_{\pm 0.14}$
M2M	$80.476_{\pm 0.29}$	$88.368_{\pm 0.12}$
Ensemble of 2	81.417	89.068
Ensemble of 4	81.932	89.139
Ensemble of 9	**82.241**	**89.344**

Table 1. 1:N search and 1:1 verification TAR (in percentage) at fixed FAR values. Symmetric retrieval, $\phi_g = \phi_q$, was evaluated on ten ResNet18 models. Unified and M2M were evaluated on ten different ϕ_g ResNet18 models and the same ϕ_q ResNet18 model in each evaluation. Symmetric retrieval, unified and M2M TARs are shown as mean and std values of ten evaluations. Our method was evaluated using the same ϕ_q model as Unified and M2M and ensemble sizes of 2, 4 and 9 for comupting the gallery.



(a) ResNet100 (1:N) (b) ResNet100 (1:1) (c) MBF (1:N) (d) MBF (1:1)

Figure 3. 1:1 verification and 1:N search accuracy vs. increasing ensemble sizes for models based on ResNet100 and MobileFaceNet.

4.2. Accuracy gains by ensemble size

Fig. 2 and Table 1 provide results of our approach and the methods it builds upon. We first present results for the symmetric retrieval setting ($\phi_g = \phi_q$), *i.e.*, the standard use case. We then present results for the unified embedding space approach [47] (Unified) and the M2M approach described in 3.2.2. Note that M2M corresponds to an ensemble of size one. The results suggest that there is no significant difference between the two approaches, indicating that the M2M approach is a valid variant of the Unified approach for us to build upon. Finally we combine multiple M2M models and show the significant accuracy gains achieved by the ensemble approach. The ensembles of variable sizes were created incrementally by adding a single model each time. *I.e.*, ensemble of size $M+1$ contains exactly the same set of M2M models as for the ensemble of size M with the addition of one new M2M model. Note that knowledge distillation based approaches [5, 11, 36, 51], are upper-bounded by the performance of the stronger model (out of ϕ_g and ϕ_q) which in Table 1 corresponds to the result of symmetric retrieval. Fig. 3 shows that the trend of improved accuracy as a function of increased ensemble size is preserved across different architectures. Interestingly, in some cases the asymmetric transformation based approaches that do not use ensembles (Unified, M2M) receive a slightly better accuracy than the symmetric counterpart. This phenomenon was also reported by Wang *et al.* [47] and might happen because of the additional transformation model parameters or by implicit regularization.

4.3. Diversity of ensemble components

In this section we provide insights on how the diversity of the ensemble components impacts accuracy. We consider three levels of component diversity:

1. **Diversity of transformation models (D-T):** We use a single gallery model ϕ_g and train different transformation models mapping embedding vectors computed by ϕ_g to the query model's embedding space.

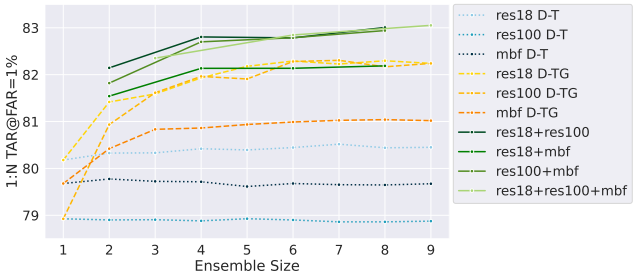
2. **D-T + diversity of gallery models (D-TG):** We use N different gallery models $\{\phi_g^i\}_{i=1}^N$ and train one transformation model per gallery model. Hence, both gallery models and transformation models are diverse.

3. **D-TG + diversity of gallery model architectures (D-TGA):** We further increase the diversity by allowing the gallery models to have different architectures.

