EXPLOITING DISTRIBUTION CONSTRAINTS FOR SCALABLE AND EFFICIENT IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Image retrieval is crucial in robotics and computer vision, with downstream applications in robot place recognition and vision-based product recommendations. Modern retrieval systems face two key challenges: scalability and efficiency. State-of-the-art image retrieval systems train specific neural networks for each dataset, an approach that lacks scalability. Furthermore, since retrieval speed is directly proportional to embedding size, existing systems that use large embeddings lack efficiency. To tackle scalability, recent works propose using off-the-shelf foundation models. However, these models, though applicable across datasets, fall short in achieving performance comparable to that of dataset-specific models. Our key observation is that, while foundation models capture necessary subtleties for effective retrieval, the underlying distribution of their embedding space can negatively impact cosine similarity searches. We introduce Autoencoders with Strong Variance Constraints (`AE-SVC`), which, when used for projection, significantly improves the performance of foundation models. We provide an in-depth theoretical analysis of `AE-SVC`. Addressing efficiency, we introduce Single-shot Similarity Space Distillation ($(SS)_2D$), a novel approach to learn embeddings with adaptive sizes that offers a better trade-off between size and performance. We conducted extensive experiments on four retrieval datasets, including Stanford Online Products (SoP) and Pittsburgh30k, using four different off-the-shelf foundation models, including DinoV2 and CLIP. `AE-SVC` demonstrates up to a 16% improvement in retrieval performance, while $(SS)_2D$ shows a further 10% improvement for smaller embedding sizes.

1 INTRODUCTION

Image retrieval involves finding the closest match of a given image (query) in a vast database of images (often called the reference set). It has numerous applications, from product recommendations in e-commerce to place recognition in robotics. In general, state-of-the-art (SOTA) image retrieval systems are dataset-specific. They are trained in a contrastive manner on a training split of the data, typically hand-labeled for positive and negative pairs for each query. Imagine a scenario where a server has a vast reference database of fashion clothing images and receives query images from users to find the closest matches in the reference set. Creating a separate split and hand-labeling positive and negative pairs to train a dataset-specific model is infeasible. A more scalable solution is to use off-the-shelf feature extractors like Dino (Caron et al., 2021), DinoV2 (Oquab et al., 2024), CLIP (Radford et al., 2021), and ViT (Dosovitskiy et al., 2021). However, the performance of off-the-shelf feature extractors is generally lower compared to dataset-specific models. **Q1 (Scalability): Is there a way to enhance the performance of these off-the-shelf models in a completely unsupervised way by using just the reference set?**

Another key problem in retrieval is that the retrieval speed is directly proportional to the size of the embeddings. For an embedding of size d and a reference set with N image vectors, the retrieval speed for a single query is $O(d \times N)$. This retrieval time becomes significant as the size of the reference set (N) increases, with modern retrieval systems containing millions of images. Additionally, the embedding size directly affects the storage space required for the reference set and the communication bandwidth needed to send the query vector to the server in distributed image retrieval systems, which are common in product recommendations and robotic place recognition. Standard dimensionality reduction techniques, such as Principal Component Analysis (PCA), aim to preserve

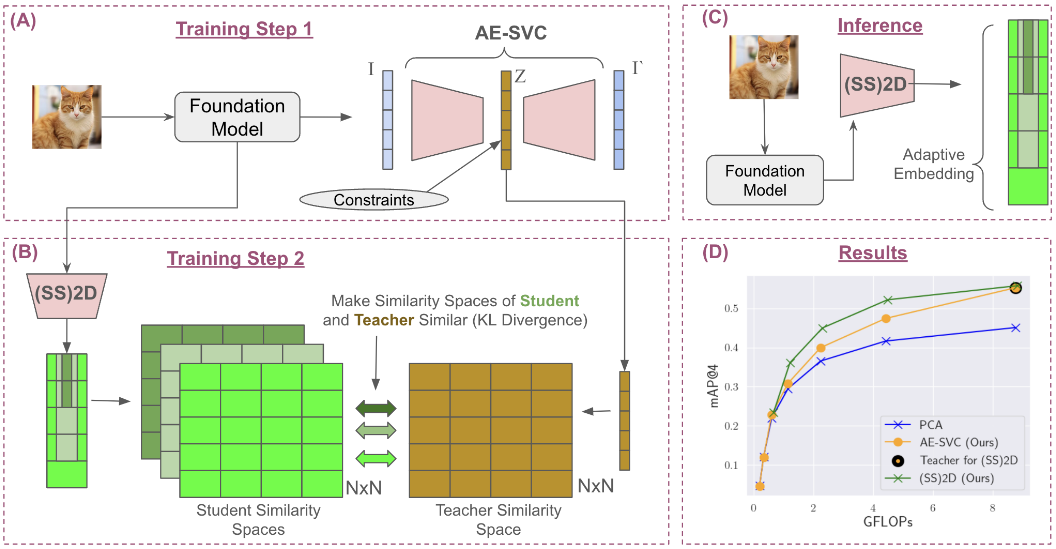


Figure 1: **Two-step pipeline for the proposed approach. (A)** AE−SVC (discussed in Sec. 3.1) trains an autoencoder with our constraints to improve foundation model embeddings. **(B)** (SS)$_2$D (discussed in Sec. 3.2) uses the improved embeddings from AE−SVC to learn adaptive embeddings for improved retrieval at any embedding size. **(C)** Once trained, (SS)$_2$D can be directly applied to foundation model embeddings to generate adaptive embeddings for improved retrieval. **(D)** AE−SVC (orange) boosts performance significantly, while (SS)$_2$D (green) further enhances results with smaller embeddings. Dino (blue) achieves optimal performance at 9 GLOPs, whereas (SS)$_2$D on top of AE−SVC achieves similar performance at only 2.5 GLOPs.

