MULTI-SOURCE MULTI-VIEW (MSMV): A ROBUST SSL FRAMEWORK FOR IMBALANCED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

One of the most challenging problems in self-supervised learning (SSL) is to achieve robustness of the learned representations irrespective of whether the data is balanced or imbalanced. SSL methods address this problem through techniques like feature reweighting, regularization or by making use of out of distribution(OOD) data. However, contrastive self supervised learning (CSSL) is more sensitive to imbalance, as frequent samples produce disproportionately more contrastive pairs leading to biased representations. Our analysis on the long-tail behavior in SSL clearly shows that most existing methods are biased in terms of *gradient ratio* which could lead to degradation of representation quality as well as poor decision boundary separation. We believe that the core reason for such quality degradation comes from the traditional *multi-view* design on which most existing SSL frameworks are based. Building on these findings, we design a novel contrastive learning framework (MSMV) in which multiple sources are considered to promote balanced information flow across views rather than a single source as in multi-view learning. The proposed model is enhanced by introducing a new contrastive loss function that helps in better representation learning for tail classes. Additionally, we adopted a unique augmentation strategy to improve the overall performance. Extensive experiments achieve consistent improvements across multiple benchmarks, including a 5% increase in linear evaluation and an 8% increase in few-shot accuracy on Cifar100-LT with ResNet-18, highlighting its effectiveness in imbalanced scenarios. We also find consistent gains on other datasets.

1 INTRODUCTION

For image datasets, learning visual representations from unlabeled data is the primary objective of self-supervised learning(SSL) Chen et al. (2020a;c); Khosla et al. (2020). A variety of frameworks support self supervised learning, including *contrastive* Chen et al. (2020a;c); Jiang et al. (2021), *non-contrastive* Grill et al. (2020); Chen & He (2021); Bardes et al. (2022a) and *hybrid* Kukleva et al. (2023); Zhou et al. (2022). Despite their success in balanced datasets, these methods often struggle under challenging conditions such as long-tailed data distributions. In such scenarios, learning dynamics become skewed, leading to biased representations and degraded performance on underrepresented components.

Previous works Lin et al. (2023); Zhou et al. (2022); Kukleva et al. (2023); Bai et al. (2023) have addressed the dataset imbalance problem in SSL using feature weighting, regularization, and out-of-distribution data. However, each of these methods has its own limitations. In feature weighting, the balancing act is achieved through reweighting of features, and this often leads to overfitting due to their sensitivity to noise. Regularization tricks do not directly address the skewness in the representation and, therefore, can only do partial correction of the imbalance. This could result in optimization instability/limited performance gains. Out-of-distribution data require careful selection of auxiliary data, as random data can mislead the representation and lead to the learning of irrelevant patterns.

We argue that the degradation in the representation quality is not solely due to data imbalance but also arises from an inherent optimization imbalance. Specifically, gradient contributions across classes become skewed during training, leading to a mismatch between optimization dynamics and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the resulting representation quality. Through empirical analysis, we observe that such an imbalanced optimization does not translate into improved representations for tail classes, resulting in degraded feature separability and decision boundaries. Our contributions are summarized as follows:

- We identify that as the environment shifts from balanced to imbalanced, the gradient ratio (head gradient / tail gradient) also shifts towards either the head or the tail. The primary reason for degradation of representation quality is this shift in nature of gradient ratio, which to the best of our knowledge has not been addressed within the SSL community.
- We design an experimental framework based on gradient ratios and representation metrics (KNN, Margin) to learn the characteristics of environment shifts in SSL.
- We propose a novel framework(MSMV) that balances the feature space by its design, resulting in a balanced gradient ratio and better representation learning of the tail classes.
- We introduce a new contrastive loss that filters the feature space, improves robustness, and enhances feature quality. Additionally, we adopt an augmentation strategy to improve our model's overall performance.
- Extensive experiments were carried out to validate our claim in long-tail benchmark datasets, particularly for tail classes, including a 5% increase in linear evaluation and an 8% increase in few-shot accuracy on Cifar100-LT with ResNet-18. We gain a 2% increase in linear evaluation on Cifar-10 LT and Imagenet-LT. In a more imbalanced environment prepared using Imagnet-LT, we achieve an almost 10% gain.

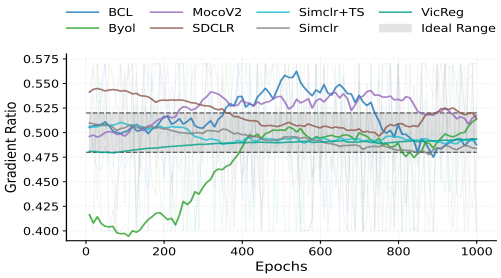
2 Problem 1: Analysing Long-Tail Behavior in SSL

We first examine the gradient allocation patterns across head and tail classes in SSL. To compute the gradient ratio from the gradient norm of the head and tail groups, as shown in Figure 1, we need to fully train a model on a long-tail dataset (we used CIFAR-10-LT as the long-tail dataset to calculate gradient ratio). In order to remove frequency bias, we artificially created a balanced dataset using the same Cifar-10LT dataset. We selected 40 images per category based on the fact that the lowest number of images carried by a particular tail class is 50 in Cifar-10LT. Further details related to the experimental setup are given in the appendix. The artificially generated balanced dataset is then passed to the trained model. The loss is computed for the head and tail classes separately and their gradient norms are obtained using backpropagation. The gradient ratio is determined by the following equation:

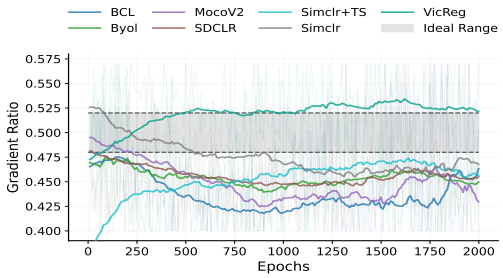
$$G_r = \frac{\parallel G_{head} \parallel_2}{\parallel G_{head} \parallel_2 + \parallel G_{tail} \parallel_2} \tag{1}$$

Here, G_{head} and G_{tail} are the gradient norms of the head and tail groups, respectively. A ratio close to 0.5 represents balanced optimization (Figure 1a) while a shift from this balanced optimization represents biased behavior in the long-tail environment (Figure 1b). These observations reveal a critical mismatch between gradient allocation patterns and representation quality: due to unstable gradient allocation, tail classes do not achieve improved feature separability or margins (Figure 1c, 1d). The above analysis of long-tail behavior in SSL prompted us to also look into the multi-view setting on which most SSL frameworks are based. The reason for this intuition is that we could observe a similar gradient behavior in two highly relevant works Zhou et al. (2022); Kukleva et al. (2023) that have contributed to solving the long-tail problem in SSL. In Figure 1b these two works are denoted as BCL and SimCLR+TS. In a multi-view setting it is common to generate N views from a single source image. In a balanced environment, the number of views will be uniformly distributed even after generating N views. On the other hand, in a long-tail environment, generating N views will result in more imbalance. For example, if $N = 2$, then a head category that has 5000 images will generate 10000 views, while the tail category that has 50 images will only have 100 views. At first, this change looks even, but if we observe the difference in the number of generated views, then we find that the head is getting 5000 more views while the tail is getting only 50 extra views. As the number of views N increases, the under representation of tail classes grows even more comparatively. The analysis of the gradient ratio as well as K-nn/margin comparison between balanced/imbalanced data as shown in Figure 1 clearly demonstrates that most existing SSL methods suffer from the long-tail problem as minority classes are underrepresented.

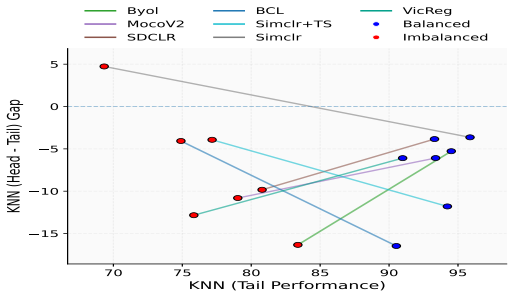
In Section 4, we propose a contrastive self supervised learning framework (MSMV) to address these concerns.



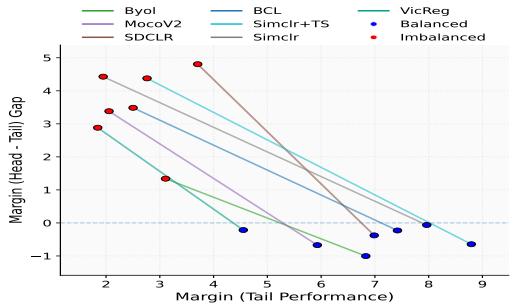
(a) Analysis of gradient Ratio on balanced data



(b) Analysis of gradient Ratio on imbalanced data



(c) KNN ((head minus tail) versus tail) comparison in balanced and imbalanced data.



(d) Margin ((head minus tail) versus tail) comparison in balanced and imbalanced data.

Figure 1: Gradient allocation pattern(gradient ratio) is shown in Figure 1(a) (with respect to balanced data) and Figure 1(b) (with respect to imbalanced data). Balanced gradient allocation patterns are reported near 0.5 (highlighted in grey). From Figure 1(b) it can be clearly seen that none of the existing SSL methods falls in the grey area. Representation quality (Knn and Margin) with respect to balanced and imbalanced data is shown in Figure 1(c) and 1(d). The interpretation can be done as follows: going left on the X-axis shows the degradation in tail performance . Length of the line represents robustness of the model. Ideal scenario is to keep minimum length and maximum tail. For instance, from Figures 1(c) and 1(d) it can be inferred that Byol has better representation quality when compared to others.

