Refining Triplet Sampling for Improved Self-Supervised Representation Learning[⋆]

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

[1]*Department of Informatics, Universidad Técnica Federico Santa María, Valparaíso, Chile*

[2]*Department of Computer Science and Engineering, University of Bologna, Bologna, Italy*

Abstract

Self-supervised representation learning extracts meaningful features from data without explicit supervision, building a space with desired properties. Contrastive learning has emerged as the predominant approach to clustering similar data points and separating dissimilar ones within the embedding space. Although creating different views of the same data (e.g., cropping, rotation) emphasizes similarities without labels, current methods struggle to define negative examples. Several algorithms only consider positive examples or integrate dissimilarity measures into their loss functions by computing average distances within the same batch. However, they do not capture nuanced differences effectively, risking collapsing data points in a single location. In this paper, we propose a novel technique, termed "Refined Triplet Sampling" (ReTSam), to generate synthetic negative vectors for contrastive learning. Mechanically, for each element in the batch, we identify its k-nearest neighbors and designate the centroid as a hard negative for a triplet loss methodology. We test ReTSam on two widely used image datasets, namely CIFAR-10 and SVHN, considering content-based image retrieval and classification tasks. Our findings demonstrate that, despite its simplicity, ReTSam not only promotes the learning of similarity but also significantly improves that of dissimilarity (with a +5% increase in Mean Average Precision on CIFAR10), resulting in superior performance in practical scenarios.

Keywords

Self Supervised Learning, Representation Learning, Triplet Loss, Negative Sampling

1. Introduction

Lately, representation learning has become a crucial element in the development of modern AI agents, largely propelled by significant advancements in self-supervised learning (SSL). SSL is a paradigm where representations are obtained through pre-training tasks using unlabeled data, playing a pivotal role in contemporary AI. These acquired representations are then utilized in subsequent tasks like classification or content-based retrieval of images. Importantly, the attractiveness of SSL stems from its capability to leverage abundant and cost-effective unlabeled data, often surpassing its supervised counterpart, as observed in certain instances [1, 2]. Many contrastive learning approaches hinge 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 [3], directs the learned representation to map

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy

[*] Corresponding author.

[†] These authors contributed equally.

manuel.goyo@sansano.usm.cl (M. Goyo); giacomo.frisoni@unibo.it (G. Frisoni); gianluca.moro@unibo.it (G. Moro); claudio.sartori@unibo.it (C. Sartori)

positive pairs to close locations and negative pairs to distant ones. Alternative objectives have also been explored [4]. The effectiveness of these methods relies on the formulation of information for the positive and negative pairs, as they cannot leverage genuine similarity information due to the absence of supervision. Certain authors opt not to explicitly generate dissimilar data. Instead, they compute distances to all other data points [4] or their closest neighbors [5], calculate the average of these similarities, and use it as a dissimilarity measure in a loss function. However, the drawback of this approach lies in the inadequacy of the average to effectively represent dissimilarity. An alternative approach addresses the issue by focusing solely on positive instances and implementing diverse parameter updates [6, 7]. Nevertheless, this method fails to endow the algorithm with the capability to construct a robust decision boundary for effectively discerning differences within the data, leading to overlaps with different categories. Some authors pursue explicit negatives by considering different views (augmentations) for each image to identify real negatives and discard false negatives [8] or by estimating a sample from the distribution over negative pairs [9]. This approach stems from metric learning settings, where "hard" (true negative) examples can expedite the correction of mistakes in the learning process [10, 11]. In representation learning, informative negative examples are intuitively those pairs that are mapped nearby but should be far apart. This concept is successfully applied in metric learning, where true pairs of dissimilar points are available, in contrast to unsupervised contrastive learning. Our methodology hinges on the generation of a hard negative, inspired by the findings of Cai et al. (2020) [12], who assert that "... a small minority of negatives were both necessary and sufficient for the downstream task to reach full accuracy." In light of this insight, we propose an approach centered around triplet loss. In this setup, the positive pairs are generated in a conventional manner, employing transformations that preserve semantic content. However, the negative element is uniquely crafted considering only the k nearest neighbors of the remaining batch of positives to the anchor. The negative is then derived by computing the centroid. This approach emphasizes that the centroid serves as an excellent representation of the negative, owing to its ability to encapsulate information from all vectors in close proximity to the anchor.

Particularly, the main contributions of this work are as follows:

- We design a simple but effective sampling strategy based on similarity to create negative elements.
- We propose a general self-supervised training method based on triplet loss for representation learning.
- We are the first to evaluate state-of-the-art self-supervised algorithms in the context of Content-Based Image Retrieval (CBIR) in different datasets.
- Our experiments across two datasets demonstrate that our approach surpasses existing methods in both Content-Based Image Retrieval (CBIR) and classification tasks, as indicated by superior performance metrics such as Mean Average Precision (MAP) for CBIR and Accuracy, Recall, Precision, and F1 for classification.

The rest of the paper is organized as follows. Section 2 presents a review of the work related to this approach. In section 3, we will describe our proposed method. Section 4 will show the results of applying our method to different datasets. Section 5 will present conclusions and future works.

2. Related Works

2.1. Representation Learning

In the realm of unsupervised representation learning, the approaches are predominantly categorized into generative and discriminative methods [13, 4]. Generative strategies involve constructing a distribution over data and latent embeddings, utilizing these embeddings as representations for images. Techniques such as auto-encoding of images [14, 15] and adversarial learning [16] are commonly employed in generative methods. While these approaches provide comprehensive pixel-level representations, the computational demands can be significant, and the generation of highly detailed images may not be essential for effective representation learning. Discriminative methods, particularly contrastive methods [4, 6, 5], currently stand at the forefront, showcasing state-of-the-art performance in self-supervised learning. Some alternative methodologies opt for auxiliary handcrafted prediction tasks to guide representation learning. However, their efficacy often falls short in comparison to contrastive methods. Noteworthy techniques, such as relative patch prediction [13, 17], colorizing grayscale images [18, 19], image inpainting [20], image jigsaw puzzle [21], image super-resolution [22], and geometric transformations [23, 24], have been explored for their utility. Despite the integration of well-structured architectures [25], these approaches consistently underperform when juxtaposed with the superior performance demonstrated by contrastive methods [26, 27].