information, while Autoencoders (AEs) (Hinton & Salakhutdinov, 2006) and Variational Autoencoders (VAEs) (Kingma & Welling, 2013) focus on reconstruction quality. However, these methods are not optimized for retrieval tasks. Additionally, learning-based approaches like AEs and VAEs require separate training for each embedding size, which is impractical given varying compute, space, and communication constraints. **Q2 (Efficiency): Is there an effective unsupervised dimensionality reduction method that strongly preserves the similarity structure of the full embeddings, and is adaptive, i.e., does not need to be trained for each dimension separately?**

Contributions: To address **Q1 (Scalability)**, we propose Autoencoders with Strong Variance Constraints (AE−SVC). Our primary idea is that, while foundation models capture the necessary details for effective retrieval, the underlying distribution of their embeddings can negatively impact cosine similarity searches. AE−SVC trains an autoencoder while enforcing three constraints on the latent space: an orthogonality constraint, a mean centering constraint, and a unit variance constraint. We empirically show and mathematically prove that these constraints cause a shift in the cosine similarity distribution, making it more discriminative. These constraints not only lead to more effective dimensionality reduction but also **outperform the complete embeddings of foundation models**. We discuss the motivation and technical details of AE−SVC in section 3.1. We also provide an in-depth mathematical analysis of AE−SVC in section 4. To address **Q2 (Efficiency)**, we propose Single Shot Similarity Space Distillation ((SS)$_2$D). (SS)$_2$D aims at reducing embeddings to smaller ones while preserving their similarity relationships. The embedding learned with (SS)$_2$D is adaptive, *i.e.*, smaller segments of the output embedding also perform well in retrieval tasks. In summary, our approach consists of two steps (as shown in Fig. 1): first, we use AE−SVC to enhance the baseline performance of foundation models. Then, we leverage the improved embeddings as a teacher for efficient dimensionality reduction through (SS)$_2$D.

Fig. 1(D) also shows the retrieval performance versus retrieval speed in giga-FLOPs (floating operations) on the Pittsburgh30K dataset (Torii et al., 2013) using the Dino model (Caron et al., 2021). Note that retrieval speed is a function of the dimension d and the reference set size N, $O(d \times N)$, with $N = 17,000$ for Pittsburgh30K. The performance of PCA (blue) represents the default off-the-shelf Dino model. We can see that AE−SVC (orange) significantly enhances the performance of the off-the-shelf Dino model (blue). Notice that it outperforms the complete Dino embedding. Applying (SS)$_2$D on top of AE−SVC further improves retrieval performance at smaller embedding sizes (green). Section 5 experimentally shows the advantages of the proposed approaches in detail on four different datasets using different foundation models.

2 RELATED WORK

Deep Metric Learning: Image retrieval is modeled as a Deep Metric Learning (DML) problem (Wang et al., 2017). DML focuses on learning latent representations where similar items are close and dissimilar items are far apart. Various techniques have been proposed (Oh Song et al., 2016; Schroff et al., 2015; Ustinova & Lempitsky, 2016; Sohn, 2016; Wang et al., 2019) to learn meaningful latent representations for effective retrieval, including contrastive loss (Hadsell et al., 2006) and triplet loss (Schroff et al., 2015). More recent works use advanced N-pair loss (Sohn, 2016) or multi-similarity loss (Wang et al., 2019). Analogous to image retrieval, Visual Place Recognition (VPR) is also modeled as a DML problem (Garg et al., 2021). Various training objectives have been proposed for VPR (Ge et al., 2020; Xiao et al., 2023; Leyva-Vallina et al., 2023; Berton et al., 2022) to learn latent representations for effective retrieval. However, all of these approaches are dataset-specific, meaning they train specific neural networks for each dataset.

Foundation Models in Image Retrieval: The rise of foundation models, such as DINOv2 (Oquab et al., 2024), CLIP (Radford et al., 2021), ViT (Dosovitskiy et al., 2021), has allowed recent works to explore the applicability of these off-the-shelf models in image retrieval. As shown in (Keetha et al., 2023; Lu et al., 2024), off-the-shelf models, in some cases, can have comparable retrieval performance to dataset-specific models. However, these approaches either use the off-the-shelf models as-is (Keetha et al., 2023) or adapt them using a labeled held-out dataset (Lu et al., 2024). In stark contrast, our approach, `AE-SVC`, exploits the inherent properties of the distribution to improve the performance of these off-the-shelf models without requiring any labeled data.

Distribution Constraints in Semi-supervised Representation Learning: Semi-supervised learning has seen significant advances with methods that leverage self-supervised objectives to learn robust representations. A notable approach is Barlow Twins, which employs an additional cross-correlation loss to align representations while minimizing redundancy across dimensions, promoting feature decorrelation (Zbontar et al., 2021). Extending this idea, VICReg introduces an additional variance constraint that ensures sufficient spread in the learned features, preventing collapse and further enhancing representation quality (Bardes et al., 2022). However, to date, no work has specifically explored the impact of such constraints in the context of autoencoders for retrieval tasks. In this paper, we are the first to address this gap, providing a comprehensive theoretical analysis of these constraints and their impact on retrieval.

Similarity Preserving Knowledge Distillation: Similarity preserving knowledge distillation (SPKD) focuses on transferring knowledge from a teacher model to a student model, similar to standard distillation (Hinton et al., 2015; Buciluǎ et al., 2006; Ba & Caruana, 2014; Zagoruyko & Komodakis, 2017; Zhang et al., 2019; Passalis & Tefas, 2018), while maintaining the structure of similarity of the data in the feature space. One of the pioneering works in this area is the FitNets approach, which uses intermediate representations to guide the student model (Romero et al., 2014). The attention transfer method extends this by aligning the attention maps of the student and teacher models (Zagoruyko & Komodakis, 2017). More recent advancements include (Tung & Mori, 2019), which explicitly preserves pairwise similarities between data points in the feature space during distillation. Contrastive Representation Distillation (CRD) further improves performance by using contrastive loss to maximize mutual information between teacher and student representations (Tian et al., 2020). Additionally, the Relational Knowledge Distillation (RKD) approach emphasizes the importance of preserving the relational knowledge among data points (Park et al., 2019). All of these works focus on training a lightweight student network to mimic a heavy teacher network. We are the first to explore this idea in a dimensionality reduction setting for image retrieval.