3 Problem 2: Contrastive learning is biased towards class frequency

In contrastive learning-based SSL frameworks, each image is optimized independently until the loss computation is carried out. A closer look at the contrastive loss of SSL reveals that it can be seen as a *one-versus-rest* identification task. This is because while calculating the loss, the augmented views of an *anchor* image needs to be distinguished from all other images. For contrastive learning, the NT-Xent (InfoNCE) loss for an anchor sample i is

$$\mathcal{L}_i = -\log \frac{\exp\big(\text{sim}(z_i, z_i^+)/\tau\big)}{\sum_{k \neq i} \exp(\text{sim}(z_i, z_k)/\tau)}, \tag{2}$$

where z_i denotes the anchor representation, z_i^+ denotes the positive representation corresponding to the same image, τ is the temperature parameter, and $\text{sim}(\cdot, \cdot)$ is a similarity function (typically cosine similarity). From the perspective of the loss function: There is one anchor representation, z_i. There is one positive representation, z_i^+. All remaining representations $\{z_k\}_{k \neq i}$ in the denominator

are treated collectively as "the rest" (negative samples). Thus, the objective can be interpreted as a one-vs-rest discrimination task, where the model learns to identify the positive pair among all other samples in the batch. The SGD update rule is given by

$$\theta_{t+1} = \theta_t - \eta \frac{1}{B} \sum_{i=1}^{B} \nabla_\theta \mathcal{L}_i, \tag{3}$$

where η is the learning rate, B is the batch size, and $\mathcal{L}_i$ denotes the loss associated with sample i. Over the entire training process, the cumulative parameter update can be expressed as

$$\Delta\theta \propto \sum_{i \in \text{all appearances}} \nabla_\theta \mathcal{L}_i. \tag{4}$$

From Equation 4 it follows that the parameter/s update is heavily dependent on the frequency of samples and this behavior is reflected in Figure 1b. We provide a toy example to make this notion clear. The following example illustrates how unequal observation frequencies naturally induce a frequency-dependent learning bias. Consider three classes n_A, n_B and n_C with different sample frequencies 1000, 100 and 10. Let each class has a true mean feature value μ_A, μ_B and μ_C. By separately learning each class, one can have a class mean estimation as $\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_j$ and the variance of the estimated mean as $\mathrm{Var}(\hat{\mu}_i) = \frac{\sigma^2}{n_i}$. where σ^2 denotes the data variance. Thus,

$$\mathrm{Var}(\hat{\mu}_A) = \frac{\sigma^2}{1000}, \qquad \mathrm{Var}(\hat{\mu}_B) = \frac{\sigma^2}{100}, \qquad \mathrm{Var}(\hat{\mu}_C) = \frac{\sigma^2}{10}.$$

Since $\frac{\sigma^2}{10} > \frac{\sigma^2}{100} > \frac{\sigma^2}{1000}$, the rare class C suffers from substantially higher estimation uncertainty. Consequently, learning quality is directly governed by the sample frequency of each class.

Real-world observations typically exhibit a combination of shared structure and entity-specific variations, rather than being entirely independent of one another. Now assume that mean all classes share a common latent component:

$$\mu_i = \mu_0 + \delta_i,$$

where μ_0 is a shared mean estimated from all classes, δ_i is a class-specific deviation. The total number of samples becomes $N = n_A + n_B + n_C = 1000 + 100 + 10 = 1110$. The shared mean and variance are estimated as

$$\hat{\mu}_0 = \frac{1}{N} \sum_{i=1}^{N} x_i, \qquad \mathrm{Var}(\hat{\mu}_0) = \frac{\sigma^2}{1110}.$$

Each class estimate is then obtained as $\hat{\mu}_i = \hat{\mu}_0 + \hat{\delta}_i$. Unlike independent estimation, the dominant component $\hat{\mu}_0$ is learned using all 1110 samples. Therefore, even the rare class C benefits from information contributed by the frequent classes. When learning is decomposed into a shared component and a class-specific component, the shared component can be estimated using all available data. Consequently, rare classes inherit information from the entire dataset, making learning substantially less sensitive to individual class frequencies than fully independent estimation.

The above observation serves as the foundation for developing a SSL paradigm that mitigates frequency-dependent biases through shared representation learning. Let X_1, X_2 be two images taken from the dataset $\mathbb{D}$. Let us also consider that $x_1^{a_1}$ *and* $x_1^{a_2}$ and $x_1^{a_2}$ *and* $x_2^{a_2}$ be the two augmented views of X_1 *and* X_2 while $z_1^{a_1}$ *and* $z_1^{a_2}$ and $z_1^{a_2}$ *and* $z_2^{a_2}$ are their latent representations. The key observation of using shared component as well as specific component can be interpreted in terms of inter relation(shared component) and intra relation(specific component) as given below

$$\underbrace{(z_1^{a_1}.z_1^{a_2})(z_2^{a_1}.z_2^{a_2})}_{\textbf{intra relation}} \underbrace{(z_1^{a_1}.z_2^{a_1})(z_1^{a_2}.z_2^{a_2})}_{\textbf{inter relation}}$$

where dot(.) is the similarity between the two representations. We need to come up with a representation that can accommodate both intra and inter relation in a compact manner. For this we make use of the fusion operation Li et al. (2018) as given below where $\circledast$ is a fusion operator.

$$\begin{aligned} &(z_1^{a_1}.z_1^{a_2}) \circledast (z_1^{a_1}.z_2^{a_1}) \circledast (z_1^{a_2}.z_2^{a_2}) \circledast (z_2^{a_1}, z_2^{a_2}) \\ &(z_1^{a_1} \circledast z_2^{a_2}).(z_1^{a_2} \circledast z_2^{a_1}) \end{aligned} \tag{5}$$

In order to accommodate inter and intra relationship in SSL framework, we need to move away from the traditional multi-view pradigm as it is limited to capture only multiple views of a single source. While our idea of using both inter and intra relation from Equation 5 advocates using a multi-source paradigm, which we explain in detail in Section 5

4 Related Work

SSL falls into the Multi-view Self supervised learning family (MVSSL) which in turn is classified into three families, viz. contrastive Chen et al. (2020a;c); He et al. (2020); Caron et al. (2021); Bardes et al. (2022b), clustering Caron et al. (2018; 2020) and distillation-based Grill et al. (2020). The proposed work belongs to the contrastive family van den Oord et al. (2018); Chen et al. (2020a;b); He et al. (2020); Chen et al. (2020c); Tian et al. (2020a); Chuang et al. (2020); Khosla et al. (2020); Chen et al. (2021); Jiang et al. (2021). The basic assumption in SSL is to generate two or more views(multi-view) for each data sample by using augmentations van den Oord et al. (2018); Tian et al. (2020b) so that the semantic information shared between the views remains as intact as that of the original sample.

All major SSL frameworks are supported by the Multi-view Information bottleneck principle(MVIB) Tishby et al. (2000); Sridharan & Kakade (2008); Tishby & Zaslavsky (2015); Federici et al. (2020); Tsai et al. (2021); Gálvez et al. (2023); Wang et al. (2023); Louizos et al. (2024), which aims to learn representations that maximize *shared/relevant* information across views of the same sample while minimizing *unnecessary/redundant* information. The traditional MVIB principle is designed for a single image multi-view perspective which may not work in scenarios wherein multiple images and multiple views are involved as in our case. To this end, we introduce the MSMV bottleneck principle.

Recent work on SSL has demonstrated that compared to fully supervised models, architectures that leverage self-supervised pretraining are more resistant to class imbalance Yang & Xu (2020); Jiang et al. (2021); Liu et al. (2022); Lin et al. (2023); Bai et al. (2023); Kukleva et al. (2023). These methods advocate using out-of-distribution (OOD) data or in-domain (ID) data samples to balance the minority class to boost the long-tailed learning performance of SSL. Most of the methods mentioned above Yang & Xu (2020); Liu et al. (2022); Lin et al. (2023); Bai et al. (2023); Kukleva et al. (2023) make use of the additional data (ID or OOD) to re-balance the features. Other methods address the long-tailed problem by strengthening minority features through sampling or reweighting techniques Liu et al. (2022). There is also a work that addresses the issue of data set imbalance in SSL using a prototypical re-balancing strategy Lin et al. (2023). Of these, except for Bai et al. (2023), all other works address the problem of data imbalance by sampling with extra domain data that can re-balance the minority class. Contrastive Learning by nature is biased towards the head classes due to the large number of negatives.