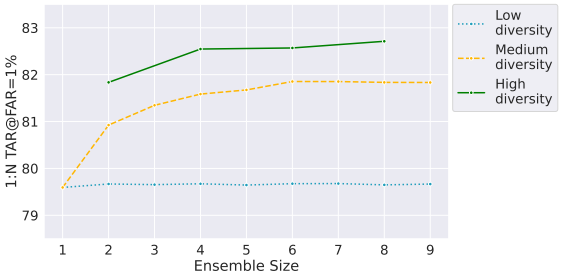
Fig. 4 shows the accuracy for various ensembles with different levels of component diversity. In all configurations the same ResNet18-based query model is used. The results imply that component diversity plays an important role in the ensemble's performance. Ensembles with high diversity consistently produce higher accuracy rates for the same number of transformation models than their less diverse counterparts. Even when the ensemble is comprised of only three gallery models with different architectures (1xResNet18, 1xResNet100 and 1xMBF), we observe better results than using an ensemble of eight gallery models with the same ResNet18 architecture. This result suggests that striving towards diverse ensembles is more beneficial than increasing the number of models in the ensemble. We provide the following hypotheses to explain this phenomenon: (1) each gallery model learns a different representation of the observed data [23, 30, 50] by focusing on slightly different features, (2) increased diversity in the gallery models results in more diverse representations, and (3) combining diverse representations of the data leads to better generalization that is translated to better performance. In other words, our method allows to capture more aspects of the data and combine them effectively.

4.4. Comparison to other ensemble alternatives

To further validate our ensemble design choices, we present alternative approaches to our proposed method of using independently trained transformation models, one for each gallery model, and averaging the transformed embeddings. Instead of the above, we learn a single combined transformation model that takes embeddings from all gallery models as inputs, and outputs a transformed embedding in the query model's embedding space. Since creating a single combined transformation model is not straightforward, we propose several variants:



(a) 1:N TAR@FAR=1%



(b) 1:N TAR@FAR=1% averaged by diversity level

Figure 4. **Diversity of ensemble components.** Dotted, dashed and solid lines correspond to D-T, D-TG and D-TGA results, respectively. For D-TGA the number of gallery models trained with each architecture is the same for each ensemble size, *e.g.*, res18+res100 of ensemble size 6 means three ResNet18 models and three ResNet100 models. (a) Shows the TAR for each ensemble type. (b) Shows the TAR for each diversity level by averaging the results of all ensemble types with the same diversity level and same ensemble size. The res18+res100+mbf D-TGA version was not used during averaging since this version ensemble sizes do not match the other D-TGA versions.

Ensemble version	1:N TAR@FAR=1%
End-to-end averaging	81.195%
Weighted end-to-end averaging	81.268%
Concatenation	81.410%
Ours	**81.932%**

Table 2. **Comparison to ensemble alternatives.** 1:N search TAR@FAR=1% results for each of the alternative ensemble variants. All experiments were conducted using the same query and same four gallery models. All models are based on ResNet18.

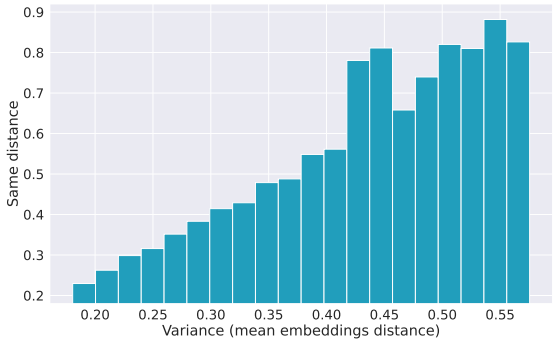


Figure 5. **Same-distance vs. variance.** Each bar in the plot represents a range of variance levels (x-axis). The y-axis corresponds to the mean same-distance (*i.e.*, the distance between the query embedding to a matching gallery embedding) of all matches inside the range. The bars' width is 0.248 corresponding to 20 bins spanning from the lowest variance to the highest variance observed.

1. **End-to-end averaging:** We train a transformation model from each gallery model to the query embedding space simultaneously. Specifically, instead of training each transformation model independently, we train all transformation models together and perform the averaging during training.

2. **Weighted end-to-end averaging:** In a similar manner, we jointly train all transformation models except that we use weighted averaging instead of simple averaging. Each transformation model outputs a single scalar value, in addition to the embedding output, that is used for weighted averaging.

3. **Concatenation:** All gallery embeddings are first concatenated and inserted into a linear layer that reduces the dimensions. The output of the linear layer is then transformed by a single transformation model to the embedding space of the query model.