Adaptive Feature Embeddings: A core requirement for dimensionality reduction in retrieval systems is adaptive embeddings due to varying compute constraints. Previous works on adaptive feature embeddings have created efficient and adaptive embedding spaces for compression (Li et al., 2023) and image classification (Kusupati et al., 2022), while ours focuses on image retrieval. The key idea is to use only one neural network model to output a feature embedding ensuring that smaller sub-chunks of the embedding also perform well in retrieval tasks.

3 METHODOLOGY

We now formally define the image retrieval problem:

Suppose we have a query set Q of M query image feature vectors and a reference set R of N reference image feature vectors, where each vector is d-dimensional. Given $q_j \in \mathbb{R}^d$ randomly drawn from Q, the task is to identify the feature vector r_i in R that has the smallest distance or highest similarity with q_j. This can be mathematically expressed as:

$$r_i = \arg\min_{r \in R} \operatorname{dis}(q_j, r),$$

with the distance dis generally based on cosine similarity, i.e., $\operatorname{dis}(q_j, r_i) = 1 - \cos(q_j, r_i)$, where cos denotes the cosine similarity.

We will now describe our approach in detail. First, we will explain the step-by-step procedure for AE−SVC. Next, we will present an in-depth explanation of (SS)$_2$D. Note that both AE−SVC and (SS)$_2$D utilize only the reference set R during the training time. The query set Q is assumed to be unavailable to the user during training. A standard assumption when doing dimensionality reduction in image retrieval settings (Keetha et al., 2023).

3.1 AUTOENCODERS WITH STRONG VARIANCE CONSTRAINTS (AE−SVC)

Motivation: Our primary motivation for AE−SVC is that, while foundation models capture necessary subtleties for effective retrieval, the underlying distribution of their embedding space can negatively impact cosine similarity searches. In Sec. 4, we discuss in mathematical detail that the discriminative power of the retrieval task is maximized when the variance of the cosine similarity distribution is minimized. AE−SVC introduces three constraints on the latent space: an orthogonality constraint, a mean centering constraint, and a unit variance constraint. In Sec. 4, we prove that these constraints on the latent space minimize the variance of the cosine similarity distribution. We now provide a more intuitive explanation of why these constraints are helpful. Consider the example illustrated in Fig. 2. We have a reference set comprising four categories of clothing: men's tank-tops, men's half-shirts, women's tank-tops, and women's half-shirts. The task is to match the query with the correct clothing type. After applying dimensionality reduction (like a standard PCA) to the reference set, we obtain two orthogonal dimensions: a person dimension (with high variance) and a shirt shape dimension (with low variance). If similarity is overly dependent on the high-variance dimension, as shown mathematically in Sec. 4, retrieval performance suffers. For instance, a query for a white man in a tank-top might incorrectly match with a white man in a half-shirt due to the dominant person dimension. Ideally, we need a space where both the person and shirt dimensions equally influence the distinction between men and women, as well as between half-shirts and tank-tops. The goal of AE−SVC is to learn a projection from the reference set R such that the latent dimensions are orthogonal and have equal variance.

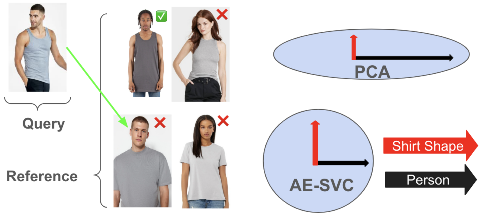


Figure 2: **In standard dimensionality reduction (say PCA), cosine similarity is disproportionately influenced by high-variance dimensions, leading to poor retrieval.** Given a task to match the query with the correct clothing type. A query image of a white man in a tank-top may be incorrectly matched to a white man in a half-shirt due to the dominant person dimension. Ideally, both orthogonal dimensions should have an equal influence on cosine similarity.

Approach:. AE−SVC (as shown in Fig. 1(**A**)) takes an input embedding $I \in \mathbb{R}^d$ coming from the foundation model, encodes it to a latent representation $z \in \mathbb{R}^d$, and produces a reconstruction $I' \in \mathbb{R}^d$. We have three additional losses to enforce the constraints mentioned above: an orthogonality constraint, a mean centering constraint, and a unit variance constraint. Note that we can choose any size for the latent representation z. We show the results for multiple sizes of z. However, since AE−SVC is used to guide (SS)$_2$D later, we will use the best performing embedding, i.e., $z \in \mathbb{R}^d$, for this guidance.

Reconstruction Loss: We minimize the Mean Squared Error (MSE) between the input embedding I and the reconstruction I':

$$\mathcal{L}_{\text{rec}} = \frac{1}{n}\sum_{i=1}^{n} \|I_i - I'_i\|_2^2, \tag{1}$$

where n is the number of samples.

Covariance Loss: To promote orthogonality, we penalize off-diagonal terms of the covariance matrix:

$$\mathcal{L}_{\text{cov}} = \left\| \frac{1}{n}(Z-\mu)^\top (Z-\mu) - \mathbb{I} \right\|_F^2, \tag{2}$$

where $Z \in \mathbb{R}^{n\times d}$ is the matrix of latent representations, $\mu \in \mathbb{R}^d$ is the mean of Z, and $\mathbb{I}$ is the identity matrix.

Variance Loss: We constrain the variance of each latent dimension to be one:

$$\mathcal{L}_{\text{var}} = \frac{1}{d}\sum_{i=1}^{d} \left(\text{Var}(z^i) - 1\right)^2, \tag{3}$$

where z^i is the i-th latent dimension of z and $\text{Var}(z^i)$ is the variance of the i-th latent dimension across all n samples.