5 The Framework

As can be seen from Fig. 2, we start with an anchor image (X_A) and it's counterpart X_C. Further, x_A^n, x_C^n and x_A^a, x_C^a are the normalized and augmented versions of anchor X_A and it's counterpart X_C respectively. We make use of both invertible (normalized) and non-invertible (augmented) functions for transforming views whereas any traditional self-supervised framework is designed to make use of only non-invertible functions in the form of augmentations. We initially carried out experiments to find the sanctity of this intuition, the results of which is given in Section 6. We adopted the momentum learning He et al. (2020) paradigm for pretraining as it has been the preferred choice of various self-supervised frameworks Chen et al. (2020c); Grill et al. (2020); Caron et al. (2020); Chen & He (2021). Momentum learning uses two parallel networks simultaneously for pretraining, known as the online encoder and the target encoder. In the Fig. 2, we denote the online encoder as $encoder_q$ and the target encoder as $encoder_k$. The online encoder representations are denoted as $z_A^c = encoder_q(x_A^n)$ and $z_C^a = encoder_q(x_C^a)$. On the other hand, target encoder representations are denoted as $z_A^a = encoder_k(x_A^a)$ and $z_C^n = encoder_k(x_C^n)$. $encoder_k$ is exponential moving average(EMA) of $encoder_q$'s parameters, while $encoder_q$ is used to update the gradients

Algorithm 1 MSMV:

Input: Batch size N, Normalized Images x_A^n, x_C^n, Augmented Image x_A^a, x_C^a, $encoder_q$, $encoder_k$,
for batch in train˙loader **do**
  $x_A^n, x_C^n, x_A^a, x_C^a$ = batch
  $z_A^n, z_C^a = encoder_q(x_A^n), encoder_q(x_C^a)$
  with no˙grad():
    $momentum_update_encoder_k()$
    $z_A^a, z_C^n =$
      $encoder_k(x_A^a)$,$encoder_k(x_C^n)$
  $S = sim\left((z_A^n \circledast z_C^a), (z_A^a \circledast z_C^n)\right)$
  $S(i,j) = \{S_{i,j}.1_{[\lambda_l,\lambda_h]}S_{i,j}\}$
  $\mathcal{L} = -log\left[\frac{e^{S/\tau}}{\sum_{k=1}^{2N} 1_{\{i\neq k\}} e^{S/\tau}}\right]$
end for
return $\mathcal{L}$

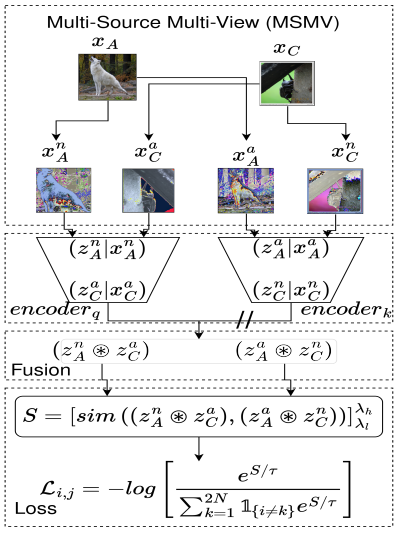


Figure 2: MSMV illustration

by backpropagation. The overall loss over these representation pairs can be computed as

$$\mathcal{L}_{i,j} = -log\frac{e^{(sim((z_A^n \circledast z_C^a)_i\ (z_A^a \circledast z_C^n)_j)/\tau)}}{\sum_{k=1}^{2N} 1_{\{i\neq k\}} e^{\left(sim\left((z_A^n \circledast z_C^a)_i\ (z_A^a \circledast z_C^n)_k\right)/\tau\right)}} \tag{6}$$

5.1 Filtered Contrastive Loss(FCL):

This loss function given in Equation 15 is a variant of the NT-Xent Chen et al. (2020a) loss function, which adopts a compact representation to form the positives and negatives. The loss function given above is modified to eliminate extreme features that are within a certain threshold (λ_l (lower limit), λ_h (higher limit)). Two extreme cases that could result are when we have extremely similar/dissimilar images.

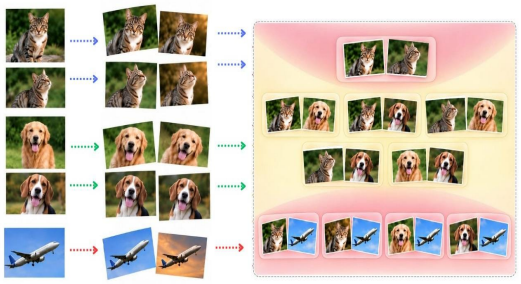


Figure 3: FCL illustration: (Case 1: upper pink area) and Case 2: lower pink area) represents the least informative relation while mid area (Case 3: yellow area) represents the most informative relations according to FCL.

Case 1: Extremely similar Let z_A and z_C be almost similar image representations such that the similarity between both representations is close to one ($sim(z_A, z_C) \backsimeq 1$). In this case of extreme similarity (Figure 3,upper pink area), intra similarity ($z_A^n.z_A^a\ and\ z_C^n.z_C^a$) becomes irrelevant. It can be high or low, but it does not make any difference because the inter similarity $z_A^n.z_C^n\ and\ z_C^a.z_A^a$ already dominates if $z_A^n.z_A^a \circledast z_A^n.z_C^n \circledast z_C^a.z_A^a \circledast z_C^a.z_C^n \backsimeq 1$. These types of cases can be ignored because good examples are those that exist closer in latent space but are symmetrically different. This is not possible in extreme cases.

Case 2: Extremely dissimilar Similarly, if we have extremely dissimilar views (Figure 3, lower pink area) i.e. z_A and z_C are completely different from each other, then $sim(z_A, z_C) \backsimeq 0$. Here $(z_A^n.z_C^a)\ and\ (z_A^a.z_C^n)$ are inter similarity pairs. Since z_A and z_C are completely distinct from each other, we end up with

$$(z_A^n.z_C^a) \backsimeq 0\ and\ (z_A^a.z_C^n) \backsimeq\ 0$$

and therefore in these kinds of scenario, decision making is highly dependent on intra-similarity pairs $x_A.x_A'\ and\ z_C.z_C'$. Even though these pairs are instances from the same image, their similarity

may decrease because of more challenging augmentation of the images and may result in extreme cases as follows:

$$z_A^n.z_A^a \circledast z_A^n.z_C^n \circledast z_C^a.z_A^a \circledast z_C^a.z_C^n \backsimeq 0$$

This is a case of extreme dissimilarity and these types of case are also not important and should be ignored.

Case 3: Semantically related (most useful) Given a certain threshold, the loss function proposed in Eq. (7) takes care of eliminating the extreme cases given in Case 1 and Case 2 so that only the most useful semantically related pairs of images are considered as shown in the Figure 3 (yellow area). The loss function in Eq. (7) can be explained as follows:

$$\begin{aligned} \text{Let } S &= Sim\left((z_A^n \circledast z_C^a).(z_A^a \circledast z_C^n)\right) \\ S(i,j) &= \{S_{i,j}.1_{[\lambda_l,\lambda_h]}S_{i,j}\} \\ \text{where } 1_{[\lambda_l,\lambda_h]}S_{i,j} &= \begin{cases} \mathcal{S}_{i,j}, & \text{if } \lambda_l \leq S_{i,j} \leq \lambda_h \\ 0, & \text{otherwise} \end{cases} \\ \text{Final Loss: } \mathcal{L}_{i,j} &= -log\left[\frac{e^{S/\tau}}{\sum_{k=1}^{2N} 1_{\{i \neq k\}} e^{S/\tau}}\right] \end{aligned} \tag{7}$$

First, we compute the pairwise similarity between our aggregated representations. In addition, using a certain threshold, extreme similarities are eliminated. Finally, these refined similarities are fed into our final loss function. This new loss function becomes more robust as the training progresses, as it is able to eliminate more extreme features and this characteristic of the loss function is shown in the Appendix. The detailed algorithm related to our Multi-Source Multi-View (MSMV) approach is given in Algorithm. 1. The significance of MSMV over multi-view paradigm is given in Section A.10.2, A.11, A.11.1 with respect to the imbalance scenario.

6 Results and Discussion

We use exponential distribution to create Cifar10-LT Krizhevsky et al. (2009) and Cifar100-LT Krizhevsky et al. (2009) with an imbalance factor of $r = 0.01$, taken from Liu et al. (2022). For ImageNet, we used a Pareto distribution to create ImageNet-LT Russakovsky et al. (2015). The Pareto distribution is more likely to generate real-world data. We followed ImageNet-LT and their subset construction as suggested by Lie et al. Liu et al. (2022). An imbalance factor of $\alpha = 0.004$ produces a challenging imbalance dataset. Furthermore, we used subsampling to make some more challenging subsets of the dataset, while its structure remains intact. We use SubsetRandomSampler with the Stratified Sampling technique to create samples.

Augmenetation Strategy The experimental results in Figure 5 shows that the addition of the normalized view in conjunction with augmentation alone improves SimCLR Chen et al. (2020a) performance by 2.8% over the test data. We will adopt this augmentation strategy throughout our proposed framework. MSMV alone with augmented views outperforms SimCLR by a margin of 3.5% over test data. Furthermore, by combining MSMV with normalized and augmented views, an improvement of more than 8% is achieved over the test data as shown in Fig. 5. All experiments were carried out on Cifar100-LT with 2000 epochs.

