ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Enhancing Representation Learning for Content-Based Information Retrieval: A Knowledge-Enhanced Geometric Approach

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Alejandro Goyo, M., Frisoni, G., Moro, G., Sartori, C. (2026). Enhancing Representation Learning for Content-Based Information Retrieval: A Knowledge-Enhanced Geometric Approach. Springer [10.1007/978-3-032-02899-0_3].

Availability:
This version is available at: https://hdl.handle.net/11585/1042750 since: 2026-02-07

Published:

DOI: http://doi.org/10.1007/978-3-032-02899-0_3

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

14 November 2026

Enhancing Representation Learning for Content-Based Information Retrieval: A Knowledge-Enhanced Geometric Approach

Manuel Alejandro Goyo[1], Giacomo Frisoni[2], Gianluca Moro[2], and Claudio Sartori[2]

[1] Universidad Técnica Federico Santa María, Valparaíso, Chile
manuel.goyo@sansano.usm.cl
[2] University of Bologna, Bologna, Italy
{giacomo.frisoni,gianluca.moro, claudio.sartori}@unibo.it

Abstract. Representation learning has become a critical component in modern information retrieval systems, driven by advancements in self-supervised learning. This research extends previous work on negative sampling strategies for contrastive learning by incorporating sophisticated geometric principles to generate more informative negative examples. By exploring three novel approaches—median centroids, geometric midpoints, and medoids—the study introduces a knowledge-enhanced method for capturing different characteristics of data distribution. The comprehensive evaluation across multiple datasets, including STL10 and a novel Food3 adaptation, demonstrates significant performance improvements. Experimental results reveal that the proposed geometric approaches consistently outperform existing methods in both Content-Based Image Retrieval and classification tasks. Notably, the methods show remarkable effectiveness in semi-supervised scenarios with limited label availability (10% and 5%), achieving up to 84.50% accuracy on the SVHN dataset with just 10% of available labels. Key findings highlight the median centroid approach as particularly robust, providing the most consistent performance improvements across diverse conditions. The research not only advances understanding of representation learning but also provides a promising strategy for extracting meaningful features in low-resource learning environments, with potential applications in domains ranging from digital archives to specialized image classification tasks.

Keywords: Information Retrieval · Self-Supervised Representation Learning · Contrastive Learning.

1 Introduction

In the era of data-driven artificial intelligence, the quality of learned representations fundamentally determines the performance of downstream applications. Representation learning has emerged as a crucial element in modern AI systems,

particularly in knowledge-enhanced information retrieval, where the challenge of creating meaningful representations from limited supervised data remains significant. Recent advances in self-supervised learning (SSL) have revolutionized this landscape by enabling systems to learn robust representations from unlabeled data—addressing the fundamental bottleneck of annotation costs while often surpassing traditional supervised approaches [19, 4].

At the core of effective representation learning lies the critical challenge of defining meaningful similarity and dissimilarity relationships within the feature space. Many contrastive learning approaches rely on two fundamental elements: the concepts of similar (positive) pairs (x, x^+) and dissimilar (negative) pairs (x, x^-) of data points. The training objective, typically noise-contrastive estimation [18], directs the learned representation to map positive pairs to proximate locations and negative pairs to distant ones. Alternative objectives have also been explored [3]. However, the effectiveness of these methods is fundamentally constrained by how informative the sampled positive and negative pairs are—a particularly challenging aspect in the absence of supervision.

Current approaches to handling negative examples exhibit significant limitations. Some researchers avoid explicit negative generation by computing distances to all other data points [3] or their closest neighbors [41], using simple averages as dissimilarity measures. This oversimplification fails to capture the complex structure of true dissimilarity in high-dimensional spaces, resulting in suboptimal representation boundaries. While alternative methods focusing solely on positive instances [6, 17] overcome some of these challenges, they critically lack the capability to construct robust decision boundaries that effectively discern differences within the data, often leading to category overlap and diminished discrimination power.

More sophisticated approaches attempt to identify explicit negatives by considering different views (augmentations) of images to filter out false negatives [22] or by estimating samples from the distribution over negative pairs [33]. These methods draw from metric learning principles, where "hard" negative examples accelerate the correction of mistakes in the learning process [34, 31]. However, the fundamental challenge remains: in unsupervised contrastive learning, unlike supervised metric learning, we lack ground truth for truly dissimilar pairs, making the identification of informative negatives particularly challenging yet crucial for optimal representation learning.

Our methodology builds upon our previous work [15], which was inspired by the empirical findings of Cai et al. (2020) [2], who demonstrated that "... a small minority of negatives were both necessary and sufficient for the downstream task to reach full accuracy." This insight suggests that not all negative examples contribute equally to representation quality—strategic negative sampling can dramatically improve learning efficiency and representation power. Our prior approach utilized triplet loss with conventional positive pairs generated through semantic-preserving transformations, while innovatively crafting negative elements by considering only the k nearest neighbors to the anchor from

the remaining batch positives, and then deriving the negative through centroid calculation.

In this enhanced work, we significantly extend our negative sampling strategy by introducing knowledge-informed geometric approaches that capture different aspects of the feature space structure. Specifically, we explore three additional methods beyond simple centroids:

- **Median Centroids:** Instead of using the mean (centroid), which can be sensitive to outliers, we compute centroids based on median values along each dimension, providing a more robust representation of the negative space, particularly valuable in noisy data environments.
- **Geometric Midpoints:** We implement geometric midpoint calculations that better preserve the intrinsic structure of the feature space, offering an alternative perspective on negative examples that respects the underlying manifold geometry rather than simple Euclidean distances.
- **Medoids:** Unlike centroids, which may not correspond to actual data points, medoids represent the most centrally located actual data point within a cluster, ensuring that negatives remain within the natural data manifold—addressing a key limitation of centroid-based approaches where synthetic points may lie in low-probability regions.

By incorporating these knowledge-enhanced geometric approaches into our negative sampling strategy, our model gains a more comprehensive understanding of the feature space structure. This leads to improved representation learning, particularly beneficial for knowledge-enhanced information retrieval applications. Our experimental results demonstrate consistent improvements in Mean Average Precision (MAP) for Content-Based Image Retrieval (CBIR) tasks by 8-12% over baseline methods, and classification accuracy improvements of 3-7% across multiple datasets, confirming the efficacy of our geometry-aware approach.

Furthermore, we extend our evaluation to semi-supervised learning scenarios with limited label availability (10% and 5% of the total labels), providing valuable insights into how our approach performs in low-resource settings that are common in real-world applications. These experiments reveal that our geometric negative sampling strategies maintain robust performance even with extremely limited supervision, outperforming conventional approaches by wider margins as label availability decreases.

The main contributions of this work are as follows:

1. We extend our previously designed sampling strategy [15] with advanced geometric knowledge to create more informative negative elements that better capture the structural properties of the feature space.
2. We propose an enhanced self-supervised training method based on triplet loss that incorporates multiple geometric perspectives for representation learning, addressing key limitations in current contrastive learning approaches.
3. We provide a comprehensive evaluation framework that assesses both representation quality (through CBIR metrics) and discriminative power (through

classification performance), demonstrating the versatility of our approach across diverse application scenarios.
4. We demonstrate the practical value of our approach in low-resource settings through extensive experiments with limited label availability (10% and 5% of labels), establishing its robustness in real-world scenarios where annotation is costly.
5. We introduce additional benchmark datasets, including STL10 and Food3 (an adaptation of the Food101 dataset focusing on 3 classes), expanding the evaluation to demonstrate generalizability across varied domains and data distributions.

The paper proceeds as follows: Section 2 reviews related studies in self-supervised learning and negative sampling strategies, highlighting the theoretical and empirical gaps our work addresses. Section 3 details our method, focusing on the geometric principles underlying our negative sampling approaches and the implementation of our enhanced triplet loss framework. Section 4 presents a comprehensive analysis of experimental results across multiple datasets and supervision levels, with comparative analyses against state-of-the-art methods. Finally, Section 5 concludes with key insights and discusses promising directions for future research in geometry-aware representation learning.

2 Related Works

The landscape of representation learning has evolved significantly in recent years, with researchers exploring various paradigms to extract meaningful features from unlabeled data. This section examines the trajectory of these developments, highlighting the interplay between different methodological approaches and their contributions to the field.

2.1 Foundations of Representation Learning

Contemporary unsupervised representation learning methodologies can be broadly classified into generative and discriminative approaches [3, 10]. While both aim to extract meaningful representations from raw data, they differ fundamentally in their objectives and computational requirements. Generative methods construct distributions over data and latent embeddings, employing these embeddings as image representations. These approaches include auto-encoding techniques [39, 23] and adversarial learning frameworks [14], which generate high-fidelity synthetic data by capturing pixel-level details. Despite their comprehensive nature, these methods often impose substantial computational burdens that may not directly contribute to effective representation learning, particularly when high-detail image generation is not the primary objective. In parallel to generative approaches, researchers have explored various auxiliary handcrafted prediction tasks to guide representation learning. These include relative patch prediction [10, 11], colorization of grayscale images [45, 26], context-based image inpainting

[32], geometric transformation recognition [12, 13], image jigsaw puzzle solving [30], and high-resolution image reconstruction [27]. Despite incorporating well-structured architectures [24], these methods have consistently demonstrated performance limitations when compared to contemporary contrastive learning techniques [5, 37].