2.2. Contrastive Learning

Contrastive learning stands as a compelling alternative to the computationally intensive task of pixel-level image generation. Shifting its focus from image creation, contrastive learning aims to minimize the distance between representations of different views of the same image (positive pairs) and maximize the distance between representations of views from different images (negative pairs) [17, 28, 6]. Contrastive methods often capitalize on comparisons with multiple examples, and in some cases, they exhibit effectiveness even without explicit negative examples [4, 5, 7]. Several noteworthy algorithms have been proposed for contrastive learning of visual representations. SimCLR [4], for instance, utilizes augmented views of other items in a minibatch as negative samples. MoCo [1, 26], on the other hand, incorporates a momentum-updated memory bank of old negative representations, enabling the use of large batches of negative samples. Tri Huynh et al. [8] tackle a fundamental issue in contrastive learning—the mitigation of false negatives. The introduction of false negatives poses challenges such as discarding semantic information and slow convergence. The authors propose novel approaches to identify false negatives, introducing two strategies—false negative elimination and attraction—to mitigate their effects. Their work involves systematic evaluations to comprehensively understand and address this issue. Robinson et al. [9] present an unsupervised method based on a simple distribution over hard negative pairs for contrastive representation learning. They construct this distribution over hard negatives with the assumption that the most useful negative samples are those that the embedding currently believes to be similar to the anchor. A noteworthy approach to learning image representation is introduced by [29]. This involves computing the cross-correlation matrix between the outputs of two identical networks, which receive distorted versions of a sample. The objective is to make this cross-correlation matrix as similar to the identity matrix as possible. This ensures that the embedding vectors of the distorted versions of a sample become more similar to each other while reducing redundancy among the components of these vectors.

2.3. Triplet Loss Approach

The triplet loss approach, initially introduced by Ding et al. for person re-identification and independently adopted by Schroff et al. for face recognition [30, 10], has undergone substantial evolution, becoming a transformative paradigm in contrastive learning. In building upon the foundational concept of triplet loss, researchers have dedicated efforts to enhance the generation and selection of valuable triplets. Hermans et al. [31] contributed significant strategies to identify and leverage informative triplets, thereby bolstering the robustness and effectiveness of the triplet loss methodology. Seeking further refinement, Wang et al. [32] delved into the application of cross-batch triplet loss, with the objective of augmenting generalization capabilities and stabilizing the triplet loss approach. This extension demonstrates a nuanced understanding of inter-batch relationships and their pivotal role in shaping the learning process. Furthermore, researchers have ventured into adapting the triplet loss approach to weakly supervised scenarios. Wang et al. [33] made notable contributions in this domain, exploring methods to harness weak supervision signals and extend the applicability of the triplet loss paradigm to scenarios where labeled data may be scarce. Turpault et al. [34] took a unique approach by integrating unsupervised triplet loss-based learning into a self-supervised representation learning framework. Their variant involves obtaining positive samples for triplets with unlabeled anchors by applying a transformation to the anchor. The negative sample for these triplets is then chosen as the sample in the training set that is closest to the anchor and distant from the positive sample. Another noteworthy contribution to the triplet loss approach comes from Wang et al. [5], who introduced a truncated triplet loss methodology. In their approach, the negative pair is constructed by selecting a negative sample deputy from all negative samples. This strategic choice aims to mitigate false negatives and prevent the model from over-clustering samples of the same actual categories into different clusters. Finally, Li et al. [35] introduce an algorithm called Trip-ROMA, based on a simple Triplet loss with RandOm MApping (ROMA) strategy, which consists of mapping random samples into other spaces and requiring these randomly projected samples to satisfy the same relationship indicated by the triplets. Finally, integrating the triplet-based loss with random mapping, we obtain the proposed method.

3. Algorithm

We first are going to show a motivation and then we present the algorithm.

3.1. Motivation

In the past year, the prominence of Self-Supervised Representation Learning has experienced significant growth, primarily driven by the challenges posed by the absence of labeled data. A prevalent strategy involves applying augmentations to generate different views of the same data, effectively emphasizing similar or closely related data points [36] (see Fig. 1).

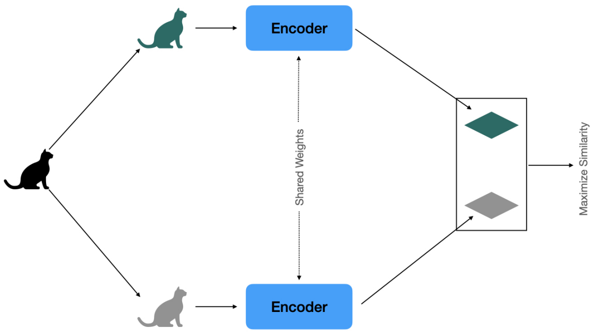


Figure 1: schema self-supervised with augmentation strategy

However, a critical challenge emerges in creating dissimilar data, as failure to do so may lead to a collapsing solution where all data points cluster at a single location [13]. Addressing the challenge of dissimilar data, some authors calculate distances to all other data points [4] or their closest neighbors [5], computing the average of these similarities and using it as a dissimilarity measure in a loss function. Nevertheless, the inadequacy of the average to effectively represent dissimilarity poses a drawback to this approach, so requires a large batch size. An alternative method tackles the issue by solely considering positive instances and implementing diverse parameter updates [6, 7]. However, this method falls short in enabling the algorithm to construct a robust decision boundary for effectively discriminating differences within the data, leading to overlaps with different categories. The crux of our motivation lies in selecting a robust representation of the negative within the data (hard negative). This representation should effectively challenge the model in differentiating it from the positive. Leveraging the triplet loss approach, commonly employed in contrastive learning for SSL, becomes a natural choice, in the Fig. 2, you can see the schema.