Mean-Centering Loss: To ensure the latent space is centered around zero, we minimize the mean of the latent dimensions:

$$\mathcal{L}_{\text{mean}} = \frac{1}{d}\sum_{i=1}^{d} (\mu^i)^2, \tag{4}$$

where μ^i is the mean of the i-th latent dimension across all n samples.

Total Loss: The final objective combines all four losses:

$$\mathcal{L}_{\text{AE-SVC}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{cov}}\mathcal{L}_{\text{cov}} + \lambda_{\text{var}}\mathcal{L}_{\text{var}} + \lambda_{\text{mean}}\mathcal{L}_{\text{mean}} \tag{5}$$

where $\lambda_{\text{rec}}, \lambda_{\text{cov}}, \lambda_{\text{var}}$, and λ_{mean} are hyperparameters discussed in A.6. By optimizing this loss, AE-SVC learns a latent space that improves retrieval performance through accurate reconstruction, decorrelated and normalized feature dimensions, and a zero-centered mean. The latent representation, $z \in \mathbb{R}^d$, is then used to guide the training of `(SS)`$_2$`D`.

3.2 SINGLE SHOT SIMILARITY SPACE DISTILLATION (`(SS)`$_2$`D`)

Motivation: As previously mentioned, embedding size significantly impacts retrieval speed. For an embedding of size d and a reference set with N image vectors, the retrieval speed for a single query is $O(d \times N)$. Therefore, it is crucial to reduce the embedding size while maintaining performance. Standard dimensionality reduction techniques, such as PCA, Autoencoders (AEs), and Variational Autoencoders (VAEs), are not optimized for retrieval. Additionally, AEs and VAEs require separate training for each embedding size, which is impractical given varying computational constraints. Therefore, there is a need for adaptive embeddings that maintain retrieval performance and are flexible, ensuring that smaller sub-chunks of the embedding also perform well in retrieval tasks.

Approach: Previous works have employed Similarity-Preserving Knowledge Distillation (SPKD) both in hidden layer activations (Smith et al., 2023) and output layers (Wu et al., 2022). SPKD trains a student network to ensure that input pairs producing similar (or dissimilar) activations in the teacher network yield similar (or dissimilar) activations in the student network. However, these studies have not explored similarity-preserving distillation for efficient dimensionality reduction. Our approach is conceptually similar to SPKD, but rather than focusing on network distillation (teacher to student), we perform embedding-level distillation, reducing a large embedding to a smaller one while preserving similarity relationships. Using SPKD directly for dimensionality reduction requires training separate networks for each dimension size. Previous works on adaptive embeddings have utilized random projections in a distributed compression setting (Li et al., 2023) or optimized multi-class classification loss for each nested dimension (Kusupati et al., 2022). Following the approach in (Kusupati et al., 2022), we optimize a similarity-preserving loss for each nested dimension. By combining a similarity-preserving loss with an adaptive embedding training pipeline, we introduce our approach, Single Shot Similarity Space Distillation, or `(SS)`$_2$`D`. Fig. 1**(B)** illustrates the central idea of `(SS)`$_2$`D` which we will now describe in detail.

Let $z \in \mathbb{R}^d$ be the latent representation of the reference set obtained via `AE-SVC`. Our goal is to reduce the dimensionality of z so that the performance of the smaller embedding approximates that of the full embedding. In essence, the complete embedding z serves as a guide for creating smaller embeddings. Although `(SS)`$_2$`D`could theoretically be applied directly to the foundation model embedding I, the latent representation z is shown to perform better in retrieval tasks, thus providing stronger guidance.

Consider a set $\mathcal{M} \subset [d]$ of embedding sizes and a neural network $\mathcal{F}(\cdot;\theta)$ parameterized by θ. We want to learn a projection $\hat{z} = \mathcal{F}(z;\theta)$ such that each of the first m dimensions, for $m \in \mathcal{M}$, of the learned embedding vector $\hat{z}^{1:m} \in \mathbb{R}^m$ independently preserves the retrieval performance of z. To learn the function $\mathcal{F}$, we first compute a cosine similarity matrix $\mathrm{C} \in \mathbb{R}^{N\times N}$ by calculating the cosine similarities of every z with every other z in the reference set R. We refer to this matrix as the cosine similarity space of the teacher Z. Let C_i denote the i^{th} row of C.

Next, for each $m \in \mathcal{M}$, we define a student similarity space $\tilde{\mathrm{C}}^m \in \mathbb{R}^{N\times N}$. We then introduce a Kullback-Leibler (KL) divergence loss l^m between $\tilde{\mathrm{C}}^m$ and C for each m as follows:

$$l^m = \sum_i \mathrm{D}_{\mathrm{KL}}(\tilde{\mathrm{C}}_i^m \| \mathrm{C}_i) \tag{6}$$

The overall loss L is then expressed as:

$$L_{\texttt{(SS)}_2\texttt{D}} = \sum_m l^m = \sum_m \sum_i \mathrm{D}_{\mathrm{KL}}(\tilde{\mathrm{C}}_i^m \| \mathrm{C}_i) \tag{7}$$

4 THEORETICAL ANALYSIS OF `AE-SVC`

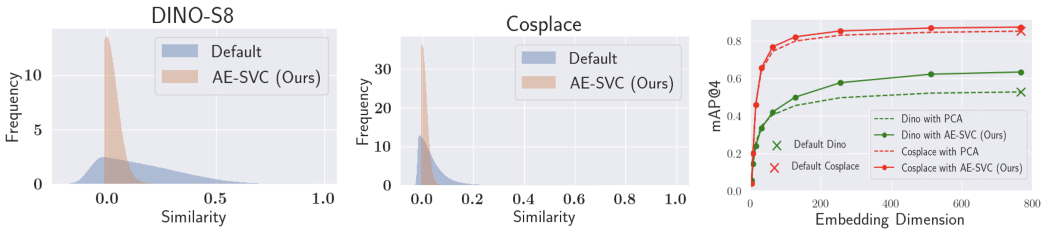
In Sec. 3.1, we stated that the discriminative power of the retrieval task is maximized when the variance of the cosine similarity distribution is minimized. This discriminative power can be modeled as the probability that the cosine similarity between two randomly drawn vectors S_a and S_b from the same class exceeds a given threshold τ:

$$P\left(\cos(S_a, S_b) > \tau\right) = \operatorname{erf}\left(\frac{\mathbb{E}[\cos(S_a, S_b)] - \tau}{\sqrt{\operatorname{Var}[\cos(S_a, S_b)]}}\right), \tag{8}$$

where erf is the error function. The vectors S_a and S_b are assumed to be drawn from the same Gaussian distribution. As evident from Eq. 8, the discriminative power increases as the variance of the cosine similarity distribution decreases. For a theoretical proof of this and alternative methods for modeling discriminative power, we refer the reader to Smith et al. (2023).