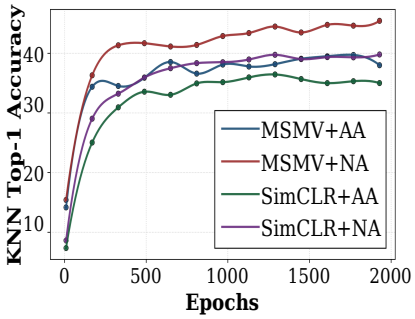
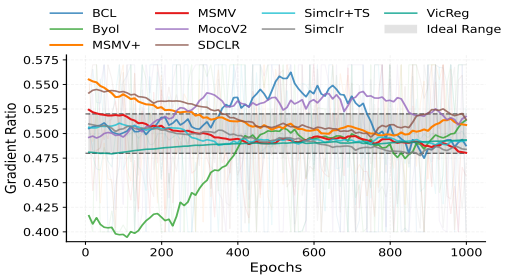
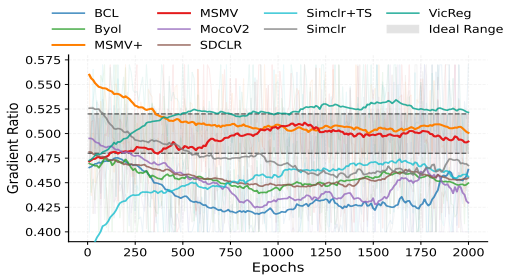


Figure 5: SimCLR and MSMV analysis (NA-Normalized, Augmented pair), (AA-Augmented, Augmented pair)

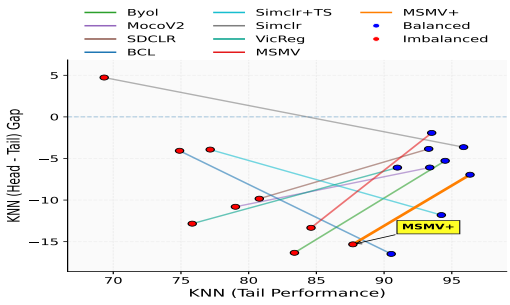
Pretraining To underline the precision of our approach, we present and validate our results on the ImageNet dataset. We used Resent18 He et al. (2016) and Resnet50 He et al. (2016) as our backbone architectures. We employed a three-layer MLP consisting of three fully connected layers with three batch normalization units and two RELU activations for the projection head. We used a batch size of 1024 for Cifar10-LT and Cifar100-LT, while a batch size of 256 was used for Imagenet-LT. We employ stochastic gradient descent with a learning rate of 3.0



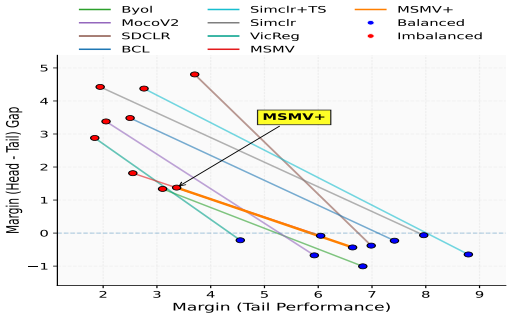
(a) Analysis of gradient Ratio on balanced data



(b) Analysis of gradient Ratio on imbalanced data



(c) KNN ((head minus tail) versus tail) comparison in balanced and imbalanced data.



(d) Margin ((head minus tail) versus tail) comparison in balanced and imbalanced data.

Figure 4: Performance of the proposed MSMV and MSMV+FCL(MSMV+) with respect to Figure 1. From Figure 1(a) and 1(b), we can clearly identify that the only method that is consistent is MSMV and MSMV+ when there is a shift from balanced to imbalanced data. From Figures 1(c) and 1(d) it can be inferred that MSMV+ (highlighted) has better representation quality as compared to others.

for Cifar10-LT and Cifar100-LT, while 0.5 for Imagenet-LT subsamples made by sampling factor $s = 0.125, 0.25, 0.50$ and 0.1 for remaining subsamples $s = 0.75, 1.0$. We used a cosine decay learning rate with a warming of 10% of the max epochs with a weight decay of $1e-4$ to learn a more robust representation. To update the weights of the target network ($encoder_k$), we used a momentum of 0.9. All metrics reported here are trained with 300 epochs in Imagenet-LT and 2000 epochs in Cifar10-LT and Cifar100-LT, and an average of three runs. To eliminate extreme features by the limit of (low) λ_l, λ_h $(high)$, we use $\lambda_l = 0.1$ and $\lambda_h = 0.9$ for Cifar10-LT and Cifar100-LT. For Imagenet-LT and their subsets, we use $\lambda_l = 0.1$ and $\lambda_h = 1.0$. We observed that we must carefully use this limit as it may result in numerical instability. We noticed that λ_h is more sensitive to numerical instability.

Linear Evaluation We demonstrate the model's versatility by freezing the model and letting the linear layer learn for our downstream task. The linear layer, with dimensions of (feature size, number of classes), is a testament to this adaptability. We adopted cosine learning weight decay without warm-up with 10% of the max epochs and a learning rate of 0.005 for this task. Detailed results of the linear evaluation of various methods are given in Tables 1- 3.

Table 1: Cifar10-LT (exponential) Imbalance Dataset Results(Resnet-18)

Model	Imbalance Type - exp				
	Acc↑	Frequent↑	Medium↑	Rare↑	Std↓
MoCoV2 He et al. (2020)	74.76	80.70	74.36	69.8	4.46
Byol Grill et al. (2020)	79.33	82.63	80.56	75.95	2.79
VicReg Bardes et al. (2022a)	73.32	76.26	74.46	70.27	**2.51**
BCL-I Zhou et al. (2022)	76.48	-	-	-	-
SimCLR Chen et al. (2020a)	76.77	82.16	76.9	71.20	4.47
SDCLR Jiang et al. (2021)	80.49	**88.30**	78.07	75.10	5.66
FASSL Lin et al. (2023)	80.69	86.55	76.30	78.80	4.23
SimSiam Ren et al. (2022)	81.40	-	-	-	-
MSMV+(Concatenate)	**81.22**	85.10	80.36	78.20	2.69
MSMV+(Sum)	**83.66**	87.30	**84.36**	**80.40**	2.82

Table 2: Cifar100-LT (exponential) Imbalance Dataset Results(Resnet-18)

Model	Imbalance Type - exp				
	Acc↑	Frequent↑	Medium↑	Rare↑	Std↓
SimCLR Chen et al. (2020a)	44.85	47.81	41.48	44.17	2.59
MoCoV2 He et al. (2020)	46.37	-	-	-	-
Byol Grill et al. (2020)	53.43	55.63	52.03	52.64	**1.57**
VicReg Bardes et al. (2022a)	45.26	48.21	43.27	44.35	2.12
BCL-I Zhou et al. (2022)	52.22	55.35	53.03	48.27	2.95
SDCLR Jiang et al. (2021)	54.94	58.79	55.03	51.00	3.18
FASSL Lin et al. (2023)	55.27	57.74	54.52	53.55	1.79
MSMV+(Concatenate)	**58.89**	**61.54**	**56.18**	**58.94**	2.19
MSMV+(Sum)	**60.18**	**62.24**	**56.75**	**61.50**	2.42

Table 3: Imagenet-LT subsamples results (Resnet-50)

Model	Subsampling Ratio (s)				
	s=1	s=0.75	s=0.50	s=0.25	s=0.125
SimCLR Chen et al. (2020a)	46.81	-	-	-	-
MoCoV2 He et al. (2020)	49.50	43.20	39.5	36.60	30.50
SDCLR Jiang et al. (2021)	46.62	-	-	-	-
Byol Grill et al. (2020)	43.16	-	-	-	-
VicReg Bardes et al. (2022a)	38.24	-	-	-	-
MSMV+(Sum)	**52.90**	**50.25**	**49.04**	**45.51**	**39.68**

7 Conclusions and Future Work

We propose a robust *Multi-Source Multi-View* (MSMV) framework for contrastive self-supervised learning (CSSL) to address the problem of dataset imbalance. We start with an in depth experimental study related to gradient allocation patterns across head and tail classes in SSL. From this study we found that gradient bias is the root cause of lower representation quality and performance degradation in the case of imbalanced data. An analysis of the contrastive objective with respect to imbalanced data revealed that it is heavily dependent on the sample frequency of each class. Our further investigation lead us to the understanding that many classes can be visualized in the backdrop of shared(inter relation) and class specific (intra relation). Based on this division of inter and intra relationship that can exist among different classes, our proposed MSMV framework is able to mitigate the frequency-dependent bias. By eliminating the frequency dependent bias we are able to balance the gradients which in turn results in the robust performance of our framework in the case of imbalanced data. We evaluated the proposed model on various imbalanced datasets (Cifar10-LT, Cifar100-LT, Imagenet-LT(1K)) and achieved state-of-the-art results. Our code is available at https://github.com/anonymouspersoncscv/MSMV.git.

REFERENCES

Jianhong Bai, Zuozhu Liu, Hualiang Wang, Jin Hao, Yang Feng, Huanpeng Chu, and Haoji Hu. On the effectiveness of out-of-distribution data in self-supervised long-tail learning. In *ICLR*, 2023. URL https://openreview.net/forum?id=v8JIQdiN9Sh.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022a. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features. *Advances in Neural Information Processing Systems*, 35:8799–8810, 2022b.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020a.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020b.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020c. URL https://arxiv.org/abs/2003.04297.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9640–9649, 2021.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=B1xwcyHFDr.

Borja Rodrıguez Gálvez, Arno Blaas, Pau Rodríguez, Adam Golinski, Xavier Suau, Jason Ramapuram, Dan Busbridge, and Luca Zappella. The role of entropy and reconstruction in multi-view self-supervised learning. In *International Conference on Machine Learning*, pp. 29143–29160. PMLR, 2023.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Proceedings of Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Ziyu Jiang, Tianlong Chen, Bobak J Mortazavi, and Zhangyang Wang. Self-damaging contrastive learning. In *International Conference on Machine Learning*, pp. 4927–4939. PMLR, 2021.

Pranay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. `http://www.deeplearningbook.org`.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Anna Kukleva, Moritz Böhle, Bernt Schiele, Hilde Kuehne, and Christian Rupprecht. Temperature schedules for self-supervised contrastive methods on long-tail data. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL `https://openreview.net/forum?id=ejHUr4nfHhD`.

Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering*, 31(10):1863–1883, 2018.

Ci-Siang Lin, Min-Hung Chen, and Yu-Chiang Frank Wang. Frequency-aware self-supervised long-tailed learning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pp. 963–972. IEEE, 2023. doi: 10.1109/ICCVW60793.2023.00103. URL `https://doi.org/10.1109/ICCVW60793.2023.00103`.