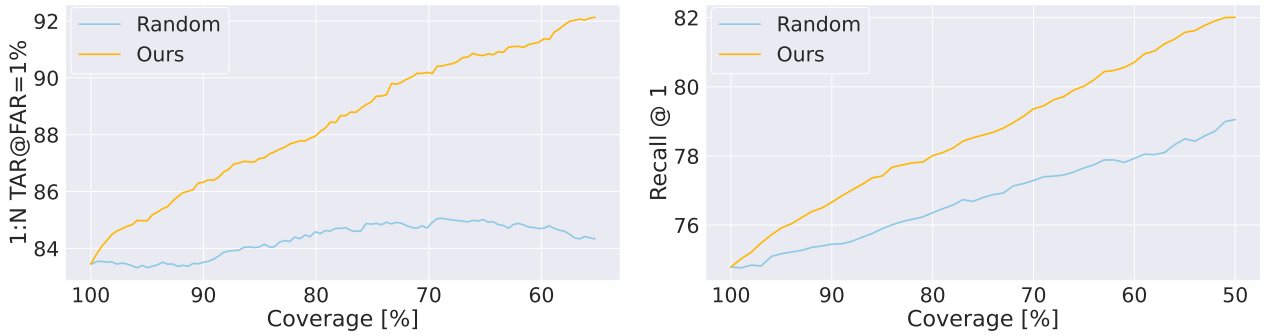
In this experiment, for all variants, and for our method, we used the same five pre-trained ResNet18 models. One for the query model, and four for the gallery models.

Table 2 presents the comparison between our ensemble approach and the combined transformation variants. The results indicate that our approach performers considerably better then the other variants. Furthermore, this implies that explicitly optimizing the ensemble performs worse than averaging independently trained models.

4.5. Uncertainty

It was previously observed that ensembles of models can be used to evaluate the uncertainty of model predictions[3] [27]. Typically, uncertainty is evaluated by calculating the variance between the predictions of the different models in the ensemble. Thus, if the predictions of the different models are inconsistent, the uncertainty rises, and the models' predictions are considered less reliable. In this section, we analyze whether the ensemble proposed in our approach can be used to reliably measure uncertainty. To evaluate the model uncertainty for a given gallery image, we measure the variance of the transformed embeddings of the gallery models in the query embedding space. The variance is calculated by measuring the mean distance between every pair of embeddings.

[3]This type of uncertainty is typically known as *epistemic* uncertainty or *model* uncertainty.



(a) IJC-B 1:N TAR@FAR=1% vs. coverage.

(b) SOP Recall@1 vs. coverage.

Figure 6. **Risk-coverage curve.** We show the accuracy increase (risk) gained by filtering out (coverage) gallery embeddings using two protocols for both the IJC-B (a) and the SOP (b) datasets. Random: removing random embeddings (as a naïve baseline). Ours: rejecting embeddings whose variance is above a gradually decreasing threshold. Note that, for Recall@1, smaller gallery sets are easier for matching, therefore, *Random* shows improvement, but not to the same extent as ours.

When querying, we consider the gallery embedding as *easy to retrieve* if the distance between the gallery embedding to the query embedding of the same class is low (*same-distance*). Fig. 5 suggests that the same-distance and the measured variance between the transformed gallery embeddings are indeed correlated. This result further implies that, upon registering a new gallery image, the variance can be used as a quality-score for accepting new gallery images, *i.e.*, if the variance is larger than some threshold the image will be rejected. In other words, if the gallery models' embeddings are far from one another, then the gallery embedding might not be easy to retrieve, thus we may choose to re-register with a new gallery image.

We follow Geifman *et al*. [12] and measure the quality of uncertainty estimation by reporting the risk-coverage curve. The curve represents the accuracy we can achieve (risk) as a function of the percent of gallery images we register (coverage). The curve is constructed as following: as we decrease the threshold for uncertainty, more images are rejected. For every point on the curve we provide the accuracy of the system given an uncertainty threshold. The risk-coverage curve is presented in Fig. 6a, and demonstrates that even rejecting as few as 10% of the gallery embeddings, reduces the error by 17.4%, while a 20% rejecting rate leads to a 27.2% error reduction. Fig. 6a shows that our filtering method is consistently better than naïvely removing gallery images, meaning that the improved accuracy is not the sheer result of reduction in the number of gallery images, rather in reduction of images that are prone to error. This suggest that the variance of gallery embeddings could be used as an embedding quality metric. In the supplementary, we provide examples for gallery images with varying uncertainty levels.