We now verify if the results of Gaussian assumptions also hold for real datasets. Fig. 3 shows the distribution of cosine similarities of the reference set feature vectors on the Pittsburgh30k dataset (Torii et al., 2013). We compute cosine similarities between all pairs of feature vectors in the reference set and plot the probability mass function (PMF) with and without `AE-SVC` for two models: the off-the-shelf foundation model Dino (Fig. 3a) and the SOTA dataset-specific model Cosplace (Berton et al., 2022) (Fig. 3b). Cosplace, which excels at retrieval, exhibits a PMF with significantly less variance compared to Dino, as seen in the blue curves in Fig. 3a and 3b. The lower variance indicates more discriminative feature vectors, which aligns with the findings from Eq. 8 that smaller variance enhances discriminative power and retrieval performance. Applying `AE-SVC` reduces the variance in both models (orange curves), with a more pronounced shift in Dino than in Cosplace. This suggests that `AE-SVC` benefits the off-the-shelf foundation model more (10%) than the dataset-specific model (2%), as reflected in the retrieval performance plot in Fig. 3c. Thus, the remainder of this work focuses on off-the-shelf foundation models.

We have mathematically and empirically demonstrated that the discriminative power of the retrieval task is maximized when the variance of the cosine similarity distribution is minimized. Next, we will



(a) (b) (c)

Figure 3: **`AE-SVC` reduces the variance of cosine similarity distributions in both foundation (a) and dataset-specific models (b), with a more significant shift in foundation models (a).** This results in greater improvement in retrieval performance for the foundation model (Dino) compared to the dataset-specific model (Cosplace), as shown in (c).

prove that the constraints introduced on the latent space in Sec. 3.1 indeed minimize this variance. Let Z be the latent space distribution of `AE-SVC`. By applying constraints in Eq. 2 and Eq. 4, the distribution Z has a zero mean, $\mu = 0$, and a diagonal covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, with eigenvalues $\lambda_i = \sigma_i^2$, where σ_i^2 is the variance of the individual dimensions of Z. Given two vectors X and Y drawn from Z, *i.e.*, $X, Y \sim Z$, the cosine similarity between X and Y is:

$$\cos(X, Y) = \frac{\sum_{i=1}^{d} X_i Y_i}{\sqrt{\sum_i X_i^2}\sqrt{\sum_i Y_i^2}} = \frac{X^T Y}{\sqrt{(X^T X)(Y^T Y)}} = \frac{X^T Y}{\|X\|\|Y\|}, \tag{9}$$

where subscript i denotes the i-th element of a vector , and $\|X\|$ and $\|Y\|$ denote the Euclidean norms of vectors X and Y respectively. Each term in the numerator of Eq. 9 is just a product of two random variables, but the denominator makes things complicated. A closed-form solution of a ratio of two random variables is difficult to calculate.

Let us introduce a relaxation that can be used to approximate the denominator and simplify the calculation. Using cosine's scale invariance (shown in A.2), we define a relaxation $r(Z)$ as:

$$r(Z) = \frac{Z}{m}, \tag{10}$$

where $m = \sqrt{\sum_i \sigma_i^2}$. Relaxation r has two properties. First, the expected value of the squared norm after relaxation, denoted as $\mathbb{E}[||r(Z)||^2]$, is equal to 1. Second, the ratio of variance to square of the mean is given by $\frac{\sum_k 2\sigma_k^2}{(\sum_k \sigma_k^2)^2}$, and it approaches 0 as the dimensionality d tends to infinity. Under the condition that the contribution of any single covariance eigenvalue is sufficiently small. Smith et al. (2023) show that these properties can then be used to approximate the cosine similarity as:

$$\cos(X, Y) = \cos(r(X), r(Y)) = \frac{\sum_{i=1}^{d} r(X)_i r(Y)_i}{\|r(X)\|\|r(Y)\|} \approx \sum_{i=1}^{d} r(X)_i r(Y)_i. \tag{11}$$

With this approximation, the moments of the cosine similarity can be easily calculated:

$$\mathbb{E}[\cos(X, Y)] = 0, \tag{12}$$

$$\mathrm{Var}(\cos(X, Y)) = \sum_{i=1}^{d} \frac{\sigma_i^4}{(\sum_{j=1}^{d} \sigma_j^2)^2}. \tag{13}$$

We can see from Eq. 13 that the cosine similarity distribution is highly dependent on the dimensions with high variance. Taking the gradient of Eq. 13 with respect to σ_i^2, we get:

$$\frac{\partial}{\partial \sigma_i^2} \mathrm{Var}(\cos(X, Y)) = 2 \left(\sum_{j=1}^{d} \sigma_j^2 \right)^{-3} \left[(\sum_{j=1}^{d} \sigma_j^2)\sigma_i^2 - (\sum_{j=1}^{d} \sigma_j^4) \right]. \tag{14}$$

We can see that the global minimum of the approximation for $\mathrm{Var}(\cos(X, Y))$ is achieved when all individual variances of the original distribution are equal, *i.e.*, $\sigma_i^2 = \sigma_j^2 \; \forall \, i, j$. This is enforced in AE-SVC by the variance constraint (Eq. 3). Therefore, we theoretically and empirically justified that AE-SVC minimizes the variance of the cosine similarity distribution, improving retrieval performance.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets: We evaluate our approach on four distinct image retrieval datasets: InShop (Liu et al., 2016), Stanford Online Products (SOP) (Song et al., 2016), Pittsburgh30K (Torii et al., 2013), and TokyoVal (Torii et al., 2015). InShop and SOP are state-of-the-art (SOTA) image retrieval datasets commonly used in metric learning research, while Pittsburgh30K and TokyoVal are recognized as SOTA place recognition datasets, frequently used in robotics research.