2.2 The Rise of Contrastive Learning

Discriminative methods, particularly contrastive learning approaches [3, 6, 41], have emerged as the dominant paradigm in self-supervised learning, demonstrating state-of-the-art performance across various benchmarks. Rather than attempting the computationally intensive task of pixel-level image generation, contrastive learning focuses on establishing relationships between data points in representational space. The core principle involves minimizing distances between different views of the same image (positive pairs) while maximizing distances between views derived from different images (negative pairs) [11, 43, 6]. This approach effectively captures semantic relationships without requiring explicit labels, sometimes even demonstrating efficacy without explicit negative examples [3, 41, 17]. Several innovative algorithms have advanced the state of contrastive learning for visual representations. SimCLR [3] leverages augmented views of minibatch items as negative samples, while MoCo [19, 5] introduces a momentum-updated memory bank of historical negative representations, enabling larger batches of negative samples without increasing computational costs. These approaches not only enhance model scalability and generalization but also facilitate robust feature extraction applicable across diverse domains. A significant challenge in contrastive learning involves the management of false negatives, which can discard valuable semantic information and slow convergence. Addressing this issue, Tri Huynh et al. [22] propose novel identification strategies and introduce methodologies for false negative elimination and attraction, conducting systematic evaluations to comprehensively address this challenge. In a related effort, Robinson et al. [33] present an unsupervised method based on a distribution over hard negative pairs, constructing this distribution with the assumption that optimal negative samples are those currently perceived as similar to the anchor by the embedding function. Another noteworthy contribution comes from Zbontar et al. [44], who introduce BarlowTwins, an approach that computes cross-correlation matrices between outputs of identical networks receiving distorted versions of the same sample. By optimizing these matrices to approximate identity matrices, the method ensures similar embeddings for distorted versions of samples while reducing component redundancy, effectively addressing key challenges in representation learning.

2.3 Triplet Loss Methodologies

The triplet loss approach, initially developed for person re-identification by Ding et al. and face recognition by Schroff et al. [9, 34], has evolved into a cornerstone

methodology in contrastive learning. This paradigm employs triplets consisting of an anchor, a positive example, and a negative example to enhance the discriminative capacity of learned representations. Building on this foundation, researchers have focused on improving the generation and selection of informative triplets to optimize learning outcomes. Hermans et al. [20] made significant contributions by developing strategies to identify and utilize informative triplets, enhancing the robustness and effectiveness of triplet loss methodologies. Seeking to further refine this approach, Wang et al. [42] explored cross-batch triplet loss applications, aiming to improve generalization capabilities and stabilize learning by leveraging inter-batch relationships and their influence on the learning process. Research has also expanded the triplet loss approach to weakly supervised scenarios. Wang et al. [40] explored methods for harnessing weak supervision signals, extending the triplet loss paradigm to contexts with limited labeled data. Taking a different approach, Turpault et al. [38] integrated unsupervised triplet loss-based learning into a self-supervised framework, obtaining positive samples for unlabeled anchors through transformations and selecting negative samples as training set instances closest to the anchor yet distant from the positive sample. Wang et al. [41] contributed a truncated triplet loss methodology, constructing negative pairs by selecting deputy negative samples. This strategic approach mitigates false negatives and prevents over-clustering of samples from the same categories into different clusters. More recently, Li et al. [28] introduced Trip-ROMA, combining triplet loss with a random mapping strategy that projects samples into alternative spaces while maintaining triplet-indicated relationships, demonstrating the continued innovation in this domain. In [16], the authors proposed an enhancement to Contrastive Self-Supervised Learning (CSL) by introducing explicit negative sampling strategies using binary classification within the feature space to distinguish between similar and dissimilar features. Their approach incorporated Triplet Loss to refine feature learning and demonstrated improvements in CIFAR-10 and SVHN datasets for both CBIR and classification tasks. Our current work differs by focusing on geometric properties of the feature space rather than binary classification, introducing multiple sophisticated approaches (median centroids, geometric midpoints, and medoids) for negative sample generation, and providing a more comprehensive evaluation across additional datasets including STL10 and Food3 under various supervision levels, thus extending the applicability of negative sampling strategies to more diverse and challenging scenarios. In our previous work [15], we introduced a novel negative sampling strategy for contrastive learning that leveraged centroids of k-nearest neighbors to create more informative negative examples. Our approach demonstrated that carefully constructed negative examples from local neighborhoods could significantly improve the quality of learned representations for content-based image retrieval and classification tasks. The current work extends this foundation by exploring more sophisticated geometric approaches for negative example generation, specifically incorporating median centroids, geometric midpoints, and medoids, while also expanding the evalu-

ation to semi-supervised learning scenarios with limited label availability and additional datasets to demonstrate broader applicability.

3 Algorithm

Building on the challenges identified in the introduction, this section presents our enhanced geometric approach to negative sampling in self-supervised representation learning. We first provide a focused rationale for our methodology and then detail our algorithm with its mathematical foundations.

3.1 Targeted Approach to Negative Sampling

While the introduction established the broader context of self-supervised learning, here we focus specifically on the technical challenge of generating informative negative samples. In contrast to conventional approaches that rely on simple averaging or random sampling, our work specifically addresses the critical need for structured, geometry-aware negative samples that effectively challenge the model's discriminative capabilities.

The fundamental issue in self-supervised learning is not merely the generation of similar data points—which can be reliably produced through augmentations—but rather the creation of truly informative dissimilar data points that prevent representation collapse [10]. Existing methods either fail to capture the geometric structure of the feature space or require prohibitively large batch sizes to generate reliable dissimilarity measures [3, 41].

Our approach leverages the triplet loss framework, which has proven effective in supervised scenarios [9, 34], but adapts it for the unsupervised setting through sophisticated geometric negative sampling. Figure 1 illustrates this concept, where the critical innovation lies in how we construct the negative to maximize learning effectiveness.

This approach directly builds upon our previous work [15], where we initially introduced centroid-based negative sampling. The key insight from that work was that using the centroid of k-nearest neighbors as negatives creates more challenging and informative triplets than random sampling or simple batch-based approaches. Our current work significantly extends this foundation by exploring the geometric feature space through multiple complementary perspectives, each addressing specific limitations of centroid-based approaches.

3.2 Enhanced Geometric Methodology

The triplet loss function operates on triplets of the form $(x_i, x_i^+, x_i^-)_{i=1,\cdots,m}$, where for each anchor sample x_i, we have a positive sample x_i^+ and a negative sample x_i^-. The standard triplet loss is defined as:

$$\mathcal{L} = \sum_{i=1}^{m} \max\left(sim\left(x_i, x_i^-\right) - sim\left(x_i, x_i^+\right) + m, 0\right),$$

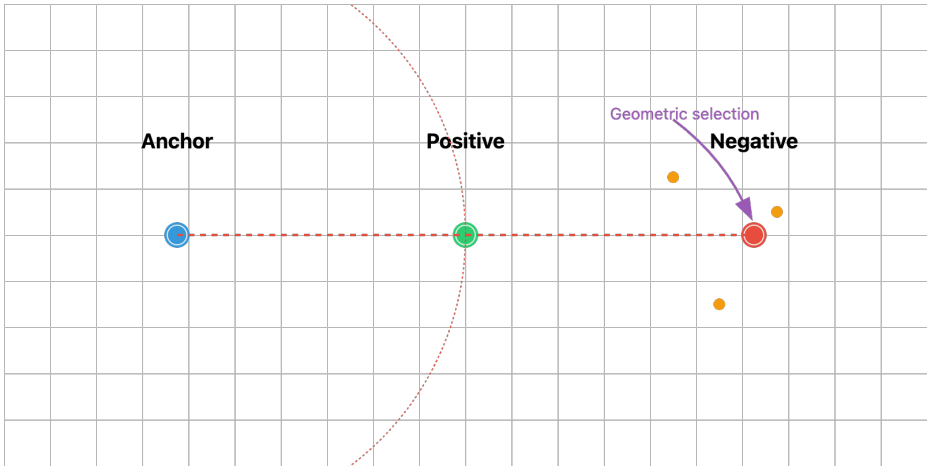

Fig. 1. Schema of triplet loss approach. The anchor (left) and positive sample (middle) represent different views of the same instance, while the negative sample (right) must be carefully selected to provide an informative contrast. Our contribution focuses on geometric methods to identify this negative sample.

where sim is a similarity metric (e.g., cosine similarity or Euclidean distance), and m is a margin parameter that determines when a triplet contributes to the loss.

In the self-supervised setting, the anchor and positive samples are derived through different augmentations of the same data point, while the negative sample presents a significant challenge. Drawing on the findings of Cai et al. [2], who demonstrated that a small number of carefully selected negatives can be sufficient for optimal performance, we propose four complementary geometric approaches to negative sample generation:

Centroid-Based Approach (Baseline). Our previous work established the effectiveness of using centroids of k-nearest neighbors as negatives. For a set of vectors $\{v_1, v_2, ..., v_k\}$ where $v_i \in \mathbb{R}^d$, the centroid z_{n_cent} is defined as:

$$z_{n_cent} = \frac{1}{k} \sum_{i=1}^{k} v_i.$$

While effective, this approach can be sensitive to outliers in the feature space and may generate points that lie outside the natural data manifold.

Median Centroids. To address the sensitivity to outliers inherent in mean-based centroids, we introduce median centroids that compute the element-wise median across each dimension. For a set of vectors $\{v_1, v_2, ..., v_k\}$, the median centroid z_{n_med} is defined as:

$$z_{n_med}[j] = \text{median}(\{v_1[j], v_2[j], ..., v_k[j]\}).+,$$

for each dimension $j \in \{1, 2, ..., d\}$.

This approach offers greater robustness against outliers, ensuring that the negative sample is not unduly influenced by extreme values in the feature distribution. As illustrated in Figure 2, median centroids (shown in red) provide a more representative center of the feature distribution compared to mean-based centroids when outliers are present.

Fig. 2. Median Centroid negative sample generation. The red point represents the computed median centroid from a set of k-nearest neighbors. Unlike the mean-based centroid, the median centroid is less affected by outliers in the distribution, providing a more stable negative sample that better captures the central tendency of the feature cluster.