4.6. Compatible query model update

In this section we evaluate the scenario where an improved query model was developed, replacing an older one and backfilling is not possible. This simulates a system that does not retain data (except of training data). We simulate the scenario as follows:

1. **Before update:** The system is composed of N gallery models and one query model. All models were trained on 50% of the VggFace2 dataset, that represents the available training dataset at some point during the lifespan of the system. At this stage the gallery set is mapped to the embedding spaces of all gallery models ($\{\phi_{g,i}^{50\%}\}_{i=1}^{N}$) and the gallery images are discarded. We transform the embeddings from the embedding spaces of $\{\phi_{g,i}^{50\%}\}_{i=1}^{N}$ to the embedding space of the query model ($\phi_q^{50\%}$), using our ensemble approach.

2. **After update:** At some point additional training data was made available[4] and was used to train an improved query model. We simulate this by training a new query model on 100% of the VggFace2 dataset ($\phi_q^{100\%}$). This time the corresponding embedding spaces of $\{\phi_{g,i}^{50\%}\}_{i=1}^{N}$ are transformed to the embedding space of $\phi_q^{100\%}$ using new transformation models.

The above scenario represents an extreme case, where the performance gap between the old models and the newer one (trained on 50% and 100% of the data respectively) is significant. This can be seen in Table 3 where the 1:N search TAR@FAR=1% of $\phi_q^{50\%}$ is 71.01% vs. 79.93% of $\phi_q^{100\%}$.

Table 3 demonstrates the performance gain of updating the query model. Interestingly, the performance increases drastically compared to the gap before the update (for ensemble size of four, the 1:N search TAR@FAR=1% increases from 73.98% to 78.96%), despite using only low-accuracy-models for gallery indexing. Furthermore, Table 3 shows that a larger ensemble size corresponds with improved performance even when the query model is much stronger than the gallery models used for indexing.

[4] Additional training data can be acquired by conducting a dedicated data collection effort for example.

	1:N TAR	1:1 TAR
Independent models in the symmetric setting		
Gallery models: $\{\phi_{g,i}^{50\%}\}_{i=1}^{4}$	$71.96_{\pm 0.3}$	$83.24_{\pm 0.17}$
Query model: $\phi_q^{50\%}$	71.01	83.00
Query model: $\phi_q^{100\%}$ †	79.93	88.28
Ensemble of size 2 ($\{\phi_{g,i}^{50\%}\}_{i=1}^{2}$)		
Before update ($\phi_q^{50\%}$)	72.90	83.78
After update ($\phi_q^{100\%}$)	78.21	86.94
Ensemble of size 4 ($\{\phi_{g,i}^{50\%}\}_{i=1}^{4}$)		
Before update ($\phi_q^{50\%}$)	73.98	84.62
After update ($\phi_q^{100\%}$)	**78.96**	**87.39**

Table 3. **Compatible query model update.** First three rows show the accuracy of the standard symmetric setting of all models used in the experiment (the accuracy of the gallery models is presented as the average of four models). "Before update", shows the results where querying is done by the old query model, $\phi_q^{50\%}$. "After update", shows the results for using the updated query model, $\phi_q^{100\%}$, for querying. In both, before and after update, the same gallery models are used for indexing. †: Note that in the no-backfilling scenario using $\phi_q^{100\%}$ in a symmetric setting is not possible.

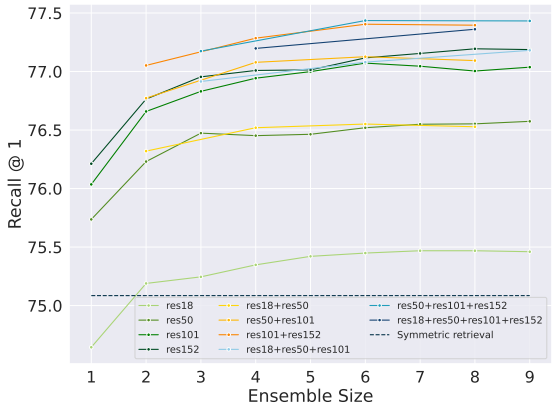
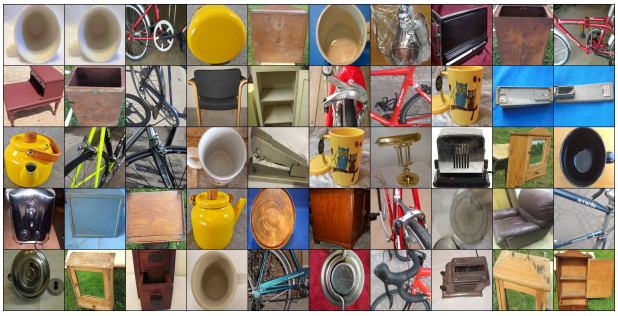