Hong Liu, Jeff Z. HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL `https://openreview.net/forum?id=4AZz9osqrar`.

Christos Louizos, Matthias Reisser, and Denis Korzhenkov. A mutual information perspective on federated contrastive learning. *arXiv preprint arXiv:2405.02081*, 2024.

Sucheng Ren, Huiyu Wang, Zhengqi Gao, Shengfeng He, Alan Yuille, Yuyin Zhou, and Cihang Xie. A simple data mixing prior for improving self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14595–14604, 2022.

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

Karthik Sridharan and Sham M Kakade. An information theoretic framework for multi-view learning. In *COLT*, number 114, pp. 403–414, 2008.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Proceedings of ECCV 2020*, volume 12356 of *Lecture Notes in Computer Science*, pp. 776–794. Springer, 2020a.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020b.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop*, 2015. doi: 10.1109/ITW.2015.7133169.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=-bdp_8Itjwp.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

Shiye Wang, Changsheng Li, Yanming Li, Ye Yuan, and Guoren Wang. Self-supervised information bottleneck for deep multi-view subspace clustering. *IEEE Transactions on Image Processing*, 32: 1555–1567, 2023.

Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. *Advances in neural information processing systems*, 33:19290–19301, 2020.

Zhihan Zhou, Jiangchao Yao, Yan-Feng Wang, Bo Han, and Ya Zhang. Contrastive learning with boosted memorization. In *International Conference on Machine Learning*, pp. 27367–27377. PMLR, 2022.

.1 Notations

.2 Elimination by Our loss function:

In order to give a detailed explanation for the MSMV loss (Eq. 8), we provide additional information. In Fig. 9, each image vividly demonstrates our loss in flow. In the first figure (Fig. 6), we have the similarities of pairs as illustrated in Fig. **??**. The white circles are those that are considered for the loss calculation as they lie within the threshold while on the other hand, black circles lie outside the threshold and denotes those extreme cases which need not be considered. The vertical and horizontal lines represents the threshold for extreme cases. Later, in the second figure (Fig. 7) and third figure (Fig. 8), our loss function continuously eliminates the features that are not necessary for learning, providing a clear visual representation of the data flow.

$$
\begin{aligned}
\text{Let } S &= Sim\left((z_A^n \circledast z_C^a).(z_A^a \circledast z_C^n)\right) \\
S(i,j) &= \{S_{i,j}.1_{[\lambda_l,\lambda_h]}S_{i,j}\} \\
\text{where } 1_{[\lambda_l,\lambda_h]}S_{i,j} &= \begin{cases} \mathcal{S}_{i,j}, & \text{if } \lambda_l \leq S_{i,j} \leq \lambda_h \\ 0, & \text{otherwise} \end{cases} \\
\text{Final Loss: } \mathcal{L}_{i,j} &= -log\left[\frac{e^{S/\tau}}{\sum_{k=1}^{2N} 1_{\{i\neq k\}} e^{S/\tau}}\right]
\end{aligned} \tag{8}
$$

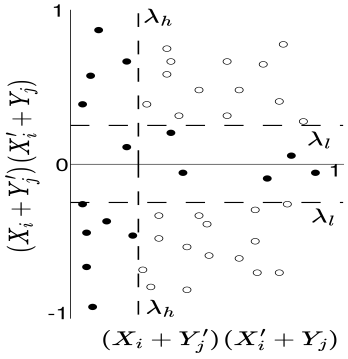


Figure 6: Flow-1

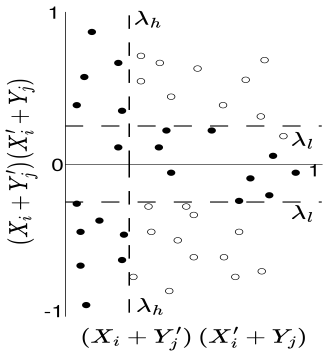


Figure 7: Flow-2

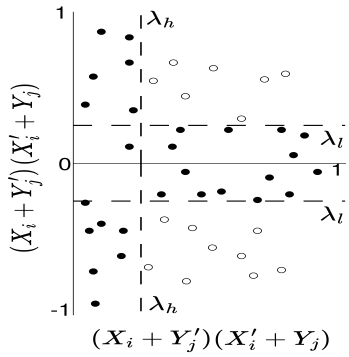


Figure 8: Flow-3

Figure 9: Cosine Similarity of MSMV pairs. Fig. 6 shows the initial cosine similarity of MSMV pairs, where λ_h and λ_l works as filter to cosider good examples for loss calculation. Later Fig. 7, and Fig. 8 shows the upcoming iterations, in which more similarity pairs are filtered out by the loss function.

.2.1 Loss threshold analysis (λ analysis)

To find the best possible threshold, several experiments were carried out using the Cifar10-LT, Cifar100-LT, ImageNet-LT dataset using different combinations of λ_h and λ_l as can be seen from 10. The values used for λ_h range from 1 to 0.7 and for λ_l range from 0 to 0.4. We observe that a higher threshold (lower λ_h and higher λ_l) results in unstable learning, while a lower threshold (higher λ_h and lower λ_l)) does not improve performance. The values of 0.1 (λ_l) and 0.9 (λ_h) give the best result, as highlighted in Fig. 10 and tab:Cifar10-LT˙sensitivity,tab:Cifar100-LT˙sensitivity. However, the value of 0.1 (λ_l) and 1.0 (λ_h) give the best result on ImageNet-LT as highlighted in Table. 6. Cifar10/100 experiments were carried out over 500 epochs, while ImageNet-LT were carried out over 300 epochs. For large data sets such as ImageNet-LT, we observe that $\lambda_l(0.1)$ is more crucial than $\lambda_h(1.0)$ as λ_h is more sensitive to numerical instability.

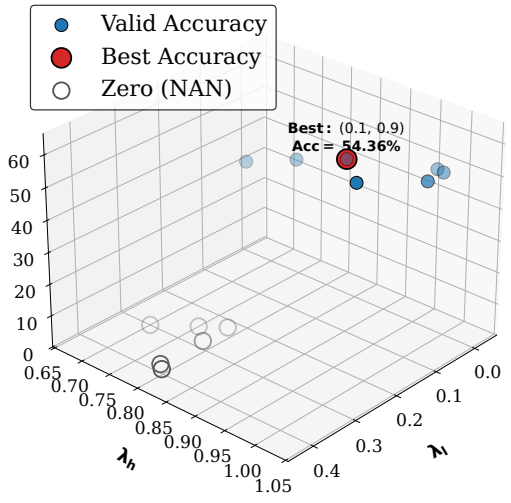


Figure 10: Threshold λ_h and λ_l analysis

Table 4: λ sensitivity analysis on Cifar10-LT

$\boldsymbol{\lambda}_\Gamma 30572$ - $\boldsymbol{\lambda}_\Gamma 30568$	**Acc**$\uparrow$
0.01 - 1.0	73.80
0.01 - 0.99	74.20
0.05 - 1.0	73.08
0.15 - 0.95	76.02
0.1 - 0.9	**76.83**
0.15 - 0.85	75.67
0.20 - 0.8	74.82
0.30 - 0.7	NAN
0.25 - 0.75	NAN
0.38 - 0.78	NAN
0.39 - 0.79	NAN
0.28 - 0.78	NAN
0.22 - 0.78	NAN

Table 5: λ sensitivity analysis on Cifar100-LT

$\boldsymbol{\lambda}_\Gamma 30572$ - $\boldsymbol{\lambda}_\Gamma 30568$	**Acc**$\uparrow$
0.01 - 1.0	51.80
0.01 - 0.99	52.09
0.05 - 1.0	51.67
0.15 - 0.95	54.00
0.1 - 0.9	**54.36**
0.15 - 0.85	53.78
0.20 - 0.8	52.73
0.30 - 0.7	NAN
0.25 - 0.75	NAN
0.38 - 0.78	NAN
0.39 - 0.79	NAN
0.28 - 0.78	NAN
0.22 - 0.78	NAN

Table 6: λ sensitivity analysis on ImageNet-LT

$\boldsymbol{\lambda}_\Gamma 30572$ - $\boldsymbol{\lambda}_\Gamma 30568$	**Acc**$\uparrow$
0.01 - 1.0	50.80
0.01 - 0.99	48.32
0.05 - 1.0	51.67
0.15 - 0.95	NAN
0.1 - 1.0	**52.90**
0.05 - 0.9	NAN
0.1 - 0.9	NAN

.3 Significance of proposed loss function with MSMV for representation learning

Fig. 11- 14 show the advantage of the proposed loss function over the traditional NT-Xent loss. In these figures, MSMV denotes MSMV plus NT-Xent loss, while MSMV(updated) refers to MSMV plus the proposed loss. These experiments were carried out on Cifar100-LT for 2000 epochs. An imbalance factor of 0.01 with exponential distribution is used in Cifar100 to generate Cifar100-LT. Cifar100-LT is divided into three parts, *Frequent, Medium, Few*. Frequent contains classes that have the highest number of images and is considered as head. Similarly, medium contains less number of images when compared to the head and few denotes the tail classes having the least number of

images[1]. From the figures (Figs. 11- 14), it can be observed that in all scenarios the proposed loss function exhibits superior behavior compared to the NT-Xent loss and performs best when it comes to tail classes (Fig. 13).