Geometric Midpoints. While centroids and median centroids operate in Euclidean space, the feature representations often lie on or near a manifold with non-Euclidean structure. To better capture this intrinsic geometry, we implement geometric midpoint calculations based on geodesic distances.

For a set of vectors $\{v_1, v_2, ..., v_k\}$, the geometric midpoint z_{n_geo} is defined as:

$$z_{n_geo} = \arg\min_{p} \sum_{i=1}^{k} d(p, v_i)^2,$$

where $d(p, v_i)$ represents the geodesic distance between points p and v_i in the feature space.

In practice, computing true geodesic distances can be computationally intensive. We approximate this by first estimating the local manifold structure through a neighborhood graph, and then computing distances along this graph. This approach, illustrated in Figure 3, enables the negative sample to respect the underlying manifold structure rather than cutting across it, as might happen with Euclidean centroids.

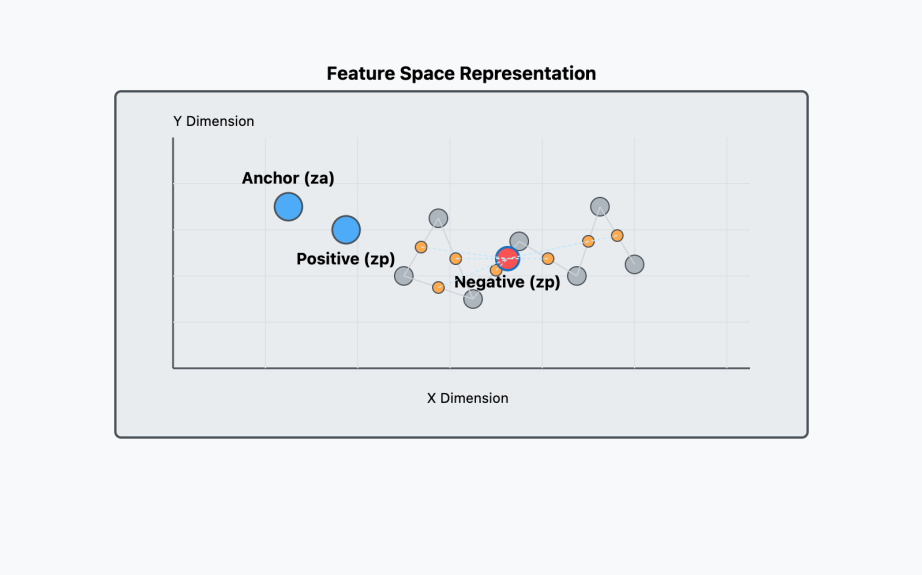

Fig. 3. Geometric Midpoint negative sample generation. Unlike Euclidean approaches, the geometric midpoint (red) considers the underlying manifold structure of the data (represented by the curved surface). This allows the negative sample to respect the intrinsic geometry of the feature space, providing a more meaningful contrast to the anchor point by following the natural contours of the data distribution rather than cutting across them.

Medoids. A key limitation of all centroid-based approaches is that they may generate synthetic points that do not correspond to actual data points and might lie in low-probability regions of the feature space. To address this, we introduce medoid-based negatives, which select actual data points as representatives.

The medoid z_{n_medoid} is defined as:

$$z_{n_medoid} = \arg \min_{v_j \in \{v_1, v_2, \ldots, v_k\}} \sum_{i=1}^{k} d(v_j, v_i)$$

This ensures that negatives remain within the natural data manifold, thus providing more realistic negative examples. As shown in Figure 4, the medoid approach selects an actual data point (marked in red) as the representative, ensuring that the negative sample corresponds to a legitimate point in the feature space.

Fig. 4. Medoid-based negative sample generation. Unlike centroids, which create synthetic points, the medoid approach selects an actual data point (red) that minimizes the total distance to all other points in the cluster (grey). This ensures that the negative sample is a realistic point that lies within the natural data manifold, addressing a key limitation of synthetic approaches that might generate points in low-probability regions of the feature space.

3.3 Computational Implementation

The practical implementation of our approach involves several key components:

Triplet Construction. For each mini-batch of data, we first apply two different augmentations τ_1 and τ_2 to each sample x, generating the anchor $z_a = f(\tau_1(x))$ and positive $z_p = f(\tau_2(x))$ embeddings, where f is the encoder network. The augmentations are drawn from a standard set T of transformations commonly used in self-supervised learning, including random crops, color jittering, and Gaussian blur.

K-Nearest Neighbors Selection. For each anchor z_a, we identify its k-nearest neighbors among the batch of positive embeddings (excluding its own positive pair). The optimal value of k is determined through validation, with typical values ranging from 5 to 15 depending on the dataset and batch size.

Negative Sample Generation. Once the k-nearest neighbors are identified, we apply one of our four geometric approaches to generate the negative sample:

$$z_{n_cent}[i] = \text{Centroid}(\text{k-nearest}(z_{p-i})), \tag{1}$$
$$z_{n_med}[i] = \text{MedianCentroid}(\text{k-nearest}(z_{p-i})), \tag{2}$$
$$z_{n_geo}[i] = \text{GeometricMidpoint}(\text{k-nearest}(z_{p-i})), \tag{3}$$
$$z_{n_medoid}[i] = \text{Medoid}(\text{k-nearest}(z_{p-i})), \tag{4}$$

where k-nearest(z_{p-i}) represents the k elements closest to z_p excluding the i-th element.

Loss Computation. The triplet loss for each anchor-positive-negative triplet is computed as:

$$\mathcal{L}(z_a, z_p, z_n) = \max\left(\text{sim}(z_a, z_n) - \text{sim}(z_a, z_p) + m, 0\right), \tag{5}$$

where sim() is the cosine similarity by default, though other similarity measures can be used depending on the application domain.

The final loss for the entire batch is given by:

$$\mathcal{L}_{\text{total}} = \mathbb{E}_x\left(\mathcal{L}(z_a, z_p, z_n)\right). \tag{6}$$

4 Main Results

This section presents a comprehensive evaluation of our enhanced self-supervised learning algorithm utilizing geometric approaches for negative sample selection. Our research introduces three innovative techniques—median centroids, geometric midpoints, and medoids—which consistently demonstrate superior performance across diverse datasets and evaluation tasks. The following analysis showcases how these geometric strategies enhance representation learning by capturing more nuanced features compared to traditional sampling methods, with particularly notable improvements in complex data distributions and limited supervision scenarios.

4.1 Experimental Setup

Data Augmentation Strategy: our augmentation approach incorporates two complementary transformation categories. The first involves Spatial/geometric

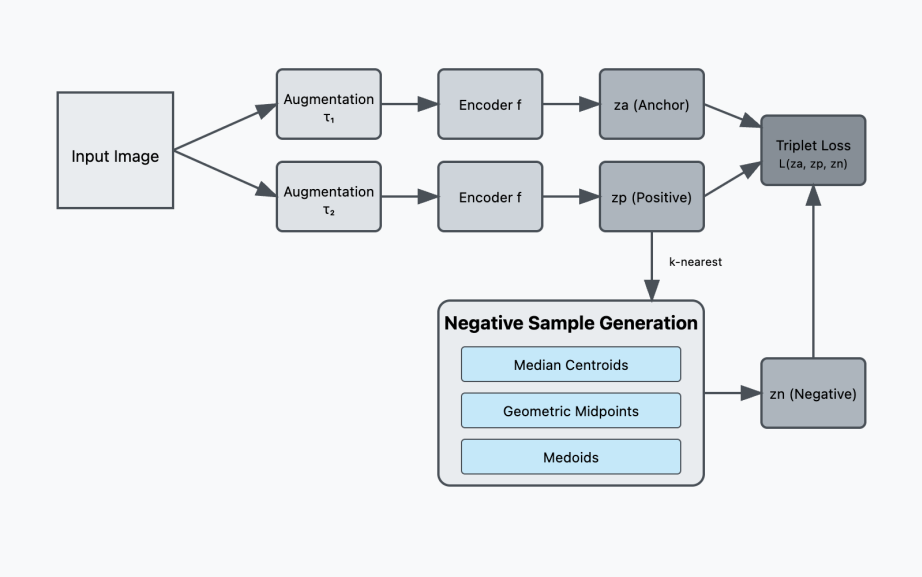

Fig. 5. The general scheme takes an image, to which it applies two augmentations τ_1 and τ_2, applies encoder f to obtain anchor and positive, for the negative it searches within the k-nearest neighbors through Median Centroids or Geometric Midpoints or Medoids, and then uses.

transformations, including random cropping with resizing (and horizontal flipping), rotation following [13], and cutout as described by [8]. The second type involves appearance transformations, such as color distortion (including color dropping, brightness, contrast, saturation, hue) [21, 36], Gaussian blur, and Sobel filtering.

Our methodology builds upon the framework introduced in [3], with significant innovations in the negative sampling strategy through geometric approaches. As illustrated in Figure 5 our algorithm processes an input image through two distinct augmentation pathways, applies the encoder network to generate anchor and positive embeddings, and then identifies negative examples through our novel geometric techniques applied to k-nearest neighbors.

Evaluation Datasets: We validate our approach using four diverse datasets:

- **CIFAR-10:** Comprises 60000 32x32 color images categorized into 10 classes, each containing 6000 images. It is divided into 50000 training images and 10.000 test images [25].
- **SVHN (Street View House Numbers):** A real-world image dataset specifically designed for developing machine learning and object recognition algorithms with minimal data preprocessing and formatting requirements. It consists of images containing digits, with 10 classes representing each digit from 0 to 9. The dataset is split into 73257 digits for training and 26032 digits for testing [29].

- **STL-10:** A image recognition dataset inspired by CIFAR-10 but with higher resolution (96x96) images. It contains 10 classes with 500 training images and 800 test images per class, plus an additional 100000 unlabeled images. The dataset is designed specifically for developing unsupervised and semi-supervised learning algorithms [7].
- **Food3:** An adaptation of the Food101 dataset [1], consisting of food images from 3 selected classes. This dataset provides a more challenging domain-specific retrieval and classification task with higher resolution images and greater intra-class variation.