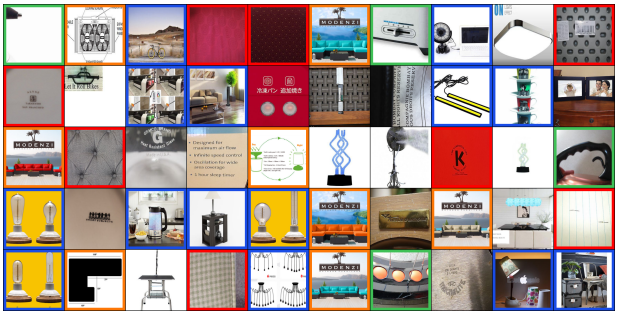
Figure 7. **Recall@1 vs. ensemble size.** "Symmetric retrieval" denotes the accuracy of the ResNet18 query model.

4.7. Product retrieval experiments

To further demonstrate the generality of our approach we conducted experiments on the Stanford online products (SOP) dataset [40]. The dataset contains 22.6K classes with 120K product images, where 11.3K classes (59.5K images) are used for training and the remaining 11.3K classes (60.5K images) are used for testing. We followed the generic protocol proposed by [53] for training query and gallery models. We used the Recall@1 [20] metric for evaluating the retrieval performance. In all the experiments we used the same ResNet18 model for querying. Fig. 7 shows that the benefits from using an ensemble generalize to the domain of product retrieval. Additionally, Fig. 7 repeats



(a) 50 images with the lowest uncertainty.



(b) 50 images with the highest uncertainty.

Figure 8. **Examples of SOP gallery images.** (a) and (b) show the images with the lowest and highest variance values respectively. Note the various factors resulting in high ambiguity within the ensemble: extreme zoom, partial views, multiple objects, non-natural images (*e.g.*, sketched, rendered). [can be zoomed-in].

the previously observed trend where diverse ensembles are generally preferable. Fig. 6b demonstrates the benefit of filtering out gallery embeddings with high uncertainty in this domain. To conform with the previous settings for which we calculated the risk-coverage curve, for each test label, we used a single image as query and a single image for the gallery set. Fig. 8 shows the images with lowest and highest variance in the gallery set. Low variance images generally include a single identifiable object, while high variance images suffer from multiple ambiguity factors.

5. Conclusions

We propose a novel embedding transformation based ensemble framework for asymmetric image retrieval. We show that embedding transformations can be leveraged for creating a non-trivial ensemble of diverse gallery models, significantly increasing the retrieval accuracy without increasing the computational cost of querying. We compared several methods for combining multiple embedding spaces and found that training the transformation models independently lead to the best performance. Additionally, we utilize the diversity between multiple transformed embeddings to estimate the uncertainty of gallery images. We propose to reject gallery images based on their uncertainty to further improve our system's accuracy.

References

[1] Artem Babenko and Victor Lempitsky. Aggregating deep convolutional features for image retrieval. *arXiv preprint arXiv:1510.07493*, 2015. 2

[2] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 584–599. Springer, 2014. 1

[3] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996. 3

[4] Jane Bromley, James W. Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. Signature verification using A "siamese" time delay neural network. *Int. J. Pattern Recognit. Artif. Intell.*, 7(4):669–688, 1993. 2

[5] Mateusz Budnik and Yannis Avrithis. Asymmetric metric learning for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8228–8238, June 2021. 1, 2, 3, 5

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74, 2018. 4

[7] Ken Chen, Yichao Wu, Haoyu Qin, Ding Liang, Xuebo Liu, and Junjie Yan. R3 adversarial network for cross model face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9868–9876, 2019. 2, 3, 4

[8] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pages 539–546. IEEE Computer Society, 2005. 2

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 1, 2

[10] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers of Computer Science*, 14(2):241–258, 2020. 3

[11] Rahul Duggal, Hao Zhou, Shuo Yang, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Compatibility-aware heterogeneous visual search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10723–10732, 2021. 1, 2, 3, 5

[12] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017. 7

[13] Pablo M Granitto, Pablo F Verdes, and H Alejandro Ceccatto. Neural network ensembles: evaluation of aggregation algorithms. *Artificial Intelligence*, 163(2):139–162, 2005. 3

[14] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 2

[15] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 4

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[18] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In Aasa Feragen, Marcello Pelillo, and Marco Loog, editors, *Similarity-Based Pattern Recognition - Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015, Proceedings*, volume 9370 of *Lecture Notes in Computer Science*, pages 84–92. Springer, 2015. 2