Triplet loss, introduced independently for various applications such as person re-identification and face recognition [30, 10], deals with sets comprising an anchor sample, a positive sample, and a negative sample. The loss function encourages the model to maximize the similarity between the anchor and positive samples while minimizing the similarity between the anchor and negative samples, subject to a margin constraint.

For simplicity, we illustrate the triplet set $(x_i, x_i^+, x_i^-)_{i=1,\cdots,m}$ using one query data and one sample. The triplet loss is defined as

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

where *sim* is a similarity metric (e.g., cosine similarity or Euclidean), and *m* is a margin determining whether to discard a triplet

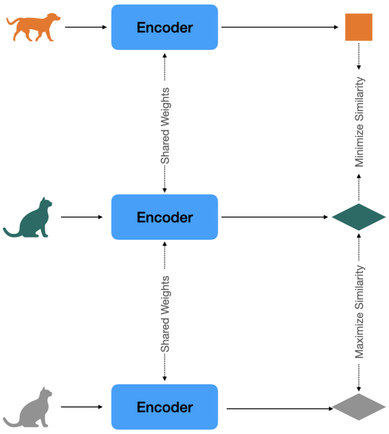


Figure 2: schema triplet loss approach

Constructing triplets for each data point poses a significant challenge, particularly in determining how to establish negative pairs accurately (dog in Fig. 2). While positive pairs can be reliably generated, identifying negative pairs involves the use of hard negative samples (points that are challenging to distinguish from an anchor point). The key challenge lies in utilizing hard negatives while remaining unsupervised, precluding the adoption of existing negative sampling strategies that rely on true similarity information.

3.2. Proposed Methodology

To overcome the challenge of creating dissimilar data and to enhance the effectiveness of the triplet loss approach, we draw inspiration from the work of Cai et al. [12]. Their findings suggest that only a small quantity of negatives is necessary for achieving full accuracy in downstream tasks. In our proposed method, we introduce a novel approach for generating negative values within a triplet set.

In this approach, the anchor represents one view of the data, and the positive is derived from the other view of the same data within a batch. Crucially, the negative is constructed by searching for the k nearest neighbors of the anchor among the positive ones. The negative value is obtained by calculating the centroid of these k vectors. This vector serves as an excellent representation of the negative since it combines elements of the negative data with characteristics of the positive data, effectively building a hard negative. This is attributed to its ability to encapsulate information from all vectors in close proximity to the anchor. Consequently, the centroid poses a challenge when differentiating it from the anchor, thereby enhancing the discriminative capability of the model.

Mathematically, the triplet loss is expressed as:

$$\mathscr{L}_1(z_a, z_p, z_n) = \max\left(\text{sim}(z_a, z_n) - \text{sim}(z_a, z_p) + m, 0\right) \tag{1}$$

Here, $z_a = f(\tau_1(x))$ represents the anchor, $z_p = f(\tau_2(x))$ represents the positive, with f

denoting an encoder neural network, and τ_1, τ_2 drawn from the set T of augmentation transform techniques, and sim() indicate a similarity measure between two vectors (cosine similarity for default). The i-th element of z_n is computed as $z_n[i]$ = Centroid(k-nearest(z_{p-i})), where Centroid denotes the centroid function, k-nearest(z_{p-i}) represents the k-elements closest to z_p excluding the i-th element.

Typically, the triplet loss is constrained by its sensitivity to the training triplets due to its reliance on a set margin [37]. Consequently, the cross-entropy loss serves as a more flexible alternative, resembling a softer version of the triplet loss with an adjustable margin [35]. This adaptation addresses the constraint of the triplet loss with a fixed margin.

$$\mathscr{L}_2(z_a, z_p, z_n) = -\log \frac{\exp\left(z_a^\top z_p\right)}{\exp\left(z_a^\top z_p\right) + \exp\left(z_a^\top z_n\right)} \tag{2}$$

Finally, the total loss function is defined as:

$$\mathscr{L}_{\text{oss}} = \mathbb{E}_x\left(\mathscr{L}_1(x_a, x_p, x_n) + \alpha\mathscr{L}_2(z_a, z_p, z_n)\right) \tag{3}$$

This proposed solution addresses the limitations of existing methods by introducing a more effective way of constructing negative representations, thereby aiming to enhance overall performance in representation learning, particularly in Content-Based Image Retrieval and Classification tasks.

4. Main Results

We are going to present the protocols to train our algorithm and our results:

4.1. Training

Data Augmentation: One type of augmentation involves spatial/geometric transformation of data, such as cropping and resizing (with horizontal flipping), rotation [24], and cutout [38]. The other type of augmentation involves appearance transformation, such as color distortion (including color dropping, brightness, contrast, saturation, hue) [39, 40], Gaussian blur, and Sobel filtering.

Algorithm: Our algorithm is based on [4]. We can see the general schema in Figure (3), and the final algorithm can be found in Algorithm 1

Datasets: We use two different datasets to validate the results. The CIFAR-10 dataset comprises 60,000 32x32 color images categorized into 10 classes, each containing 6,000 images. It is divided into 50,000 training images and 10,000 test images [41], and The SVHN (Street View House Numbers) dataset is 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 73,257 digits for training, 26,032 digits for testing [42].