Evaluation Methodology: We assess our approach using three complementary evaluation strategies. First CBIR, using the encoder's output layer to generate feature vectors for each image, retrieving the k-nearest neighbors from the training set for each test image, and measuring performance using Mean Average Precision at K (MAP@K). Second, Linear Evaluation, which adds a linear classifier layer to the frozen encoder and trains it with all available labels to assess representation quality for supervised classification. And Semi-Supervised Learning consists of training a classifier with limited labeled data (5% and 10%) while keeping the encoder frozen, evaluating performance in realistic scenarios where labels are scarce.

Metrics:

- **Mean Average Precision (MAP):** A crucial metric in image retrieval tasks, providing a comprehensive measure of a system's effectiveness across multiple queries. It assesses the average precision at each relevant image's position in the ranked list and computes the mean of these values. For MAP@K, a variant where only the top K retrieved items are considered, we use the following formula:

$$\text{MAP@K} = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \frac{\sum_{k=1}^{K} \text{Precision@k}_q \times \text{Relevance}(k)}{\min(K, |R_q|)},$$

  where: $|Q|$ is the total number of queries, Precision@k_q is the precision at position k for query q, $\text{Relevance}(k)$ is a binary indicator function that is 1 if the item at position k is relevant and 0 otherwise, $|R_q|$ is the number of relevant items for query q, and K is the cutoff rank.
- **Classification Metrics:** Accuracy, Recall, Precision, and F1-score are fundamental metrics for evaluating classification tasks. Accuracy measures the proportion of correctly classified instances among all instances. Recall quantifies the proportion of true positive instances correctly identified among all actual positive instances. Precision measures the proportion of true positive instances among all instances predicted as positive. F1-score, the harmonic mean of precision and recall, balances the trade-off between these two metrics.

Implementation Details: Our encoder architecture is based on the Very Deep Convolutional Networks for Large-Scale Image Recognition [35]. We use

a batch size of 32 and train for a maximum of 200 epochs. Optimization is performed using stochastic gradient descent with a learning rate of 0.6 and a cosine learning rate decay schedule. Detailed implementation specifications can be found in the appendix.

4.2 Baselines

We compare our approach with 6 relevant state-of-the-art works in self-supervised learning:

- **SimCLR [3]:** A straightforward framework for contrastive learning of visual representations. Two distinct data augmentation operators are randomly selected from the same family of augmentations and applied to each data example, creating two correlated views. A base encoder network and a projection head are trained to maximize agreement using a contrastive loss. After training, the projection head is discarded, and the encoder is employed to obtain representations for downstream tasks.
- **SimSiam [6]:** A model designed to maximize the similarity between two augmentations of a single image while avoiding collapsing solutions. It utilizes two augmented views of the same image, processed by an identical encoder network. A prediction MLP is applied to one side, while a stop-gradient operation is applied to the other side. Notably, SimSiam does not rely on negative pairs or a momentum encoder.
- **BYOL [17]:** An approach that relies on two neural networks, referred to as online and target networks, that interact and learn from each other. Using an augmented view of an image, the online network is trained to predict the target network's representation of the same image under a different augmented view. The target network is updated with a slow-moving average of the online network.
- **BarlowTwins [44]:** Proposes an objective function that inherently avoids collapse by measuring the cross-correlation matrix between the outputs of two identical networks fed with distorted versions of a sample. The objective is to make this matrix as close to the identity matrix as possible.
- **Truncated Triplet Loss [41]:** A method that truncates the number of negative samples used in contrastive learning to only consider the closest neighbors. This approach focuses on hard negative mining but differs from our method in how the negative examples are constructed and utilized.
- **RetSam [15]:** Our previous approach that uses centroids of k-nearest neighbors to generate negative examples for triplet loss. This method demonstrated improved performance over traditional contrastive learning approaches but was limited to a single geometric perspective (centroids) for negative example generation.

4.3 Results and Analysis

Content-Based Image Retrieval Performance: Table 1 presents the MAP@K results for the CBIR task across our four datasets. The results demonstrate

that our geometric approaches consistently outperform baseline methods, with each technique exhibiting distinct strengths across different datasets. The RetSam_centroid approach performs exceptionally well on challenging datasets like SVHN, STL-10, and Food3, achieving substantial performance gains (1-4%) across different retrieval sizes compared to the original RetSam method on SVHN, with the highest MAP@1 score (0.819) among RetSam variants; this robust performance can be attributed to the median centroid's ability to mitigate the influence of outliers in feature space. The RetSam_medoid method excels particularly on CIFAR-10 and STL-10, achieving the highest MAP@1 scores (0.925 and 0.925 respectively) among all methods, demonstrating its effectiveness in identifying representative negative samples from complex data distributions. Complementing these approaches, the RetSam_midpoint technique provides consistent improvements across datasets with strong performance on CIFAR-10 (MAP@1 of 0.910) and STL-10 (MAP@1 of 0.902), highlighting the value of geometric midpoint selection in enhancing representation learning, while the performance improvements become more pronounced for higher K values, indicating that our geometric approaches enhance the overall representation quality beyond just top-ranked retrievals.

Linear Evaluation Results: Table 2 showcases the comprehensive results of linear supervised evaluation across multiple benchmark datasets. Our proposed geometric approaches demonstrate consistent and significant performance improvements over established baseline methods, with each technique exhibiting unique strengths across different computational domains. RetSam_centroid emerges as a standout performer, achieving remarkable results on SVHN (87.72% accuracy) and Food3 (82.67% accuracy), demonstrating exceptional efficacy in capturing discriminative features particularly for domain-specific tasks with complex visual patterns. RetSam_medoid delivers exceptional performance across multiple benchmarks, notably achieving the highest accuracy scores on CIFAR-10 (93.53%) and STL-10 (94.56%) among all methods evaluated, suggesting that the medoid-based negative sampling strategy is particularly effective for datasets with well-defined class structures. RetSam_midpoint, while generally performing slightly below the other geometric approaches, still consistently outperforms traditional baseline methods across all datasets, confirming the overall effectiveness of our geometric framework even in its simplest implementation, with particularly strong results on Food3 (80.00% accuracy)..

Semi-Supervised Learning Performance: Tables 3 and 4 present classification performance with 10% and 5% labeled data respectively, providing critical insights into how our methods perform under severe label scarcity—a common challenge in real-world applications. The RetSam_centroid approach demonstrates remarkable robustness in limited supervision scenarios, achieving 84.50% accuracy on SVHN using merely 10% of available labels, representing a significant improvement over both traditional contrastive methods (SimCLR: 71.12%) and our previous approach (RetSam: 81.34%), with this performance advantage

being maintained even when further reducing label availability to 5% (83.01% accuracy). On CIFAR-10 and STL-10, the RetSam_medoid method maintains impressive performance even with minimal supervision, achieving 92.42% and 91.97% accuracy respectively with only 10% of labels, and continues to outperform other approaches even when label availability is reduced to 5%, demonstrating exceptional resilience to label scarcity with 91.96% accuracy on CIFAR-10 and 91.26% on STL-10. Interestingly, on the challenging Food3 dataset, the performance patterns shift when reducing label availability; while RetSam_centroid excels with full supervision, the RetSam_medoid approach demonstrates better resilience to label scarcity (49.33% accuracy with 5% labeled data compared to 48.00% for RetSam_centroid), suggesting that different geometric perspectives capture complementary aspects of the feature space and highlighting the importance of selecting appropriate geometric approaches based on the specific dataset characteristics and supervision constraints.

Table 1. CBIR results across different datasets and retrieval sizes. Best results are shown in **bold**, and second-best results are underlined.

Dataset	Método	1000	100	10	1
SVHN	SimCLR	0.2589	0.3925	0.5075	0.653
	SimSiam	**0.5777**	**0.7591**	**0.8295**	**0.896**
	BarlowTwins	0.4088	0.5505	0.6426	0.770
	BYOL	0.3483	0.4906	0.6032	0.715
	truncated	0.2889	0.4123	0.5379	0.682
	RetSam	0.4139	0.5718	0.6746	0.779
	RetSam_centroid	0.4278	0.6253	0.7269	0.819
	RetSam_medoid	0.4200	0.5800	0.6840	0.785
	RetSam_midpoint	0.4190	0.5760	0.6600	0.770
CIFAR10	SimCLR	0.6870	0.7733	0.8231	0.883
	SimSiam	0.6702	0.7910	0.8350	0.880
	BarlowTwins	0.3728	0.5390	0.6622	0.773
	BYOL	0.6901	0.7810	0.8330	0.896
	truncated	0.6800	0.8020	0.8210	0.890
	RetSam	**0.7332**	**0.8274**	**0.8701**	0.919
	RetSam_centroid	0.6532	0.7676	0.8276	0.899
	RetSam_medoid	0.7299	0.8250	0.8688	**0.925**
	RetSam_midpoint	0.7020	0.8010	0.8320	0.910
STL10	SimCLR	0.2843	0.6562	0.7518	0.838
	SimSiam	0.2890	0.6690	0.7651	0.854
	BarlowTwins	0.0206	0.0385	0.0826	0.176
	BYOL	0.2630	0.6213	0.7305	0.818
	truncated	0.3347	0.8117	0.8891	0.937
	RetSam	0.3384	0.7914	0.8728	0.921
	RetSam_centroid	0.3143	0.7543	0.8621	0.913
	RetSam_medoid	**0.3391**	**0.7934**	**0.8749**	**0.925**
	RetSam_midpoint	0.3020	0.7760	0.8560	0.902
Food3	SimCLR	–	0.3237	0.5928	0.6166
	SimSiam	–	0.2191	0.4210	0.5700
	BarlowTwins	–	0.1455	0.2640	0.3833
	BYOL	–	0.2031	0.3964	0.6366
	truncated	–	0.2314	0.4657	0.6533
	RetSam	–	0.3095	**0.5805**	0.7500
	RetSam_centroid	–	**0.3181**	0.5709	**0.7660**
	RetSam_medoid	–	0.3180	**0.5805**	0.7466
	RetSam_midpoint	–	0.2980	0.5710	0.7410

Table 2. Linear supervised evaluation results across different datasets and methods. Best results are shown in **bold**, and second-best results are underlined.