[19] Florian Jaeckle, Fartash Faghri, Ali Farhadi, Oncel Tuzel, and Hadi Pouransari. Fastfill: Efficient compatible model update. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[20] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:117–128, 2011. 8

[21] Cheng Ju, Aurélien Bibaut, and Mark van der Laan. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*, 45(15):2800–2818, 2018. 3

[22] John Kolen and Jordan Pollack. Back propagation is sensitive to initial conditions. *Advances in neural information processing systems*, 3, 1990. 3

[23] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019. 5

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012. 3

[25] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7, 1994. 3

[26] Ludmila I Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014. 3

[27] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty

estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016. 6

[28] Hui Li, Xuesong Wang, and Shifei Ding. Research and development of neural network ensembles: a survey. *Artificial Intelligence Review*, 49(4):455–479, 2018. 3

[29] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165. IEEE, 2018. 4

[30] Johannes Mehrer, Courtney J. Spoerer, Nikolaus Kriegeskorte, and T. Kietzmann. Individual differences among deep neural network models. *bioRxiv*, 2020. 5

[31] Qiang Meng, Chixiang Zhang, Xiaoqiang Xu, and Feng Zhou. Learning compatible embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9939–9948, 2021. 3

[32] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018. 1

[33] Ali S Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. Visual instance retrieval with deep convolutional networks. *ITE Transactions on Media Technology and Applications*, 4(3):251–258, 2016. 2

[34] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018. 3

[35] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2

[36] Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6368–6377, 2020. 2, 3, 4, 5

[37] Oriane Siméoni, Yannis Avrithis, and Ondrej Chum. Local features and visual words emerge in activations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11651–11660, 2019. 1

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 3

[40] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4004–4012, 2015. 8

[41] Yi Sun. *Deep learning face representation by joint identification-verification*. The Chinese University of Hong Kong (Hong Kong), 2015. 2

[42] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6397–6406. Computer Vision Foundation / IEEE, 2020. 2

[43] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 3

[44] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 2

[45] Giorgos Tolias, Tomas Jenicek, and Ondřej Chum. Learning and aggregating deep local descriptors for instance-level recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 460–477. Springer, 2020. 1

[46] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104:154–171, 2013. 2

[47] Chien-Yi Wang, Ya-Liang Chang, Shang-Ta Yang, Dong Chen, and Shang-Hong Lai. Unified representation learning for cross model compatibility. *arXiv preprint arXiv:2008.04821*, 2020. 2, 3, 4, 5

[48] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Process. Lett.*, 25(7):926–930, 2018. 2

[49] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017. 1, 2

[50] Liwei Wang, Lunjia Hu, Jia-Yuan Gu, Yue Kris Wu, Zhiqiang Hu, Kun He, and John E. Hopcroft. Towards understanding learning representations: To what extent do different neural networks learn the same representation. In *NeurIPS*, 2018. 5

[51] Hui Wu, Min Wang, Wengang Zhou, Houqiang Li, and Qi Tian. Contextual similarity distillation for asymmetric image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9489–9498, June 2022. 1, 2, 3, 5

[52] Lingxi Xie, Richang Hong, Bo Zhang, and Qi Tian. Image classification and retrieval are one. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 3–10, 2015. 2

[53] Andrew Zhai and Hao-Yu Wu. Classification is a strong baseline for deep metric learning. In *British Machine Vision Conference*, 2018. 8

[54] Binjie Zhang, Yixiao Ge, Yantao Shen, Shupeng Su, Chun Yuan, Xuyuan Xu, Yexin Wang, and Ying Shan. To-

wards universal backward-compatible representation learning. *arXiv preprint arXiv:2203.01583*, 2022. 3

[55] Cha Zhang and Yunqian Ma. *Ensemble machine learning: methods and applications*. Springer, 2012. 3

[56] Liang Zheng, Shengjin Wang, Lu Tian, Fei He, Ziqiong Liu, and Qi Tian. Query-adaptive late fusion for image search and person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1741–1750, 2015. 2

[57] Yifei Zhou, Zilu Li, Abhinav Shrivastava, Hengshuang Zhao, Antonio Torralba, Taipeng Tian, and Ser-Nam Lim. BT^2: Backward-compatible training with basis transformation. *arXiv preprint arXiv:2211.03989*, 2022. 2