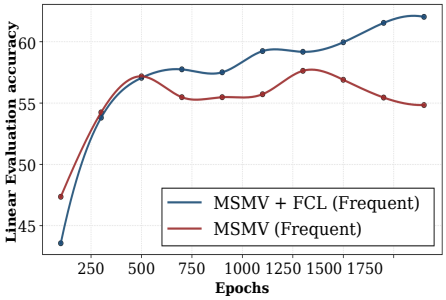


Figure 11: Frequent Classes overview

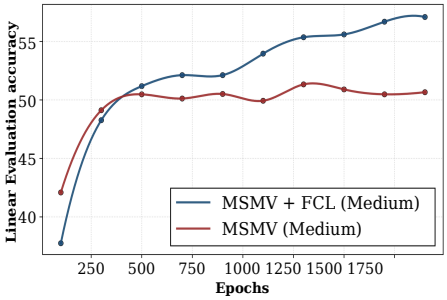


Figure 12: Medium Classes overview

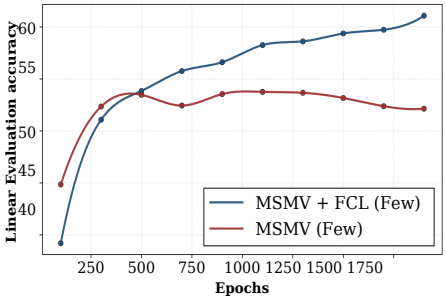


Figure 13: Few Classes overview

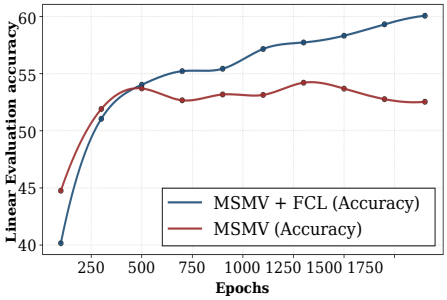


Figure 14: Linear Evaluation accuracy overview

.4 Alignment and Uniformity analysis

In order to measure representation quality, two measures (alignment and uniformity) have been suggested by Tongzhou Wang & Philip Isola (Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere, ICML-2020). The pattern of pulling positive examples closer (`alignment`) while pulling apart negatives that are irrelevant (`uniformity`) are two key properties that any contrastive loss function should satisfy. From Fig. 15- 17, it can be seen that the proposed MSMV framework with updated loss satisfies both alignment and uniformity compared to other SSL frameworks.

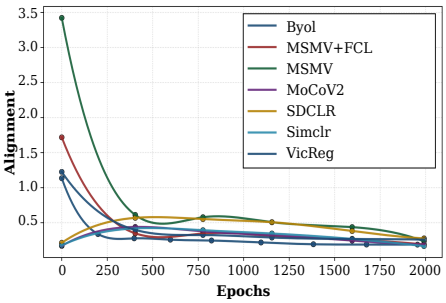


Figure 15: Alignment overview

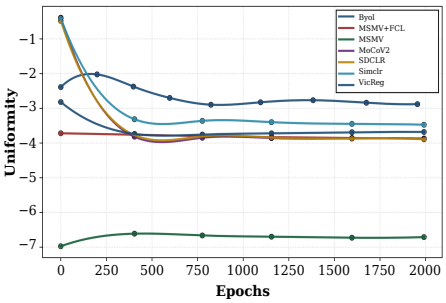


Figure 16: Uniformity overview

[1]For imbalanced factor 0.01, generally frequent classes are the first 36 classes, medium is next 34 classes and few consist of the last 30 classes.

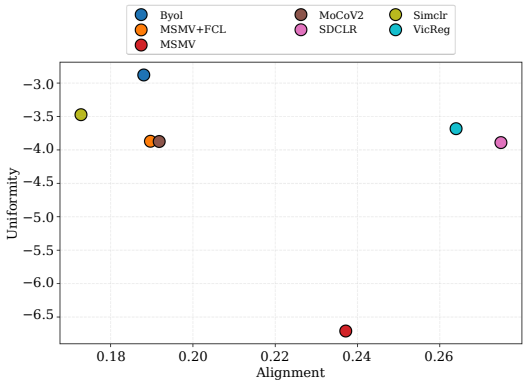


Figure 17: Alignment and Uniformity analysis overview

.5 T-SNE REPRESENTATIONS

Figures, (Figs. 18- 24) show t-SNE representations of embeddings related to Cifar10-LT. We can visualize from these figures that the proposed framework (Fig. **??**) has better discriminative representation while MSMV plus the proposed loss (Fig. 21) has even better discriminative ability.

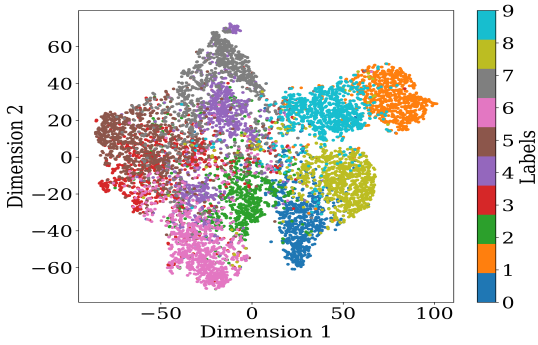


Figure 18: MSMV

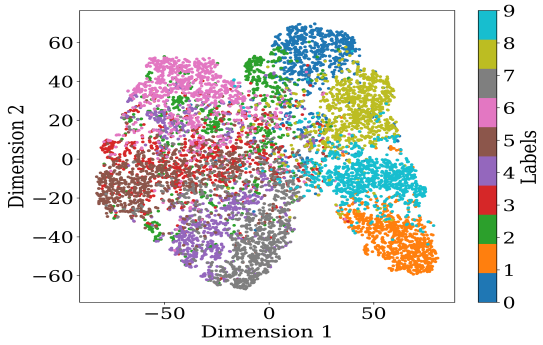


Figure 19: BYOL

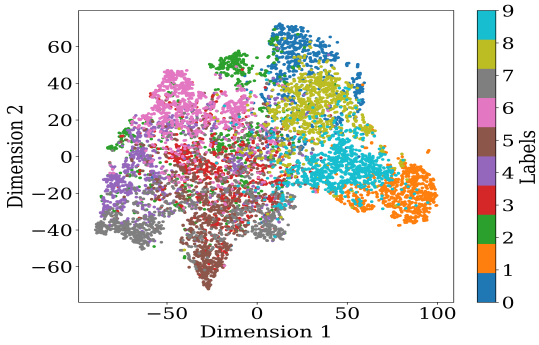


Figure 20: VicReg

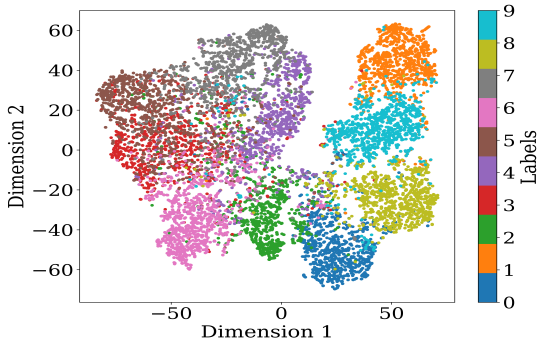


Figure 21: MSMV+FCL

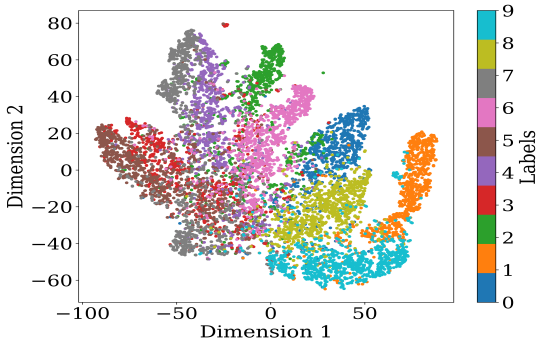


Figure 22: SDCLR

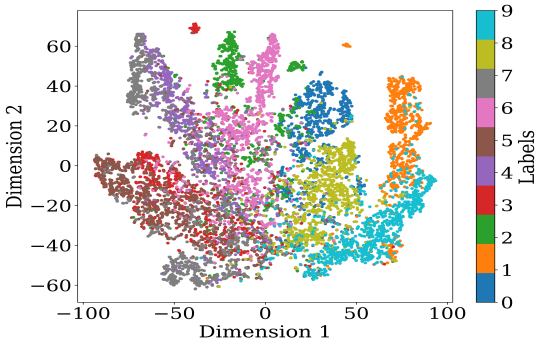


Figure 23: SimCLR

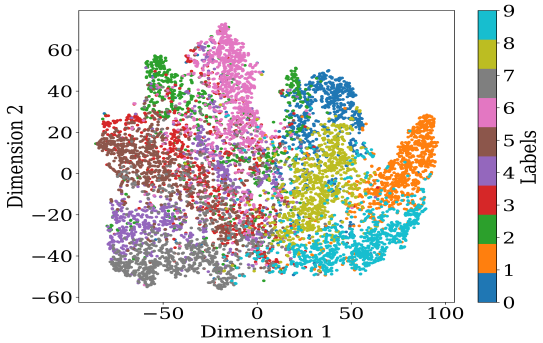


Figure 24: MoCoV2

.6 PER CLASS ACCURACY ANALYSIS ON CIFAR-10

The per class accuracy analysis on Cifar10-LT is given in Fig. 25. It can be observed that the proposed MSMV plus updated loss is consistently having better results over each class.