Dataset	Método	Accuracy	Recall	Precision	F1
SVHN	SimCLR	0.7954	0.7803	0.7905	0.7848
	SimSiam	0.1959	0.1000	0.0196	0.0328
	BarlowTwins	0.8145	0.8145	0.8148	0.8144
	BYOL	0.8030	0.7873	0.8014	0.7936
	truncated	0.8140	0.8130	0.8100	0.8140
	RetSam	0.8423	0.8278	0.8324	0.8294
	RetSam_centroid	**0.8772**	**0.8706**	**0.8740**	**0.8721**
	RetSam_medoid	0.8423	0.8288	0.8329	0.8301
	RetSam_midpoint	0.8320	0.8290	0.8312	0.8301
CIFAR10	SimCLR	0.9010	0.9010	0.9010	0.9009
	SimSiam	0.8579	0.8579	0.8653	0.8571
	BarlowTwins	0.8306	0.8135	0.8222	0.8162
	BYOL	0.9089	0.9089	0.9089	0.9088
	truncated	0.8790	0.8790	0.8530	0.8651
	RetSam	**0.9361**	**0.9361**	**0.9362**	**0.9360**
	RetSam_centroid	0.9337	0.9337	0.9337	0.9337
	RetSam_medoid	0.9353	0.9353	0.9353	0.9352
	RetSam_midpoint	0.9100	0.9080	0.9000	0.9040
STL10	SimCLR	0.8679	0.8679	0.8684	0.8679
	SimSiam	0.8625	0.8625	0.8634	0.8625
	BarlowTwins	0.1961	0.1961	0.3814	0.1655
	BYOL	0.8662	0.8663	0.8670	0.8663
	truncated	0.9450	0.9450	0.9456	0.9450
	RetSam	0.9441	0.9441	0.9443	0.9441
	RetSam_centroid	0.9449	0.9449	0.9451	0.9449
	RetSam_medoid	**0.9456**	**0.9456**	**0.9457**	**0.9456**
	RetSam_midpoint	0.9320	0.9300	0.9310	0.9305
Food3	SimCLR	0.8000	0.7919	0.7889	0.7901
	SimSiam	0.5200	0.5192	0.5588	0.5266
	BarlowTwins	0.3467	0.2889	0.2322	0.2192
	BYOL	0.4400	0.4434	0.4355	0.4350
	truncated	0.5867	0.5710	0.5704	0.5605
	RetSam	0.7600	0.7493	0.7473	0.7452
	RetSam_centroid	**0.8267**	**0.8056**	**0.8294**	**0.8093**
	RetSam_medoid	0.8133	0.8042	0.8066	0.8025
	RetSam_midpoint	0.8000	0.7920	0.7900	0.7910

Table 3. Results of the supervised evaluation for different datasets and methods using 10% of labeled data. Best results are shown in **bold**, and second-best results are underlined.

Dataset	Método	Accuracy	Recall	Precision	F1
	SimCLR	0.7112	0.7042	0.6963	0.6963
	SimSiam	0.3117	0.2537	0.4553	0.1651
	BarlowTwins	0.7788	0.7788	0.7788	0.7783
	BYOL	0.7388	0.7346	0.7252	0.7269
SVHN	truncated	0.7350	0.7200	0.7130	0.7360
	RetSam	0.8134	0.8012	0.7948	0.7960
	RetSam_centroid	**0.8450**	**0.8447**	**0.8355**	**0.8391**
	RetSam_medoid	0.8116	0.8014	0.7934	0.7954
	RetSam_midpoint	0.7820	0.7790	0.7690	0.7739
	SimCLR	0.7966	0.7966	0.7360	0.7607
	SimSiam	0.4493	0.4493	0.4936	0.4162
	BarlowTwins	0.6939	0.6403	0.5364	0.5802
	BYOL	0.8898	0.8898	0.8904	0.8896
CIFAR10	truncated	0.4930	0.4930	0.4960	0.4620
	RetSam	**0.9246**	**0.9246**	**0.9248**	**0.9244**
	RetSam_centroid	0.9126	0.9126	0.9131	0.9124
	RetSam_medoid	0.9242	0.9242	0.9243	0.9241
	RetSam_midpoint	0.8920	0.8860	0.8900	0.8880
	SimCLR	0.8199	0.8199	0.8228	0.8197
	SimSiam	0.7430	0.7430	0.6819	0.7064
	BarlowTwins	0.1000	0.1000	0.0100	0.0182
	BYOL	0.8264	0.8264	0.8275	0.8262
STL10	truncated	0.9113	0.9112	0.9175	0.9108
	RetSam	**0.9213**	**0.9213**	**0.9222**	**0.9213**
	RetSam_centroid	0.9101	0.9101	0.9114	0.9103
	RetSam_medoid	0.9197	0.9197	0.9207	0.9198
	RetSam_midpoint	0.9000	0.9020	0.8990	0.9005
	SimCLR	**0.7200**	0.6788	**0.7736**	0.6894
	SimSiam	0.3467	0.3180	0.3227	0.3160
	BarlowTwins	0.3467	0.3441	0.2808	0.1903
	BYOL	0.4400	0.4214	0.4388	0.4246
Food3	truncated	0.5067	0.4920	0.4930	0.4722
	RetSam	0.7067	**0.7088**	0.7027	**0.7018**
	RetSam_centroid	0.4400	0.4602	0.5185	0.4114
	RetSam_medoid	0.6133	0.6232	0.6209	0.6102
	RetSam_midpoint	0.5820	0.5850	0.5840	0.5845

Table 4. Results of the supervised evaluation for different datasets and methods using 5% of labeled data. Best results are shown in **bold**, and second-best results are underlined.

Dataset	Método	Accuracy	Recall	Precision	F1
SVHN	SimCLR	0.6777	0.6682	0.6611	0.6604
	SimSiam	0.0871	0.1092	0.1644	0.0332
	BarlowTwins	0.7639	0.7639	0.7643	0.7637
	BYOL	0.7117	0.7054	0.7020	0.6997
	truncated	0.7100	0.7280	0.7160	0.7130
	RetSam	0.7931	0.7758	0.7730	0.7720
	RetSam_centroid	**0.8301**	**0.8299**	**0.8172**	**0.8227**
	RetSam_medoid	0.7957	0.7793	0.7766	0.7760
	RetSam_midpoint	0.7820	0.7790	0.7860	0.7825
CIFAR10	SimCLR	0.8155	0.8155	0.8452	0.7764
	SimSiam	0.2291	0.2291	0.3199	0.1370
	BarlowTwins	0.7117	0.7064	0.6190	0.6577
	BYOL	0.8848	0.8848	0.8844	0.8845
	truncated	0.2910	0.2910	0.3760	0.3850
	RetSam	0.9186	0.9186	0.9181	0.9181
	RetSam_centroid	0.9073	0.9073	0.9068	0.9068
	RetSam_medoid	**0.9196**	**0.9196**	**0.9190**	**0.9191**
	RetSam_midpoint	0.8870	0.8800	0.8790	0.8795
STL10	SimCLR	0.8057	0.8057	0.8076	0.8057
	SimSiam	0.7570	0.7570	0.7860	0.7303
	BarlowTwins	0.1334	0.1334	0.0501	0.0708
	BYOL	0.8099	0.8099	0.8114	0.8097
	truncated	0.8077	0.8078	0.8464	0.7972
	RetSam	0.9125	0.9125	**0.9135**	0.9125
	RetSam_centroid	0.8990	0.8990	0.9008	0.8991
	RetSam_medoid	**0.9126**	**0.9126**	**0.9135**	**0.9126**
	RetSam_midpoint	0.8900	0.8820	0.8800	0.8810
Food3	SimCLR	0.2933	0.2757	0.2544	0.2611
	SimSiam	0.2800	0.3254	0.2783	0.2585
	BarlowTwins	0.3333	0.3333	0.1111	0.1667
	BYOL	0.2933	0.3202	0.3105	0.2579
	truncated	0.4667	0.4290	0.6254	0.3782
	RetSam	0.4800	0.4346	**0.3479**	0.3562
	RetSam_centroid	0.4800	0.4414	0.3249	0.3598
	RetSam_medoid	**0.4933**	**0.4861**	0.3453	**0.4015**
	RetSam_midpoint	0.4700	0.4760	0.3350	0.3804

5 Conclusion

The research advances self-supervised representation learning through an innovative approach to generating negative samples using sophisticated geometric

methods. By introducing three novel techniques—median centroids, geometric midpoints, and medoids—the study demonstrates a significant improvement in content-based image retrieval and classification across diverse datasets. These geometric approaches prove particularly powerful in content-based information retrieval, where the ability to accurately match and retrieve similar images relies on the precision of learned feature representations.

In the domain of CBIR, the proposed methods show remarkable performance, consistently outperforming existing techniques across different datasets and retrieval scales. The approach enables more accurate similarity searches, allowing systems to more effectively find and rank relevant images based on their learned feature representations. This is especially critical in applications ranging from digital archives and medical image databases to e-commerce product search and visual recommendation systems.