Foundation Models Evaluated: To demonstrate the broad applicability of our approach, we conducted experiments using four different foundational models trained with a variety of techniques and pre-training datasets: CLIP (Radford et al., 2021), DINO (Caron et al., 2021), DINOv2 (Oquab et al., 2024), and ViT (Dosovitskiy et al., 2021). The CLIP model was pre-trained on the LAION-2B dataset (Community, 2023) using a contrastive learning approach. Both DINO and ViT models were pre-trained in a self-supervised manner on the ImageNet dataset (Russakovsky et al., 2015). The DINOv2 model was pre-trained in a self-supervised manner on the LVD-142M dataset (Oquab et al., 2023). For the DINO and DINOv2 models, we evaluated different model sizes: specifically, DINO-S8 with 22.7 million parameters and DINO-B16 with 300 million parameters, as well as DINOv2-Small with 22.1 million parameters and DINOv2-Large with 300 million parameters.

Settings: Our work presents two contributions: the introduction of a novel space, AE-SVC, and a new method for learning an adaptive embedding, (SS)$_2$D, on top of AE-SVC. Consequently, our experiments are divided into two parts. First, we demonstrate that AE-SVC significantly enhances the retrieval performance of foundation models across various datasets and model choices. Second, we show that (SS)$_2$D facilitates more effective dimensionality reduction on top of AE-SVC, further improving performance at lower dimensions. We compare (SS)$_2$D with the following baselines: a variational auto-encoder (VAE) applied to AE-SVC for each dimensional size separately. Additionally, we compare (SS)$_2$D with using Similarity Space Distillation applied to each dimension separately, referred to as SSD. Note that SSD serves as a theoretical upper bound for (SS)$_2$D. We use mean Average Precision at $\mathbf{k}$ (mAP@$\mathbf{k}$), a standard metric for evaluating retrieval performance. We show additional results on the Recall@$\mathbf{k}$ metric in Appendix A.5.

5.2 RESULTS

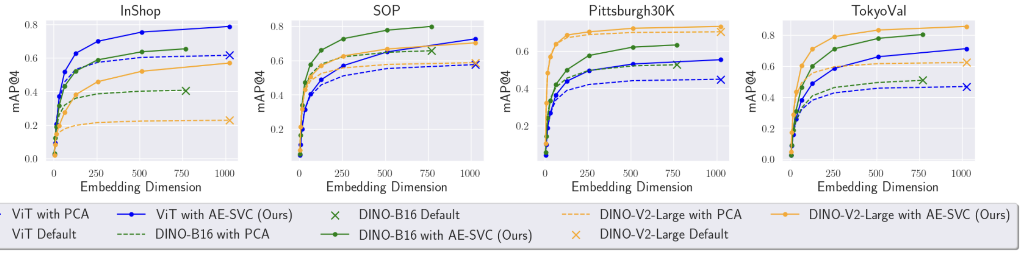


Figure 4: **AE-SVC significantly improves the retrieval performance of foundation models.** AE-SVC (solid lines) consistently outperforms the off-the-shelf foundation models, i.e., PCA (dashed lines), on four datasets, achieving a 15.5% average improvement in retrieval performance.

Fig. 4 shows the performance comparison between AE-SVC and PCA across 4 datasets, InShop, Stanford Online Products (SOP), Pittsburgh30K, and TokyoVal, using 3 foundation models (ViT, DINOv2-Large, and DINO-B16). The performance of PCA (dotted line) represents the default off-the-shelf performance of the foundation models, while the performance of AE-SVC is shown with solid lines. AE−SVC consistently outperforms PCA across all datasets and embeddings. For example, the DINOv2-Large embedding (yellow) shows a 24% improvement on InShop, 10% on SoP, 5% on Pittsburgh30K, and 22% on TokyoVal at full embedding size. Overall, AE−SVC achieves an average improvement of 15.5% across all datasets and embeddings at full size. Also, AE−SVC consistently surpasses PCA at smaller embedding sizes, providing a better size-performance trade-off. For additional results with more foundation models, see Appendix A.3.

Fig. 5 illustrates the advantages of (SS)$_2$D on two datasets: InShop and Pittsburgh30K, utilizing two foundation models, Dinov2-Large and Dino-S8. We compare (SS)$_2$D with a non-linear dimensionality reduction technique, Variational Auto Encoder (VAE) (black crosses). Additionally, we compare it with Similarity Space Distillation (SSD), which represents the theoretical upper bound of (SS)$_2$D. Note that unlike VAE (black crosses) and SSD (red crosses), which are trained for each dimension separately, (SS)$_2$D learns an adaptive embedding in one shot. From the plot, we observe that AE−SVC (orange) initially enhances the performance of the original embedding (blue), and (SS)$_2$D subsequently provides additional improvements of up to 10% at smaller embedding sizes. Also note that since SSD is trained for each dimension separately, it serves as the theoretical upper bound of (SS)$_2$D. The results indicate that (SS)$_2$D closely approaches this upper bound. For additional results involving more datasets and foundation models, please refer to Appendix A.4. As discussed in Sec. 1, a smaller embedding size directly translates to faster retrieval speeds. Therefore, the effective dimensionality reduction achieved with (SS)$_2$D results in improved retrieval speeds, as demonstrated in Fig. 1(**D**).