Metrics:

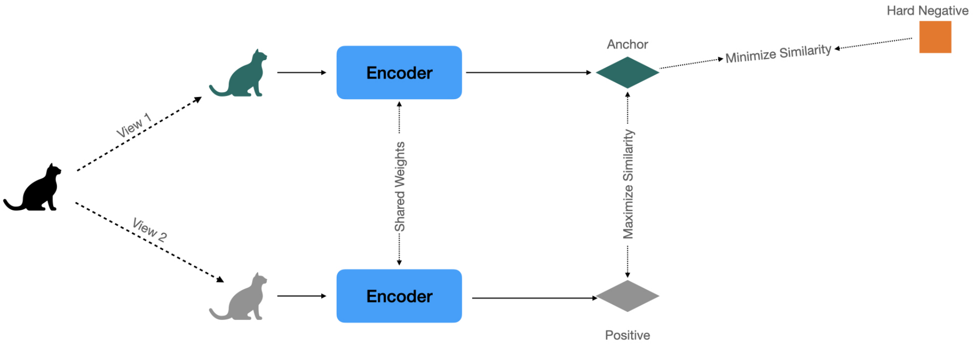


Figure 3: schema of our algorithm

- Mean Average Precision (MAP) is 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. Relevant images are defined based on query relevance, and precision is calculated by dividing the number of relevant images retrieved up to a certain position by the total number of retrieved images up to that position. To calculate MAP@K, a variant of MAP where only the top K retrieved items are considered, you can use the following formula:

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

  Where: $|Q|$ is the total number of queries, $\mathrm{Precision@k}_q$ is the precision at position k for query q, $\mathrm{Relevance}(k)$ 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.
- Accuracy, Recall, Precision, and F1-score are fundamental metrics for evaluating classification tasks. Accuracy measures the proportion of correctly classified instances among all instances, providing an overall assessment of the model's performance. Recall quantifies the proportion of true positive instances correctly identified by the model among all actual positive instances. Precision measures the proportion of true positive instances among all instances predicted as positive, offering insights into the model's precision in positive predictions. F1-score, the harmonic mean of precision and recall, balances the trade-off between precision and recall, providing a single metric that reflects both measures' performance. These metrics collectively offer a comprehensive understanding

of the classification model's effectiveness in correctly identifying instances belonging to different classes.

Evaluation: The evaluation was carried out using two different methods. Firstly, the CBIR method was employed, where the last output layer of the encoder was used to generate a feature vector for each image. Subsequently, the closest images in the training set were retrieved for each image in the test set, aiming to measure the results of the k nearest neighbors using the Mean Average Precision at K (MAP@K) metric. Secondly, a linear evaluation was conducted. In this approach, only a linear layer was added to the encoder, and then the model was retrained to perform classification using the available labels while keeping the encoder weights frozen.

Other protocols: Our encoder is based on the Very Deep Convolutional Networks for Large-Scale Image Recognition paper [43]. The batch size is 32. The maximum epoch is 200, we use stochastic gradient descent with a learning rate 0.6 and cosine learning rate decay schedule. You can observe all the details in the appendix.

4.2. Baselines

We are going to compare our approach with 4 relevant state-of-the-art works in self-supervised.

- **SimCLR [4]**: is a straightforward framework for contrastive learning of visual representations. Two distinct data augmentation operators, $\tau \sim \mathcal{T}$ and $\tau' \sim \mathcal{T}$, 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 completing the training, the projection head is discarded, and the encoder is employed to obtain a representation, denoted as h, for downstream tasks. Notably, SimCLR introduces a learnable nonlinear transformation between the representation and the contrastive loss, significantly enhancing the quality of the learned representations.
- **SimSiam [6]**: is 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 (comprising a backbone and a projection MLP). A prediction MLP is applied to one side, while a stop-gradient operation is applied to the other side. The model's objective is to maximize the similarity between both sides. Notably, SimSiam does not rely on negative pairs or a momentum encoder. The authors empirically demonstrate the existence of collapsing solutions and emphasize the critical role of the stop-gradient operation in preventing such occurrences. This suggests the presence of an underlying optimization problem different from conventional contrastive learning.
- **BYOL [7]**: is an approach to self-supervised image representation learning. It 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. Concurrently, the target network is updated with a slow-moving average of the online network. The use of a slow-moving average of the online parameters as the target

network encourages the encoding of increasing information within the online projection and mitigates the risk of collapsed solutions.

- **BarlowTwins [29]**: 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. This approach ensures that the embedding vectors of distorted versions of a sample are similar while minimizing redundancy between the components of these vectors.

4.3. Preliminary results:

The provided Tables offer a comprehensive insight into the performance metrics concerning Content-Based Image Retrieval (CBIR) and Linear Evaluation across various Self-Supervised Learning (SSL) algorithms applied to datasets CIFAR-10 [41] and The Street View House Numbers (SVHN) [42]. In the context of CBIR, the precision metric, Mean Average Precision (MAP), is computed at different values of k, indicating the number of nearest neighbors sought in the retrieval process. Each row in Table 1 corresponds to a distinct SSL algorithm, with the MAP values at different k values displayed, showcasing the algorithm's performance in retrieving relevant images. Notably, higher MAP values indicate a superior ability to retrieve relevant images in the CBIR task. In the Linear Evaluation, presented in Table 2, various performance metrics such as Accuracy, Recall, Precision, and F1-score are provided for each SSL algorithm. These Tables provide a detailed breakdown of the performance of each SSL algorithm under consideration, facilitating a nuanced understanding of their effectiveness in image retrieval and classification tasks.

Table 1
MAP Results

Model with CIFAR-10	**1000**	**100**	**10**	**1**
SimCLR [4]	0.687	0.7732	0.8221	0.881
SimSiam [6]	0.691	0.8054	0.8475	0.904
BYOL [7]	0.6917	0.7832	0.8377	0.905
BarlowTwins [29]	0.4323	0.5753	0.6689	0.791
RetSam	0.7316	0.8253	0.868	0.924
Model with SVHN	**1000**	**100**	**10**	**1**
SimCLR [4]	0.2593	0.3899	0.508	0.657
SimSiam [6]	0.5188	0.7177	0.812	0.874
BYOL [7]	0.3217	0.4636	0.584	0.715
BarlowTwins [29]	0.3758	0.5671	0.69	0.78
RetSam	0.4315	0.6004	0.6965	0.805

4.4. Analysis

Preliminary results reveal the outstanding effectiveness of our approach on two fundamental tasks: Content-Based Image Retrieval (CBIR) and Linear Evaluation.