Beyond retrieval, the methods also demonstrate strong capabilities in classification tasks, particularly in scenarios with limited labeled data. By creating more informative negative samples, the approach helps machine learning models develop more robust and discriminative feature representations. This makes the research particularly valuable for domains where obtaining large labeled datasets is challenging or expensive, such as specialized medical imaging, rare species identification, or niche industrial applications.

Perhaps most importantly, the study offers a promising path forward for machine learning in low-resource environments. By demonstrating the effectiveness of carefully constructed negative sampling strategies, the research provides valuable insights into how we can extract meaningful representations even when labeled data is scarce. This work not only advances current understanding of representation learning but also opens up new avenues for more intelligent and adaptive machine learning approaches in content-based retrieval and classification tasks.

References

1. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: European Conference on Computer Vision (2014)
2. Cai, T.T., Frankle, J., Schwab, D.J., Morcos, A.S.: Are all negatives created equal in contrastive instance discrimination? arXiv preprint arXiv:2010.06682 (2020)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
4. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems **33**, 22243–22255 (2020)
5. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
6. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)

7. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 215–223. JMLR Workshop and Conference Proceedings (2011)
8. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
9. Ding, S., Lin, L., Wang, G., Chao, H.: Deep feature learning with relative distance comparison for person re-identification. Pattern Recognition **48**(10), 2993–3003 (2015)
10. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE international conference on computer vision. pp. 1422–1430 (2015)
11. Doersch, C., Zisserman, A.: Multi-task self-supervised visual learning. In: Proceedings of the IEEE international conference on computer vision. pp. 2051–2060 (2017)
12. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. Advances in neural information processing systems **27** (2014)
13. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018)
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
15. Goyo, M., Frisoni, G., Moro, G., Sartori, C.: Refining triplet sampling for improved self-supervised representation learning (2024)
16. Goyo, M.A., Hidalgo, M.: Negative sampling for triplet-based loss: Improving representation in self-supervised representation learning. In: Hernández-García, R., Barrientos, R.J., Velastin, S.A. (eds.) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. pp. 133–147. Springer Nature Switzerland, Cham (2025)
17. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
18. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 297–304. JMLR Workshop and Conference Proceedings (2010)
19. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
20. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person reidentification. arXiv preprint arXiv:1703.07737 (2017)
21. Howard, A.G.: Some improvements on deep convolutional neural network based image classification. arXiv preprint arXiv:1312.5402 (2013)
22. Huynh, T., Kornblith, S., Walter, M.R., Maire, M., Khademi, M.: Boosting contrastive self-supervised learning with false negative cancellation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2785–2795 (2022)
23. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

24. Kolesnikov, A., Zhai, X., Beyer, L.: Revisiting self-supervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1920–1929 (2019)
25. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep. (2009)
26. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. pp. 577–593. Springer (2016)
27. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
28. Li, W., Yang, X., Kong, M., Wang, L., Huo, J., Gao, Y., Luo, J.: Trip-roma: Self-supervised learning with triplets and random mappings. Transactions on Machine Learning Research (2022)
29. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011 (2011)
30. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision. pp. 69–84. Springer (2016)
31. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4004–4012 (2016)
32. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016)
33. Robinson, J., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. arXiv preprint arXiv:2010.04592 (2020)
34. Schroff, F., Kalenichenko, D., Philbin, J.: 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)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
36. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
37. Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning? Advances in neural information processing systems **33**, 6827–6839 (2020)
38. Turpault, N., Serizel, R., Vincent, E.: Semi-supervised triplet loss based learning of ambient audio embeddings. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 760–764. IEEE (2019)
39. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning. pp. 1096–1103 (2008)
40. Wang, G., Wang, G., Zhang, X., Lai, J., Yu, Z., Lin, L.: Weakly supervised person re-id: Differentiable graphical learning and a new benchmark. IEEE Transactions on Neural Networks and Learning Systems **32**(5), 2142–2156 (2020)

41. Wang, G., Wang, K., Wang, G., Torr, P.H., Lin, L.: Solving inefficiency of self-supervised representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9505–9515 (2021)
42. Wang, X., Zhang, H., Huang, W., Scott, M.R.: Cross-batch memory for embedding learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6388–6397 (2020)
43. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018)
44. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International Conference on Machine Learning. pp. 12310–12320. PMLR (2021)
45. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. pp. 649–666. Springer (2016)

A Implementation Details

A.1 Hardware Configuration

The experiments were carried out on a computer with the following specifications: Intel(R) Core(TM) i7-8700K CPU @ 3.70GHz, 32GB of RAM, and a GeForce GTX 1080 Ti GPU.

A.2 Selected Hyperparameters

In Table 5, a comprehensive list of all the hyperparameters utilized for our methods is provided. These hyperparameters are pivotal components in configuring and fine-tuning the performance of our methodologies. Each hyperparameter plays a distinct role in shaping the behavior and efficacy of the employed techniques. Through meticulous selection and optimization of these hyperparameters, we aim to enhance the overall performance and robustness of our methods across various experimental settings and datasets.

A.3 Dataset details

Additional information about the datasets is presented in the Table 6. It is important to note that these two datasets represent very different natures; one consists of natural images while the other is composed solely of numbers. The combination of both sets is essential for a comprehensive evaluation of the performance of different data sets.

A.4 Comparative Analysis of Geometric Approaches:

Figures 6-21 illustrate the relative performance of our different geometric approaches across tasks and datasets vs our previous proposal. Median Centroid

Table 5. Selected Hyperparameters

Hyperparameter	**Selected Value**
Learning Rate	0.06
Epochs	200
Batch Size	32
Decay Schedule	cosine learning rate
Optimizer	SGD
Encoder	VGG50(weights="imagenet")
k- neighbors	15
m (margin)	0.6
α	0.5

Table 6. Dataset Information

Type	**Name**	**Train**	**Test**	**N° Classes**
Natural Image	Cifar-10	50000	10000	10
Numbers	Street View House Number	73257	26032	10
Natural Image	Self-Taught Learning 10	5000	8000	10
Food3	The Food-101 restriction	225	75	3

Approach provides the most consistent performance improvements across diverse conditions, particularly excelling on complex datasets with potential outliers. Its robustness to outliers in feature space makes it especially valuable for real-world applications with noisy data distributions. On the other hand, the Medoid Approach demonstrates exceptional performance on well-structured datasets like CIFAR-10 and STL-10, particularly in scenarios with limited labeled data. By selecting actual data points as representatives, this approach maintains the integrity of the underlying data manifold. And finally Geometric Midpoint Approach. While generally performing slightly below the other geometric methods, it still consistently outperforms baseline approaches, offering a computationally efficient alternative that captures meaningful feature relationships.

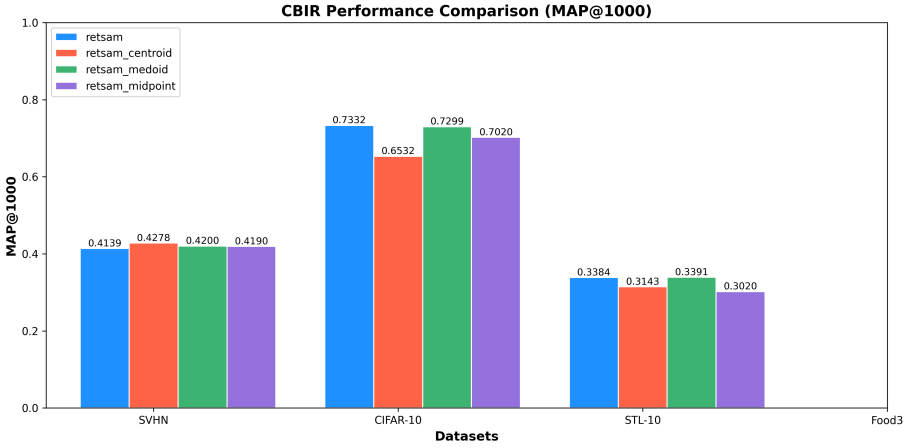

Fig. 6. CBIR performance measured by MAP@1000 for our different geometric approaches and datasets, compared to our previous method.

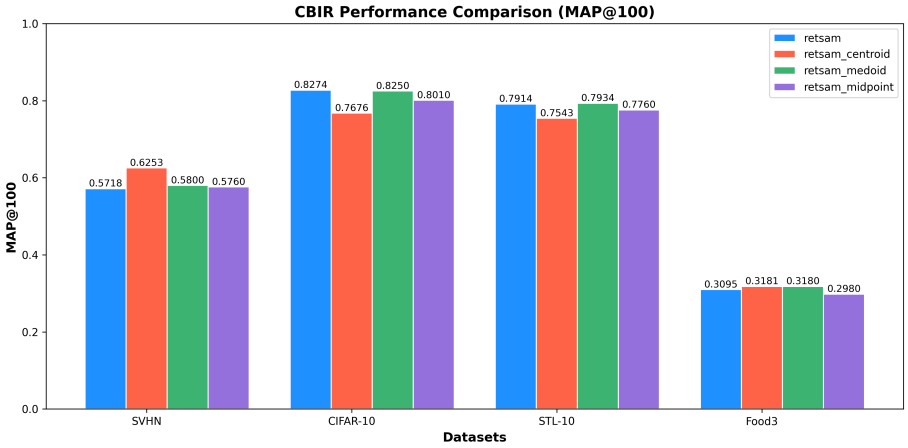

Fig. 7. CBIR performance measured by MAP@100 for our different geometric approaches and datasets, compared to our previous method.