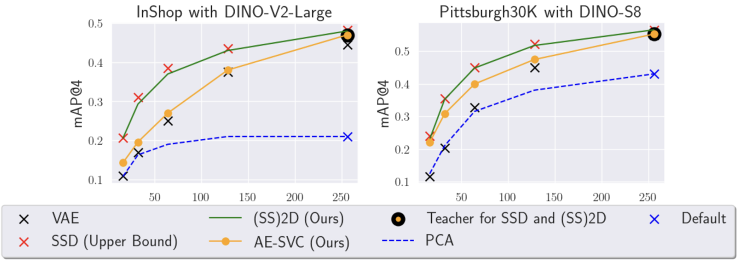


Figure 5: **Applying (SS)$_2$D over AE−SVC leads to further performance boost at lower embedding sizes.** Compared to VAE and SSD, (SS)$_2$D offers superior single-shot dimensionality reduction, achieving up to a 10% enhancement at smaller embedding sizes, closely approaching SSD's theoretical upper bound.

5.3 LIMITATIONS

While AE−SVC demonstrates significant improvements when applied to foundation models, its ability to enhance dataset-specific models is limited. The pre-trained models tailored for a specific dataset already capture the nuances required for effective retrieval, leaving less room for further optimization via AE−SVC. Additionally, (SS)$_2$D introduces a computational overhead during training due to the need to compute the loss functions for various embedding sizes. This overhead increases training time, though it is still substantially lower than the cost of retraining neural networks for different embedding sizes from scratch.

6 CONCLUSION AND FUTURE WORK

In this paper, we proposed AE−SVC and (SS)$_2$D, which improve retrieval scalability and efficiency. Our results show a significant improvement on 4 datasets, from image retrieval to place recognition tasks. AE−SVC demonstrates up to a 16% improvement in retrieval performance, and (SS)$_2$D shows an improvement of 10% for smaller embedding sizes, which further boosts retrieval efficiency. Future work includes new retrieval models that orchestrate the variance-aware property in training loss of foundation models. Although we observe the resulting difference in the cosine similarity distributions of the foundation models and the data set-specific models (Figure 3), the fundamental discrepancy between the embedding spaces remains unknown to the community.

REPRODUCIBILITY STATEMENT

In the theoretical analysis section (Sec. 4) of this work, we clearly state all the assumptions as and when necessary. Appendix A.1 and A.2 provides additional details necessary to understand the proofs in Sec. 4. We also discussed all the hyperparameters in Appendix A.6. Lastly, we release a reproducible codebase in the supplementary details. The code includes dataloaders, execution code, and links to download all the datasets and models used. We also provide detailed instructions on how to replicate our experiment setup and results, ensuring complete transparency and reproducibility.

REFERENCES

Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pp. 2654–2662, 2014.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=xm6YD62D1Ub`.

Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4878–4888, 2022.

Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

LAION Community. Laion-2b: A large-scale dataset for multimodal learning. `https://huggingface.co/datasets/laion/laion2B-en`, 2023. Accessed: 2024-05-16.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Sourav Garg, Tobias Fischer, and Michael Milford. Where is your place, visual place recognition? In *IJCAI*, volume 8, pp. 4416–4425, 2021.

Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pp. 369–386. Springer, 2020.

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, pp. 1735–1742. IEEE, 2006.

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

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *arXiv preprint arXiv:2308.00688*, 2023.

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

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

María Leyva-Vallina, Nicola Strisciuglio, and Nicolai Petkov. Data-efficient large scale place recognition with graded similarity supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23487–23496, 2023.

Po-han Li, Sravan Kumar Ankireddy, Ruihan Philip Zhao, Hossein Nourkhiz Mahjoub, Ehsan Moradi Pari, Ufuk Topcu, Sandeep Chinchali, and Hyeji Kim. Task-aware distributed source coding under dynamic bandwidth. *Advances in Neural Information Processing Systems*, 36, 2023.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1096–1104, 2016.

Feng Lu, Lijun Zhang, Xiangyuan Lan, Shuting Dong, Yaowei Wang, and Chun Yuan. Towards seamless adaptation of pre-trained models for visual place recognition. *arXiv preprint arXiv:2402.14505*, 2024.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4004–4012, 2016.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Lvd-142m: A curated dataset for self-supervised learning. `https://github.com/facebookresearch/dinov2`, 2023. Utilized in the pretraining of DINOv2 models.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 268–284, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

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*, pp. 815–823, 2015.

Ian Smith, Janosch Ortmann, Farnoosh Abbas-Aghababazadeh, Petr Smirnov, and Benjamin Haibe-Kains. On the distribution of cosine similarity with application to biology. *arXiv preprint arXiv:2310.13994*, 2023.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations (ICLR)*, 2020.

Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 883–890, 2013.

Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1808–1817, 2015.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

Evgeniya Ustinova and Victor Lempitsky. Learning deep embeddings with histogram loss. *Advances in neural information processing systems*, 29, 2016.

Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Deep metric learning: A survey. *arXiv preprint arXiv:1706.09720*, 2017.

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/CVF conference on computer vision and pattern recognition*, pp. 5022–5030, 2019.

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*, pp. 9489–9498, 2022.

Jiuhong Xiao, Gao Zhu, and Giuseppe Loianno. Visual geo-localization with self-supervised representation learning. *arXiv preprint arXiv:2308.00090*, 2023.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations (ICLR)*, 2017.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations (ICLR)*, 2019.