Table 2
Linear Evaluation Results

Model with CIFAR-10	Accuracy	Recall	Precision	F1
SimCLR [4]	0.9014	0.9014	0.9014	0.9016
SimSiam [6]	0.8587	0.8587	0.8692	0.8607
BYOL [7]	0.9028	0.9028	0.9027	0.9028
BarlowTwins [29]	0.8328	0.8328	0.8331	0.8328
RetSam	0.9322	0.9322	0.9323	0.9322
Model with SVHN	**Accuracy**	**Recall**	**Precision**	**F1**
SimCLR [4]	0.8130	0.8130	0.8138	0.8127
SimSiam [6]	0.2233	0.2233	0.4027	0.1127
BYOL [7]	0.8090	0.8090	0.8102	0.8089
BarlowTwins [29]	0.8456	0.8456	0.8468	0.8457
RetSam	0.8742	0.8742	0.8752	0.8743

- **Content-Based Image Retrieval (CBIR)**: To evaluate the performance of our method on CBIR, the CIFAR-10 and SVHN datasets were used. Looking at the Table 1:
  - **CIFAR-10**: Our method significantly outperforms the baselines for different values of k. Compared to other state-of-the-art methods such as SimCLR, SimSiam, BYOL, and BarlowTwins, our approach demonstrates considerable improvement in mean average precision (MAP). We achieved a MAP of 0.7316 for k=1000, 0.8253 for k=100, 0.868 for k=10, and 0.924 for k=1, indicating a high capacity for image representation and retrieval in the latent space.
  - **SVHN**: Although our algorithm shows notable improvement compared to baselines, including SimCLR, BYOL, and BarlowTwins, in terms of MAP, it has been outperformed by the SimSiam approach. Our method achieves a MAP of 0.4315 for k=1000, 0.6004 for k=100, 0.6965 for k=10, and 0.805 for k=1. Despite not being the best in this data set, our approach is still competitive and offers promising results.
- **Linear Evaluation** To evaluate the generalization ability of the learned representations in a linear classification task, an evaluation was performed on CIFAR-10 and SVHN. Performance metrics include precision, recall, precision, and F1-score. Analyzing the Table 2
  - **CIFAR-10**: Our method excels at this task, significantly outperforming other state-of-the-art approaches such as SimCLR, SimSiam, BYOL, and BarlowTwins. We achieved a classification accuracy of 93.22%, demonstrating the effectiveness of the learned representations in linear classification tasks on this dataset.
  - **SVHN**: Our method also shows impressive performance on the linear classification task for SVHN. Although SimSiam outperforms our approach on the CBIR task, our method outperforms both SimSiam and other baselines in terms of classification accuracy, achieving an accuracy of 87.42%.

In summary, our results indicate that our approach has outstanding performance on the CBIR task in CIFAR-10, being highly competitive in SVHN. Furthermore, it demonstrates exceptional generalization ability in linear classification tasks on both data sets. These findings support the effectiveness and promise of our method in feature extraction and representation of image data.

Algorithm 1 Algorithm

```
Input: Unlabeled dataset $\mathbf{X}$
Output: Trained model
Initialize encoder network $f$
Define hyperparameters: $\alpha$, margin m
while Training not converged do
    for every batch $\mathbf{x}$ in $\mathbf{X}$ do
        Apply data transformations $\tau_1$ and $\tau_2$ to create $\tau_1(x)$ and $\tau_2(x)$
        Compute embeddings $\mathbf{z_a}$, $\mathbf{z}_p$ using $f$ in $\tau_1(x)$ and $\tau_2(x)$ respectly
        Compute distance between $\mathbf{z_a}$, $\mathbf{z}_p$ and take the k-firsts.
        Exclude the first and compute Centroid $\mathbf{z_n}$ with the rest
        Calculate $\mathscr{L}_1$ using Equation 1 with $\mathbf{z_a}$, $\mathbf{z_p}$, and $\mathbf{z_n}$
        Calculate $\mathscr{L}_2$ using Equation 2 with $\mathbf{z_a}$, $\mathbf{z_p}$, and $\mathbf{z_n}$
        Calculate total loss $\mathscr{L}_{\text{oss}}$ using Equation 3
        Update model parameters using backpropagation
    end for
end while
Return: Trained model
```

5. Conclusion

The landscape of Self-Supervised Representation Learning (SRL) has witnessed significant advancements, and this paper contributes to the field by addressing a crucial limitation in existing methods. Traditional approaches often focus on learning similarity without adequately capturing dissimilarity nuances, leading to suboptimal representations. Our proposed method, termed "Refining Triplet Sampling", introduces a novel strategy for generating negative vectors in a batch, enhancing the triplet loss methodology for representation learning. The motivation behind our approach stems from the challenge of creating dissimilar data, a critical aspect of effective SRL. Existing methods, including those relying on the average as a measure of dissimilarity, fall short of providing robust negative representations. Our method tackles this limitation by constructing negative samples based on the k-nearest neighbors, significantly improving the model's ability to differentiate dissimilar instances.

Experimental results, particularly in Content-Based Image Retrieval (CBIR) and Linear Evaluation, consistently demonstrate the superiority of our approach over other Self-Supervised Learning (SSL) methods (baselines). The refined representations showcase higher Mean Average Precision (MAP) values in CBIR, emphasizing the effectiveness of our method in retrieving relevant images. Linear Evaluation further underscores the versatility of our learned representations, outperforming other algorithms in terms of Accuracy, Recall, Precision, and F1.

Acknowledgments

This research received partial support through an agreement with Scotiabank and Federico Santa María Technical University, as well as via a scholarship for international visits provided by Federico Santa María Technical University and the National Agency for Research and Development (doctoral scholarship 2022/21221059).