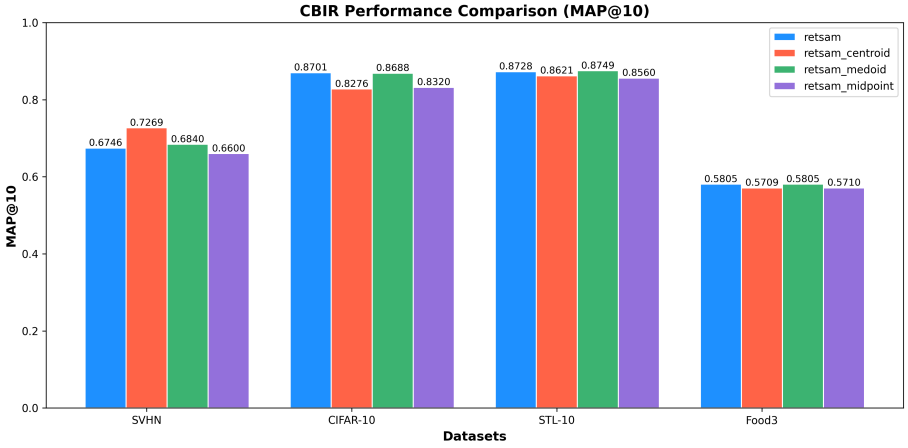

Fig. 8. CBIR performance measured by MAP@10 for our different geometric approaches and datasets, compared to our previous method.

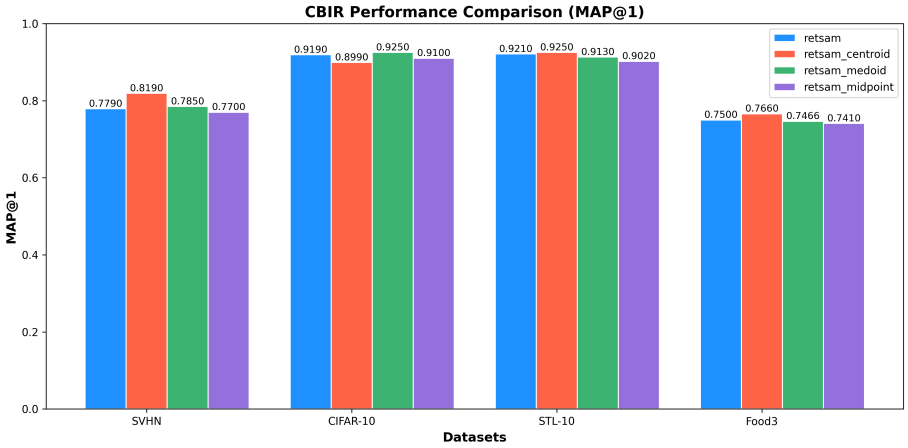

Fig. 9. CBIR performance measured by MAP@1 for our different geometric approaches and datasets, compared to our previous method.

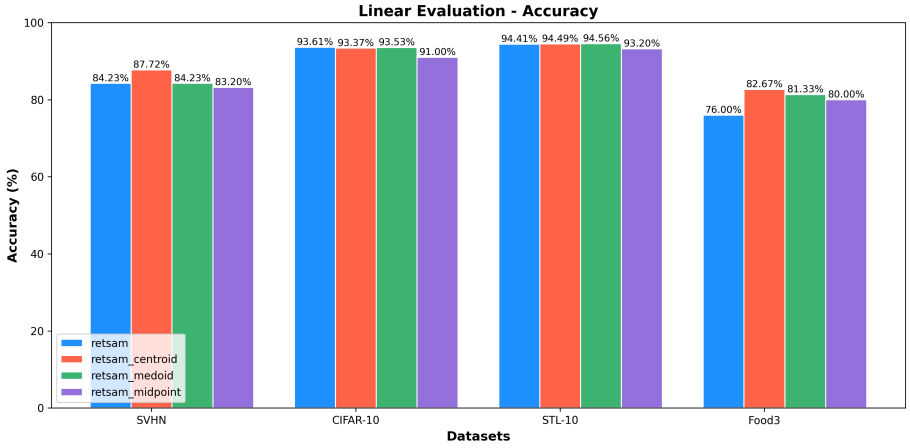

Fig. 10. Accuracy of linear evaluation for our different geometric approaches and datasets, compared to our previous method.

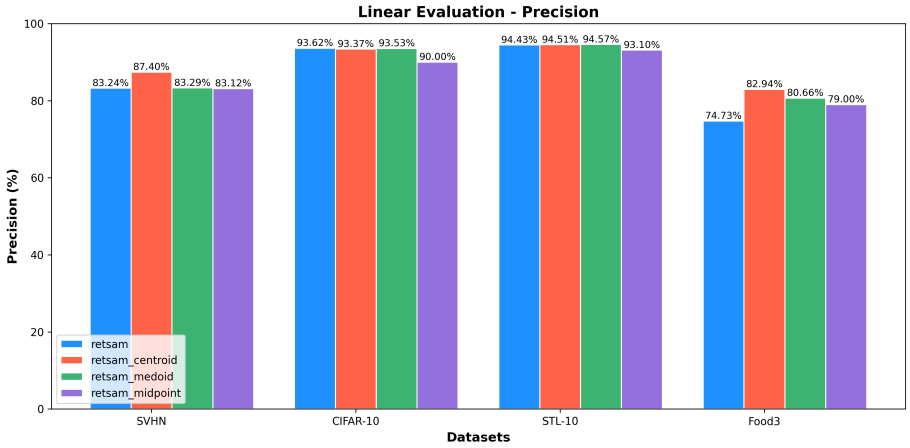

Fig. 11. Precision of linear evaluation for our different geometric approaches and datasets, compared to our previous method.

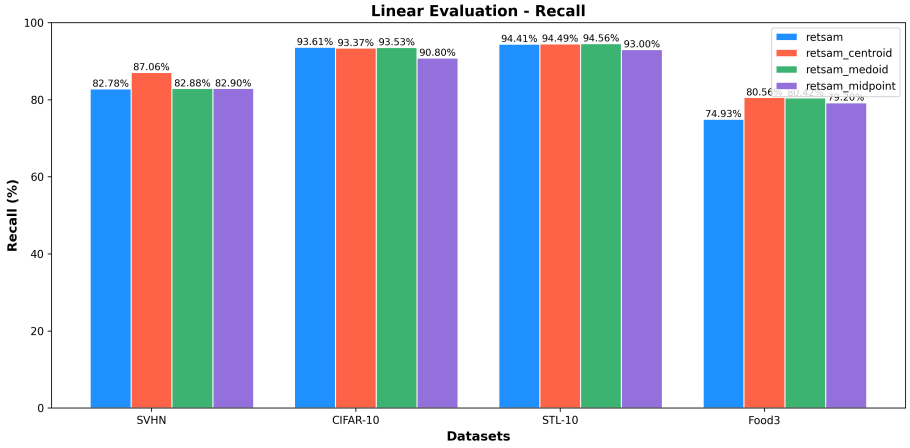

Fig. 12. Recall of linear evaluation for our different geometric approaches and datasets, compared to our previous method.

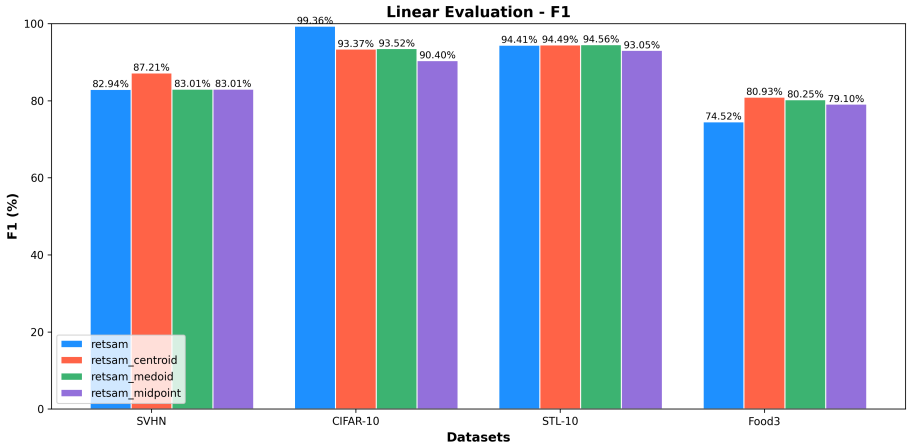

Fig. 13. F1-score of linear evaluation for our different geometric approaches and datasets, compared to our previous method.

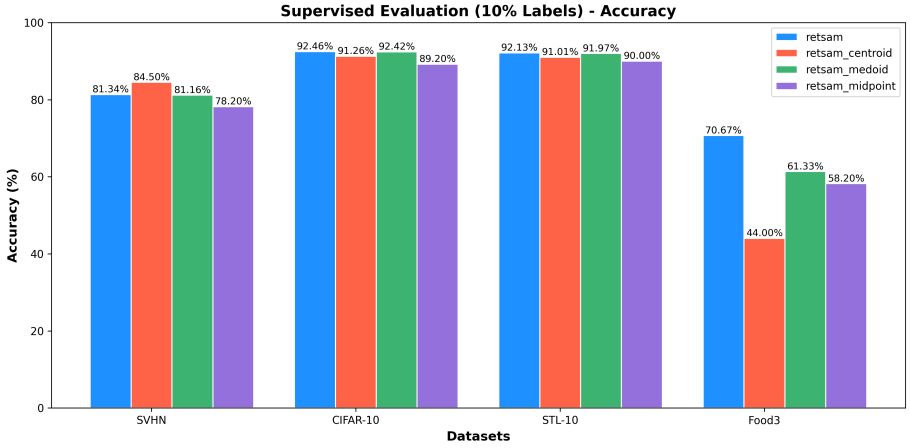

Fig. 14. Accuracy of supervised evaluation using 10% of labeled data for our different geometric approaches and datasets, compared to our previous method.