A APPENDIX

A.1 INVARIANCE OF ORTHOGONAL TRANSFORMATION

We show that cosine similarity is invariant to orthogonal transformation. Define $T_{\text{orth}} \in \{T : T^\top T = \mathbf{I}\}$ as an orthogonal transformation. Following equation 9, we see that

$$\begin{aligned} \cos(T_{\text{orth}}X, T_{\text{orth}}Y) &= \frac{(T_{\text{orth}}X)^\top T_{\text{orth}}Y}{\|T_{\text{orth}}X\|\|T_{\text{orth}}Y\|} \\ &= \frac{X^\top \mathbf{I} Y}{\sqrt{X^\top \mathbf{I} X}\sqrt{Y^\top \mathbf{I} Y}} \\ &= \frac{X^\top Y}{\|X\|\|Y\|} \\ &= \cos(X, Y) \quad \square. \end{aligned} \tag{15}$$

A.2 COSINE SCALE INVARIANCE

Similar to the derivation in A.1, we now show that cosine similarity is invariant to scaling. Again, we define a constant c and follow equation 9:

$$\begin{aligned} \cos\left(\frac{X}{c}, \frac{Y}{c}\right) &= \frac{X^\top Y / c^2}{\|X\|\|Y\|/c^2} \\ &= \frac{X^\top Y}{\|X\|\|Y\|} \\ &= \cos(X, Y) \quad \square. \end{aligned} \tag{16}$$

A.3 `AE-SVC` ADDITIONAL RESULTS WITH MORE FOUNDATION MODELS

In Fig. 4, we showed the advantage of using `AE-SVC` for embeddings of three models. Here, in Fig. 6, we show additional results with three more embeddings, namely: CLIP, DINO-V2-Small, and DINO-S8. We observe similar trends as previously noted. We release our codebase in the supplementary material for easy reproduction of all these results.

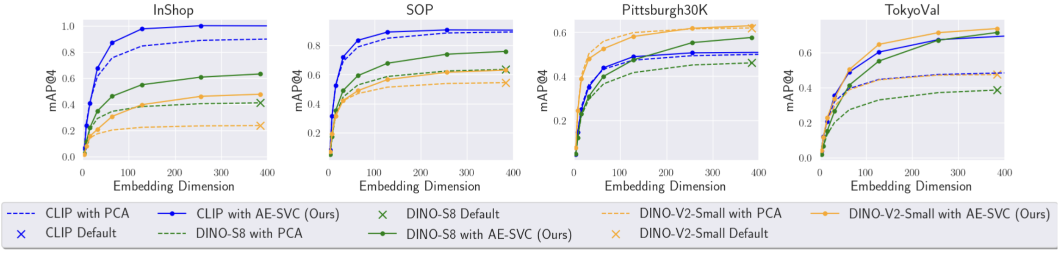


Figure 6: **`AE-SVC` significantly improves the retrieval performance of foundation models.** Here, we see a similar trend as discussed in the main paper but with three additional foundation models.

A.4 `(SS)`$_2$`D` ADDITIONAL RESULTS WITH MORE DATASETS AND FOUNDATION MODELS

In Fig. 5, we showed the advantage of using `(SS)`$_2$`D` on top of `AE-SVC` for additional improvement in retrieval performance at smaller dimensions. Here, in Fig. 7, we show additional results for `(SS)`$_2$`D` on SOP with DINO-B16 and on TokyoVal with ViT. We observe similar trends as previously noted. For trying `(SS)`$_2$`D` with more combinations of datasets and foundation models, please refer to our codebase that we release in the supplementary material.

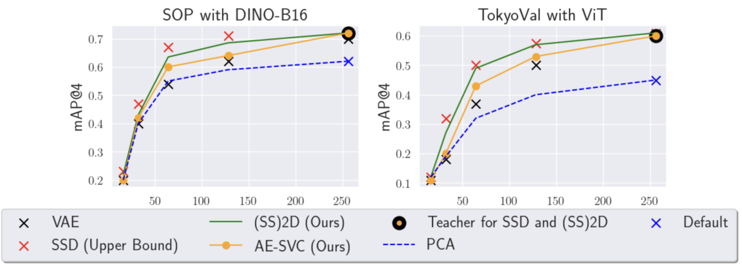


Figure 7: **Applying $(\mathtt{SS})_2\mathtt{D}$ over AE−SVC leads to further performance boost at lower embedding sizes.** Here, we see a similar trend as discussed in the main paper but with two additional foundation models and datasets.

A.5 AE−SVC ADDITIONAL RESULTS ON THE RECALL METRIC

In Fig. 4, we show the advantage of using AE−SVC for embeddings of three models on the mean Average Precision at $\mathbf{k}$ (mAP@$\mathbf{k}$) metric. Here, in Fig. 8, we show additional results on the Recall@$\mathbf{k}$ metric. We observe very similar trends on the Recall@$\mathbf{k}$ metric as we did on the mAP@$\mathbf{k}$ metric. We release a reproducible codebase for readers to experiment with both metrics.

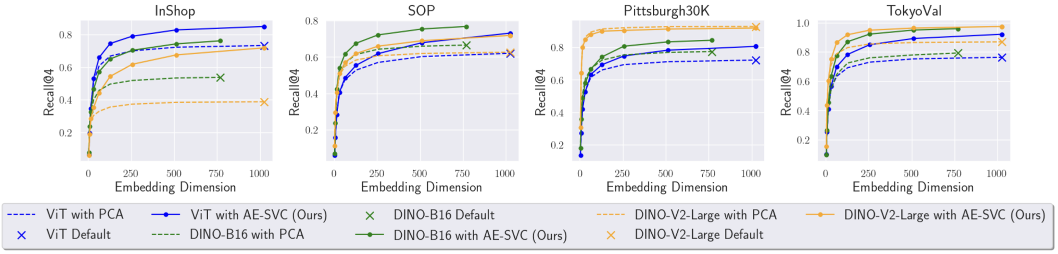


Figure 8: **AE−SVC significantly improves the retrieval performance of foundation models on the recall metric.** AE−SVC (solid lines) consistently outperforms the off-the-shelf foundation models, i.e., PCA (dashed lines), on four datasets.

A.6 HYPERPARAMETERS

We discussed four hyperparameters in Sec. 3.1: λ_{rec}, λ_{cov}, λ_{var}, and λ_{mean}. We empirically determined the values of these hyperparameters, ensuring optimal performance across all datasets. The fixed values are: $\lambda_{\text{rec}} = 25$, $\lambda_{\text{cov}} = 1$, $\lambda_{\text{var}} = 15$, and $\lambda_{\text{mean}} = 1$.