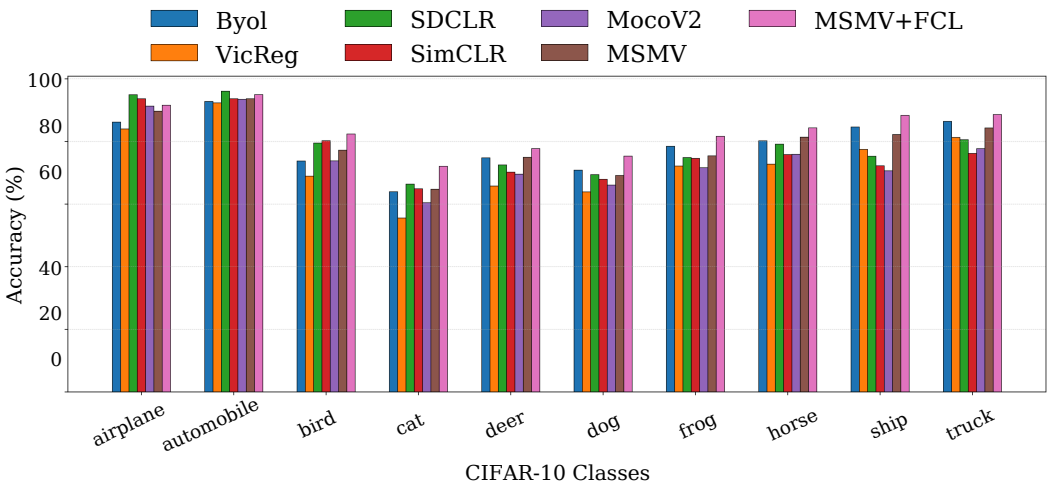


Figure 25: Per class accuracy on Cifar10-LT with SSL models and ResNet-18 as backbone

.7 MUTUAL INFORMATION WITH NT-XENT LOSS

In order to prove Mutual Information as given in Eq. (9), we provide the supporting proof as follows: When we reexamine the NT-Xent environment, it can be seen that two positive pairs are chosen among $2N$ samples. This happens because NT-Xent averages the loss in both directions, such as (z_i, z_j) and (z_j, z_i).

$$\mathcal{L}_{i,j} = -\log \frac{\exp\left(\text{sim}(z_i, z_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\text{sim}(z_i, z_k)/\tau\right)} \tag{9}$$

Mutual information between X and Y is defined as:

$$I(X;Y) = \mathbb{E}_{(p_x, p_y)} \left[\log \frac{p(x,y)}{p(x)p(y)}\right] \tag{10}$$

This measures how much knowing X reduces uncertainty about Y and vice versa.

Let x be an anchor image and our objective is to find which is the true positive out of N samples.

Formulating as a classification problem:

Given an anchor image x, we want to predict which of the samples N is the true positive y (the one drawn from the joint $p(x, y)$). The other $2N - 2$ samples are drawn from the marginal $p(y)$ — i.e., they're negatives.

Let $(x, y_1) \sim p(x, y)$ be a true pair and let $y_2, \ldots, y_{2N-1} \sim p(y)$ be negatives, sampled independently of x. Let $y_{1:2N-1} = \{y_1, \ldots, y_{2N-1}\}$ be the complete set of candidates.

We define a classification task: Given x and $\{y_1, \ldots, y_{2N-1}\}$, predict which is the positive.

This gives:

$$p(i = 1 | x, y_{1:2N-1}) = \frac{p(x, y_1)}{\sum_{k=1}^{2N-1} p(x)p(y_k)}$$

Since we do not know the exact distributions $p(x, y), p(x), p(y)$, we approximate this ratio with the Radon-Nikodym derivative:

$$f(x, y) \propto \frac{p(x, y)}{p(x)p(y)} \tag{11}$$

Our critic function in NT-Xent loss is defined as:

$$f(x, y) = \exp(\text{sim}(x, y)/\tau)$$

Using this, we can reshape our softmax classifier as follows:

$$q(i = 1 | x, y_{1:2N-1}) = \frac{f(x, y_1)}{\sum_{k=1}^{2N-1} f(x, y_k)}$$

NT-Xent loss is the negative log-likelihood of the correct class:

$$\mathcal{L}_{i,j} = -\mathbb{E}_{(x,y_1)\sim p(x,y), \{y_k\}_{k=1}^{2N-1} \sim p(y)} \left[\log \frac{f(x, y_1)}{\sum_{k=1}^{2N-1} f(x, y_k)}\right] \tag{12}$$

Now, consider:

$$I(X;Y) = \mathbb{E}_{p^-} \left[\log \frac{p(x,y)}{p(x)p(y_1)}\right]$$

We approximate the marginals using a Monte Carlo estimator.

$$p(x)p(y_1) \approx \frac{1}{2N-1} \sum_{k=1}^{2N-1} p(x)p(y_k)$$

Thus:

$$\frac{p(x,y_1)}{p(x)p(y_1)} \approx \frac{p(x,y_1)}{\frac{1}{2N-1}\sum_{k=1}^{2N-1} p(x)p(y_k)}$$

Taking log:

$$\begin{aligned}\log\left[\frac{p(x,y_1)}{p(x)p(y_1)}\right] &= \log\left[\frac{p(x,y_1)}{\frac{1}{2N-1}\sum_{k=1}^{2N-1} p(x)p(y_k)}\right] \\ &= \log(2N-1) + \log\left[\frac{p(x,y_1)}{\sum_{k=1}^{2N-1} p(x)p(y_k)}\right]\end{aligned}$$

By Jensen's inequality (since log is concave):

$$\log \mathbb{E}[Z] \geq \mathbb{E}[\log Z] \tag{13}$$

Apply this and substitute from equation 12:

$$\begin{aligned}\mathbb{E}_{p^-}\left[\log \frac{p(x,y_1)}{\frac{1}{2N-1}\sum_{k=1}^{2N-1} p(x)p(y_k)}\right] &\geq \mathbb{E}_{p^-}\left[\log \frac{f(x,y_1)}{\frac{1}{2N-1}\sum_{k=1}^{2N-1} f(x,y_k)}\right] \\ I(X;Y) &\geq \log(2N-1) + \mathbb{E}_{p^-}\left[\log \frac{f(x,y_1)}{\sum_{k=1}^{2N-1} f(x,y_k)}\right] \\ &= \log(2N-1) - \mathcal{L}_{\text{NT-Xent}}\end{aligned}$$

$$I(X;Y) = log(N) - \mathcal{L}_{NT_Xent} + c \tag{14}$$

.8 Mutual Information with MSMV loss

In the MSMV environment, we can see that two complementary representations $(z_A^n, z_C^a), (z_A^n, z_C^a)$ of normalized and augmented versions of two images are chosen as positive among the 2N samples. (z_A^n, z_A^a) are the representations of $x_A^n, x_A^a \in x_A$(anchor) and (z_A^n, z_A^a) are the representations of $x_C^n, x_C^a \in x_C$(counterpart).

$$\mathcal{L}_{i,j} = \frac{e^{sim((z_A^n, z_C^a),(z_A^a, z_C^n))/\tau}}{\sum_{k=1}^{2N} 1_{i\neq k}\, e^{sim((z_A^n, z_C^a),(z_A^a, z_C^n))/\tau}}$$

Formulating a classification problem: Given an anchor image x_A and a counterpart image x_C their representations can be given as $z_A^n, z_C^a \in x_A$ and $z_C^n, z_C^a \in x_C$. Later, they form a pair $(x_A^n \circledast y_C^a)$ of representations, and we want to predict which of the N samples is the true positive $(z_A^n \circledast z_C^a)$ (the one drawn from the joint $p(z_A^n \circledast z_C^a, z_A^a \circledast z_C^n)$). The other $2N-2$ samples are drawn from the marginal $p(z_A^a \circledast z_C^n)$ — i.e., they are negatives.

Let $(z_A^n \circledast z_C^a, z_A^a \circledast z_C^n) \sim p(z_A^n \circledast z_C^a, z_A^a \circledast z_C^n)$ be a true pair and let $(z_A^a \circledast z_C^n)_2, \ldots, y_{2N-1} \sim p(z_A^a \circledast z_C^n)$ be negatives, sampled independently of $z_A^n \circledast z_C^a$. Let $(z_A^a \circledast z_C^n)_{1:2N-1} = \{(z_A^a \circledast z_C^n)_1, \ldots, (z_A^a \circledast z_C^n)_{2N-1}\}$ be the complete set of candidates.

We define a classification task: Given $z_A^n \circledast z_C^n$ and $\{(z_A^a \circledast z_C^n)_1, \ldots, y_{2N-1}\}$, predict which is the positive.