References

[1] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9729–9738.

[2] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, G. E. Hinton, Big self-supervised models are strong semi-supervised learners, Advances in neural information processing systems 33 (2020) 22243–22255.

[3] M. Gutmann, A. Hyvärinen, Noise-contrastive estimation: A new estimation principle for unnormalized statistical models, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[4] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR, 2020, pp. 1597–1607.

[5] G. Wang, K. Wang, G. Wang, P. H. Torr, L. Lin, Solving inefficiency of self-supervised representation learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9505–9515.

[6] X. Chen, K. He, Exploring simple siamese representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 15750–15758.

[7] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al., Bootstrap your own latent-a new approach to self-supervised learning, Advances in neural information processing systems 33 (2020) 21271–21284.

[8] T. Huynh, S. Kornblith, M. R. Walter, M. Maire, M. Khademi, Boosting contrastive self-supervised learning with false negative cancellation, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022, pp. 2785–2795.

[9] J. Robinson, C.-Y. Chuang, S. Sra, S. Jegelka, Contrastive learning with hard negative samples, arXiv preprint arXiv:2010.04592 (2020).

[10] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

[11] H. Oh Song, Y. Xiang, S. Jegelka, S. Savarese, Deep metric learning via lifted structured

feature embedding, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4004–4012.

[12] T. T. Cai, J. Frankle, D. J. Schwab, A. S. Morcos, Are all negatives created equal in contrastive instance discrimination?, arXiv preprint arXiv:2010.06682 (2020).

[13] C. Doersch, A. Gupta, A. A. Efros, Unsupervised visual representation learning by context prediction, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1422–1430.

[14] P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, in: Proceedings of the 25th international conference on Machine learning, 2008, pp. 1096–1103.

[15] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27 (2014).

[17] C. Doersch, A. Zisserman, Multi-task self-supervised visual learning, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2051–2060.

[18] R. Zhang, P. Isola, A. A. Efros, Colorful image colorization, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14, Springer, 2016, pp. 649–666.

[19] G. Larsson, M. Maire, G. Shakhnarovich, Learning representations for automatic colorization, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, Springer, 2016, pp. 577–593.

[20] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A. A. Efros, Context encoders: Feature learning by inpainting, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2536–2544.

[21] M. Noroozi, P. Favaro, Unsupervised learning of visual representations by solving jigsaw puzzles, in: European conference on computer vision, Springer, 2016, pp. 69–84.

[22] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, 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, 2017, pp. 4681–4690.

[23] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, T. Brox, Discriminative unsupervised feature learning with convolutional neural networks, Advances in neural information processing systems 27 (2014).

[24] S. Gidaris, P. Singh, N. Komodakis, Unsupervised representation learning by predicting image rotations, arXiv preprint arXiv:1803.07728 (2018).

[25] A. Kolesnikov, X. Zhai, L. Beyer, Revisiting self-supervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 1920–1929.

[26] X. Chen, H. Fan, R. Girshick, K. He, Improved baselines with momentum contrastive learning, arXiv preprint arXiv:2003.04297 (2020).

[27] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, P. Isola, What makes for good views for contrastive learning?, Advances in neural information processing systems 33 (2020)

6827–6839.

[28] Z. Wu, Y. Xiong, S. X. Yu, D. Lin, Unsupervised feature learning via non-parametric instance discrimination, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3733–3742.

[29] J. Zbontar, L. Jing, I. Misra, Y. LeCun, S. Deny, Barlow twins: Self-supervised learning via redundancy reduction, in: International Conference on Machine Learning, PMLR, 2021, pp. 12310–12320.

[30] S. Ding, L. Lin, G. Wang, H. Chao, Deep feature learning with relative distance comparison for person re-identification, Pattern Recognition 48 (2015) 2993–3003.

[31] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, arXiv preprint arXiv:1703.07737 (2017).

[32] X. Wang, H. Zhang, W. Huang, M. R. Scott, Cross-batch memory for embedding learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6388–6397.

[33] G. Wang, G. Wang, X. Zhang, J. Lai, Z. Yu, L. Lin, Weakly supervised person re-id: Differentiable graphical learning and a new benchmark, IEEE Transactions on Neural Networks and Learning Systems 32 (2020) 2142–2156.

[34] N. Turpault, R. Serizel, E. Vincent, Semi-supervised triplet loss based learning of ambient audio embeddings, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 760–764.

[35] W. Li, X. Yang, M. Kong, L. Wang, J. Huo, Y. Gao, J. Luo, Trip-roma: Self-supervised learning with triplets and random mappings, Transactions on Machine Learning Research (2022).

[36] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), volume 2, IEEE, 2006, pp. 1735–1742.

[37] C.-Y. Wu, R. Manmatha, A. J. Smola, P. Krahenbuhl, Sampling matters in deep embedding learning, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2840–2848.

[38] T. DeVries, G. W. Taylor, Improved regularization of convolutional neural networks with cutout, arXiv preprint arXiv:1708.04552 (2017).

[39] A. G. Howard, Some improvements on deep convolutional neural network based image classification, arXiv preprint arXiv:1312.5402 (2013).

[40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[41] A. Krizhevsky, Learning multiple layers of features from tiny images, Technical Report, 2009.

[42] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng, Reading digits in natural images with unsupervised feature learning, in: NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011, 2011. URL: http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

[43] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

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 3, 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.