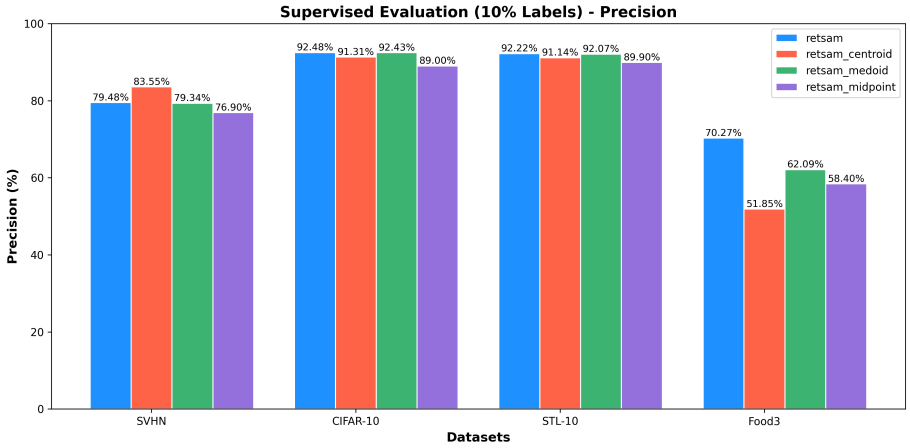

Fig. 15. Precision of supervised evaluation using 10% of labeled data for our different geometric approaches and datasets, compared to our previous method.

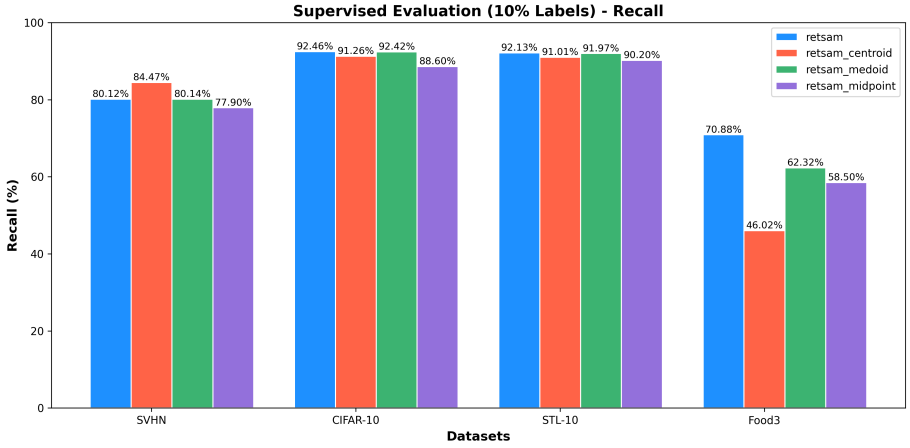

Fig. 16. Recall of supervised evaluation using 10% of labeled data for our different geometric approaches and datasets, compared to our previous method.

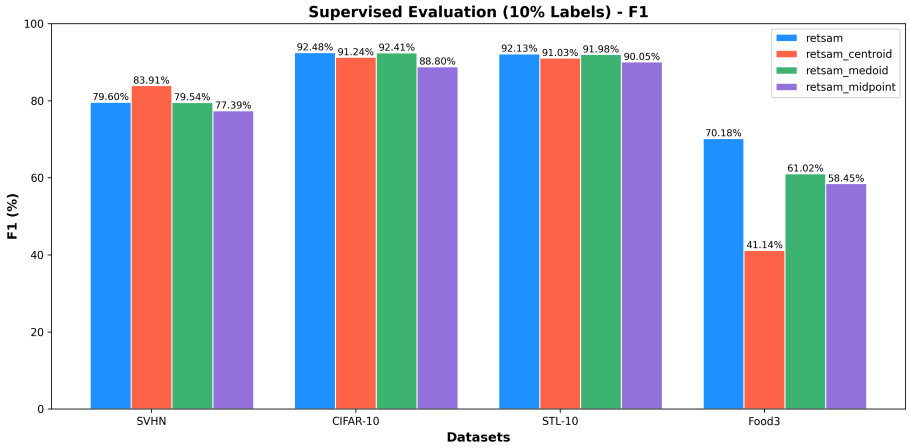

Fig. 17. F1-score of supervised evaluation using 10% of labeled data for our different geometric approaches and datasets, compared to our previous method.

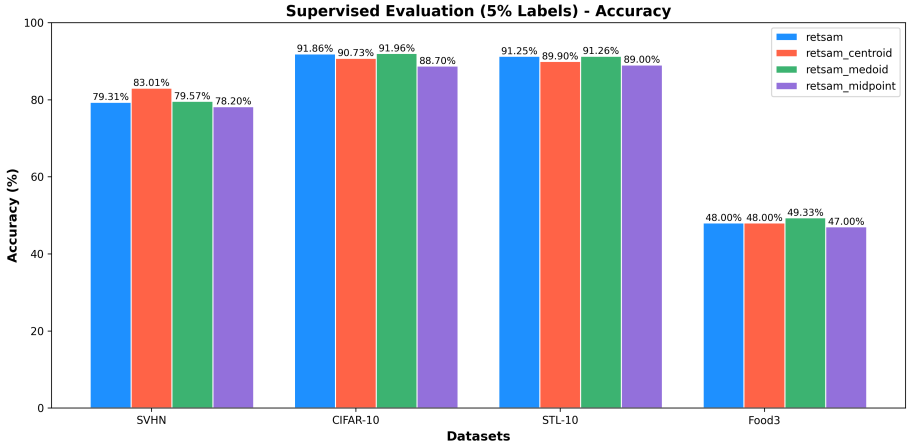

Fig. 18. Accuracy of supervised evaluation using 5% of labeled data for our different geometric approaches and datasets, compared to our previous method.

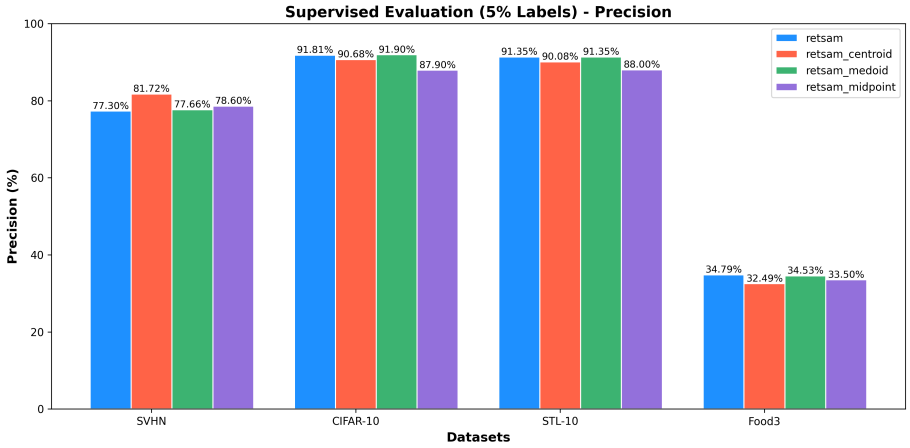

Fig. 19. Precision of supervised evaluation using 5% of labeled data for our different geometric approaches and datasets, compared to our previous method.

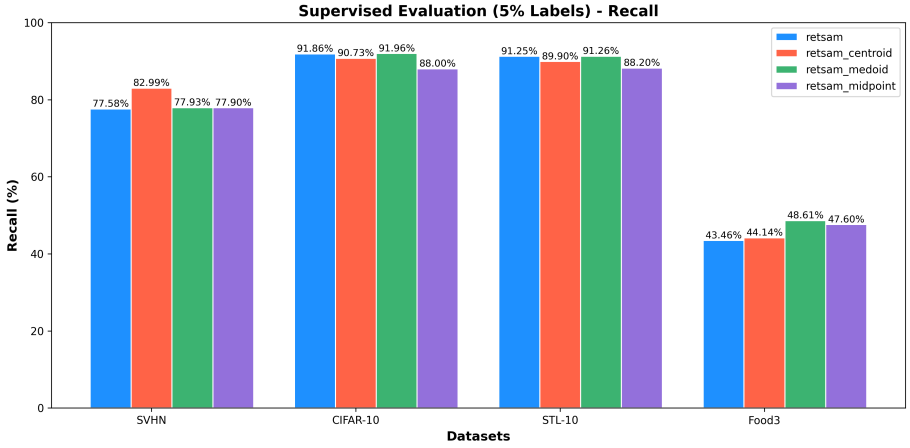

Fig. 20. Recall of supervised evaluation using 5% of labeled data for our different geometric approaches and datasets, compared to our previous method.

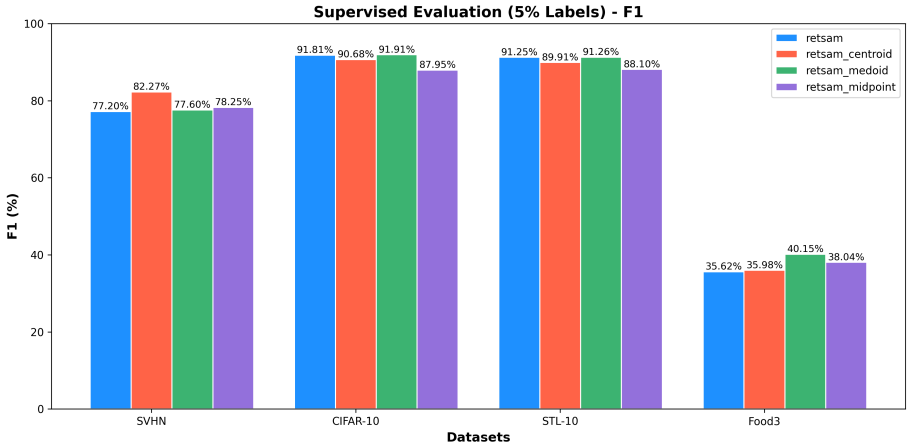

Fig. 21. F1-score of supervised evaluation using 5% of labeled data for our different geometric approaches and datasets, compared to our previous method.