Table 3
Selected Hyperparameters

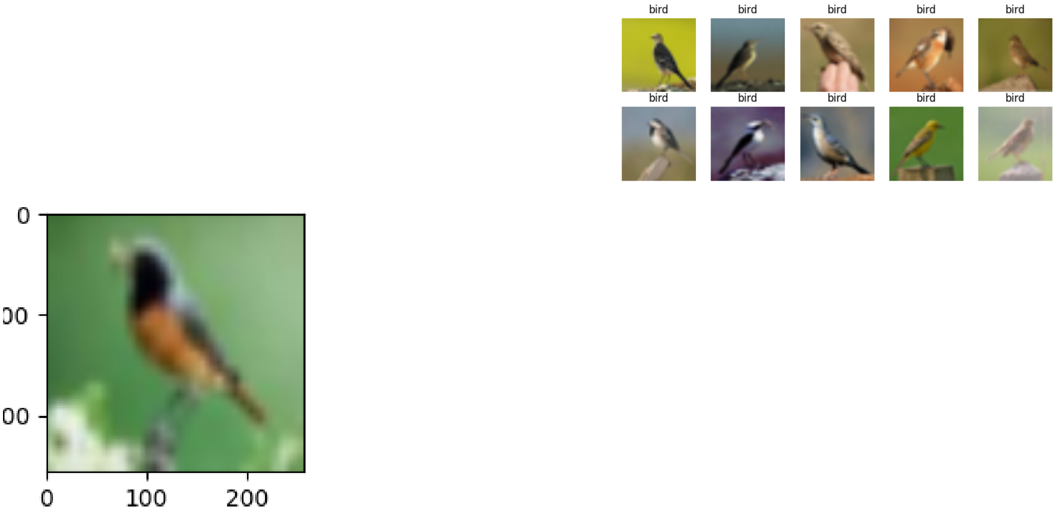
Hyperparameter	Selected Value
Learning Rate	0.06
Epochs	200
Batch Size	32
Decay Schedule	cosine learning rate
Optimizer	SGD
Encoder	VGG50(weights=”imagenet”)
k- neighbors	15
m (margin)	0.6
α	0.5

A.3. Dataset details

Additional information about the datasets is presented in the Table 4. 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.

Table 4
Dataset Information

Type	Name	Train	Test	N° Classes
Natural Image	Cifar-10	50000	10000	10
Numbers	Street View House Number	73257	26032	10



(a) Query Image

(b) Recovery Images

Figure 4: Class: Bird

A.4. Recovery Visualization

In this subsection, we present visual examples showcasing the recovery achieved by our method. These illustrations are depicted in Figures ??, ??, ??, ??, and ??. Through these images, we aim to demonstrate the effectiveness of our approach in accurately reconstructing the original content. Notably, our method excels in preserving the semantic integrity of the images during the recovery process, thereby emphasizing its robust performance in retaining crucial visual details and structures

A.5. Online Resources

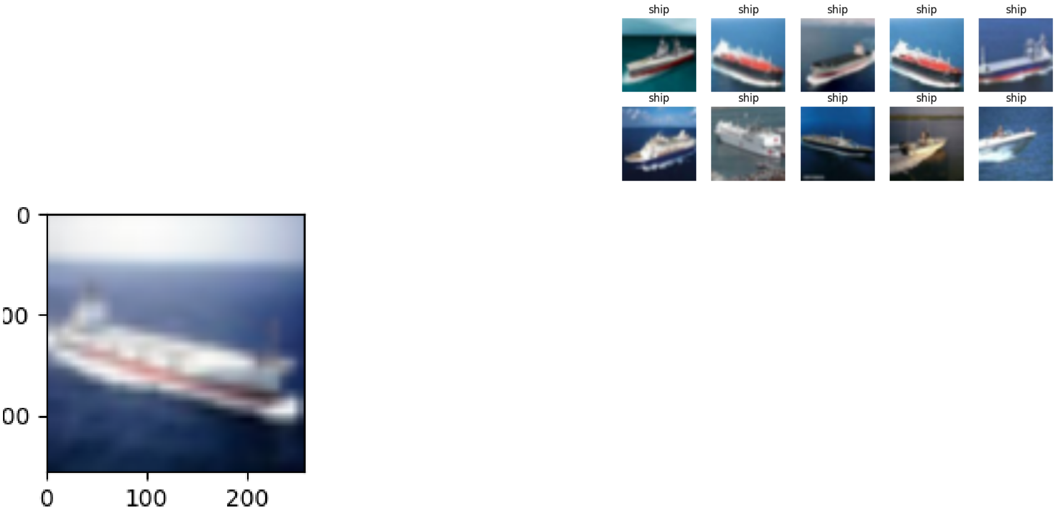
For those interested in replicating our results, the code is available on GitHub at the following link:

GitHub Repository

This repository contains the necessary resources and instructions to facilitate the replication of our findings. Feel free to explore and utilize the code to delve deeper into our methodology and validate the outcomes.

A.6. Future Work

Despite the advancements presented in this work in the domain of image retrieval and classification, there are several lines of research that can further enrich our approach and explore its applicability in different visual contexts. Below are highlighted some areas of interest for future investigations:



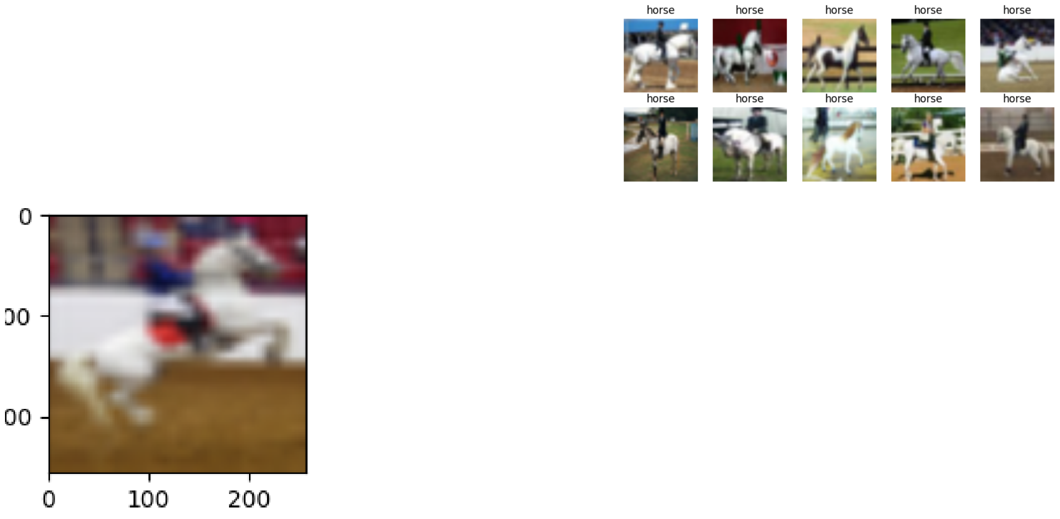
(a) Query Image

(b) Recovery Images

Figure 5: Class: Ship

- **Exploration of Diversity in Image Datasets**: To assess the robustness and generalization of our algorithm across different visual domains, we propose the inclusion of additional datasets representing diverse nature of images. This could involve datasets containing medical images, satellite data, texture images, among others. Expanding the domains of images will allow for a more comprehensive evaluation of the algorithm's ability to adapt to a variety of visual contexts.
- **Transfer Learning in Cross-Domain Scenarios**: To extend our research on transfer learning, we suggest exploring cross-domain scenarios where the model is trained on one dataset and evaluated on another with different visual characteristics. This line of investigation will help assess the algorithm's adaptation capability to different visual styles and evaluate the transferability of learned representations across different image domains.
- **Exploration of Semi-Supervised Learning Techniques**: To further improve the performance of the algorithm in image retrieval and classification tasks, we propose investigating semi-supervised learning techniques. This approach leverages both labeled and unlabeled data to train the model, which can be particularly useful in scenarios where labeled datasets are scarce or expensive to obtain. Exploring semi-supervised strategies could open up new opportunities to enhance the efficiency and accuracy of the algorithm in computer vision tasks.

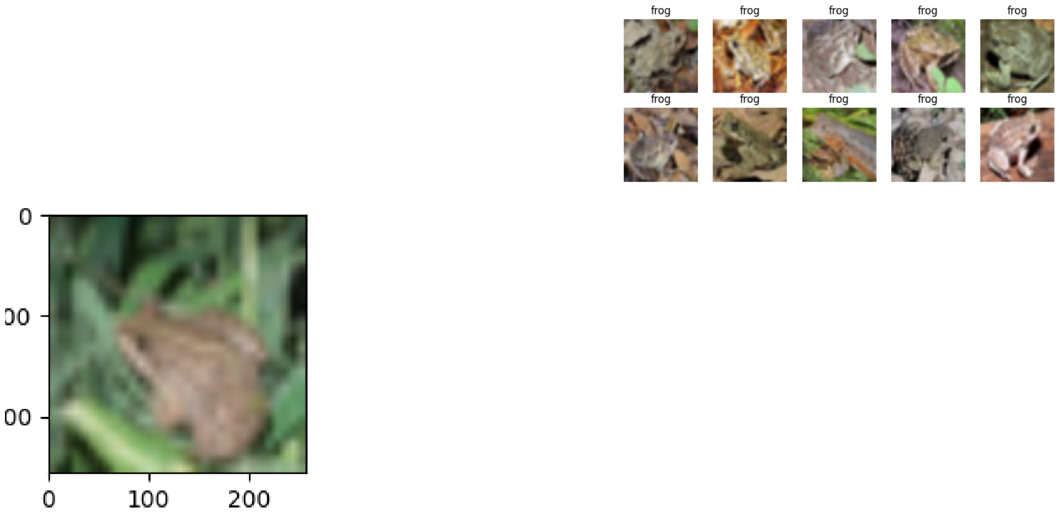
These research directions represent significant steps towards advancing our understanding of self-supervised algorithms in the field of computer vision and their application in a variety of visual domains and real-world scenarios.



(a) Query Image

(b) Recovery Images

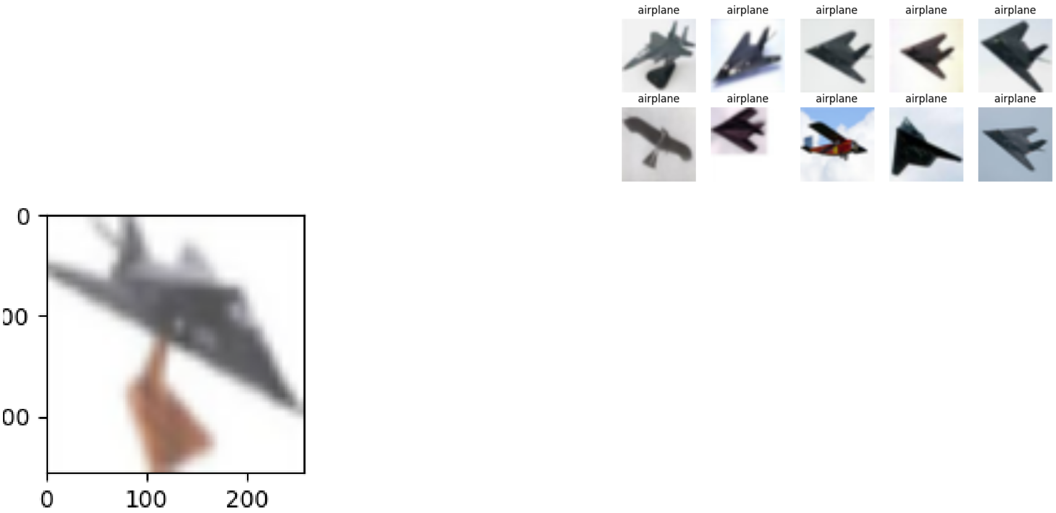
Figure 6: Class: Horse



(a) Query Image

(b) Recovery Images

Figure 7: Class: Frog



(a) Query Image

(b) Recovery Images

Figure 8: Class: Airplane