Figure 26: Top-5 image retrieval for the query "Dog" by SimCLR



Figure 27: Top-5 image retrieval for the query "Cat" by SimCLR

This gives:

$$p(i=1|(z_A^n \circledast z_C^a),(z_A^a \circledast z_C^n)_{1:2N-1}) = \frac{p((z_A^n \circledast z_C^a),(z_A^a \circledast z_C^n)_1)}{\sum_{k=1}^{2N-1} p(z_A^n \circledast z_C^a)p(z_A^a \circledast z_C^n)_k}$$

Since we do not know the exact distributions $p((z_A^n \circledast z_C^a),(z_A^a \circledast z_C^n)), p((z_A^n \circledast z_C^a)), p((z_A^a \circledast z_C^n))$, we approximate this ratio with the Radon-Nikodym derivative:

$$f((z_A^n \circledast z_C^a),(z_A^a \circledast z_C^n)) \propto \frac{p((z_A^n \circledast z_C^a),(z_A^a \circledast z_C^n))}{p((z_A^n \circledast z_C^a))p((z_A^a \circledast z_C^n))}$$

Our critic function in MSMV loss is defined as:

$$f((z_A^n \circledast z_C^a),(z_A^a \circledast z_C^n)) = \exp(\text{sim}((z_A^n \circledast z_C^a),(z_A^a \circledast z_C^n))/\tau)$$

Using this, we can reshape our softmax classifier as follows:

$$q(i=1|(z_A^n \circledast z_C^a),(z_A^a \circledast z_C^n)_{1:2N-1}) = \frac{f((z_A^n \circledast z_C^a),(z_A^a \circledast z_C^n)_1)}{\sum_{k=1}^{2N-1} f((z_A^n \circledast z_C^a),(z_A^a \circledast z_C^n)_k)}$$

Since MSMV loss is the negative log-likelihood of the correct class:

$$\mathcal{L}_{i,j} = -\mathbb{E}_{((z_A^n \circledast z_C^a),z_C^n) \sim p((z_A^n \circledast z_C^a),(z_A^a \circledast z_C^n)),\{(z_A^a \circledast z_C^n)_k\}_{k=1}^{2N-1} \sim p((z_A^a \circledast z_C^n))} \left[\log \frac{f((z_A^n \circledast z_C^a),(z_A^a \circledast z_C^n)_1)}{\sum_{k=1}^{2N-1} f((z_A^n \circledast z_C^a),(z_A^a \circledast z_C^n)_k)}\right] \tag{15}$$

Since the classification formulation reveals that the negative log-liklelihood of MSMV loss 15 is an instance of the negative log-likelihood of the NT-Xent loss 12, its mutual information lower bound is derived as follows:

$$\begin{aligned} I((z_A^n \circledast z_C^a);(z_A^a \circledast z_C^n)) &\geq \log(2N-1) - \mathcal{L}_{\text{MSMV}} \\ I((z_A^n \circledast z_C^a),(z_A^a \circledast z_C^n)) &\geq log(N) - \mathcal{L}_{MSMV} + c \end{aligned} \tag{16}$$

Our code is available at https://github.com/anonymouspersoncscv/MSMV.git.

.9 Top-K Retrieval



Figure 28: Top-5 image retrieval for the query ”Elephant” by SimCLR



Figure 29: Top-5 image retrieval for the query ”Eagle” by SimCLR

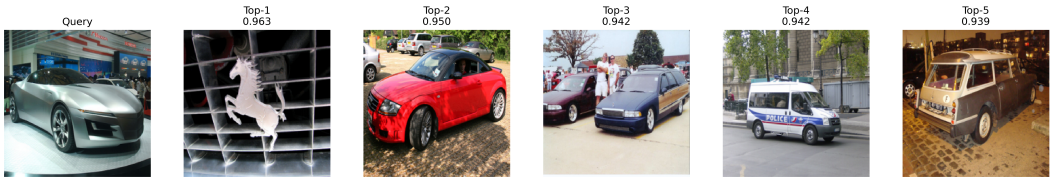


Figure 30: Top-5 image retrieval for the query ”Car” by SimCLR

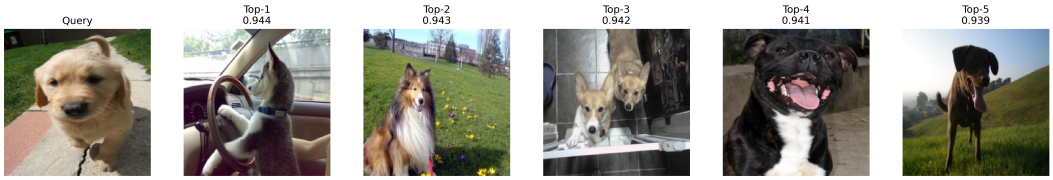


Figure 31: Top-5 image retrieval for the query ”Dog” by Byol



Figure 32: Top-5 image retrieval for the query ”Cat” by Byol



Figure 33: Top-5 image retrieval for the query ”Elephant” by Byol

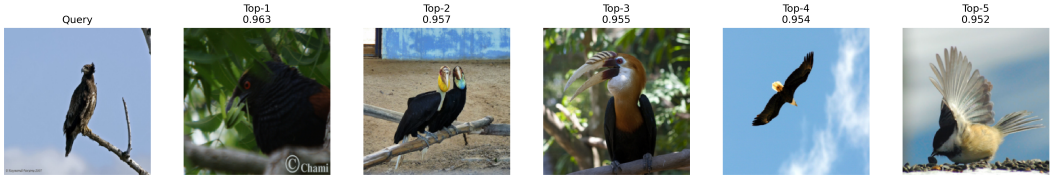


Figure 34: Top-5 image retrieval for the query ”Eagle” by Byol



Figure 35: Top-5 image retrieval for the query ”Car” by Byol



Figure 36: Top-5 image retrieval for the query ”Dog” by SDCLR

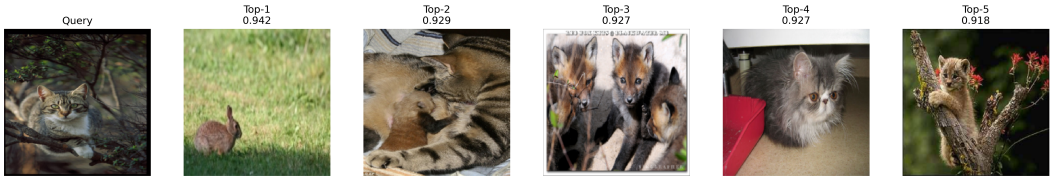


Figure 37: Top-5 image retrieval for the query ”Cat” by SDCLR

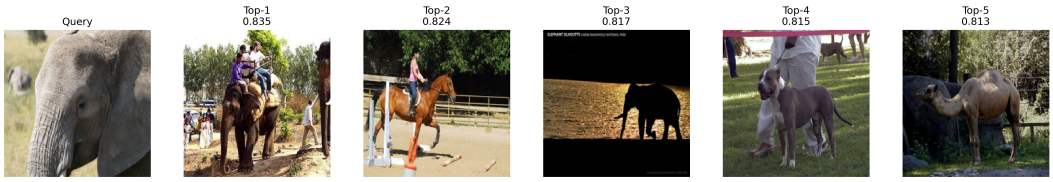


Figure 38: Top-5 image retrieval for the query ”Elephant” by SDCLR

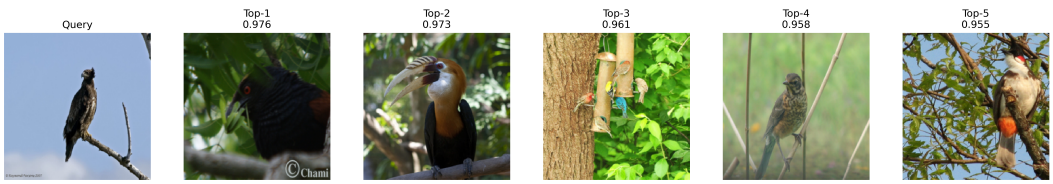


Figure 39: Top-5 image retrieval for the query ”Eagle” by SDCLR



Figure 40: Top-5 image retrieval for the query ”Car” by SDCLR

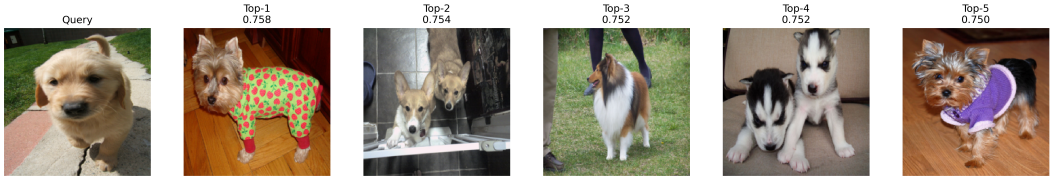


Figure 41: Top-5 image retrieval for the query ”Dog” by the proposed method MSMV + FCL

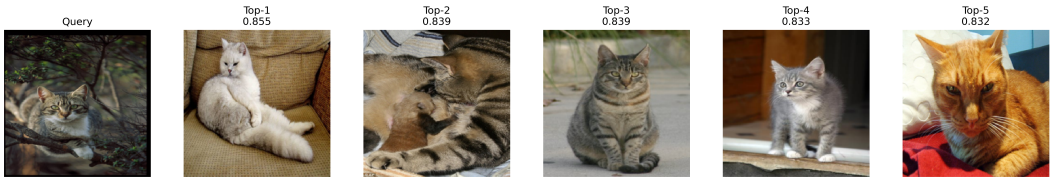


Figure 42: Top-5 image retrieval for the query ”Cat” by the proposed method MSMV + FCL

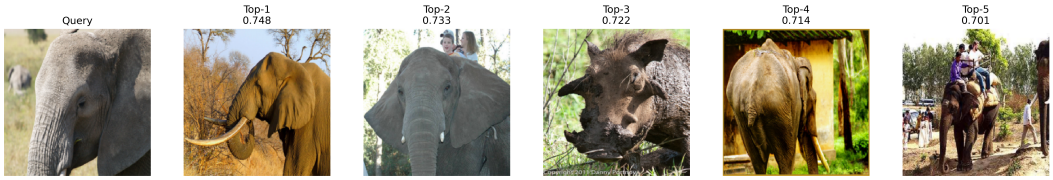


Figure 43: Top-5 image retrieval for the query ”Elephant” by the proposed method MSMV + FCL



Figure 44: Top-5 image retrieval for the query ”Eagle” by the proposed method MSMV + FCL



Figure 45: Top-5 image retrieval for the query ”Car” by the proposed method MSMV + FCL

.10 Computing Head and Tail Gradients

Let $\mathcal{H}$ and $\mathcal{T}$ denote the sets of anchors belonging to the head and tail classes, respectively. Since each image generates two augmented views, every sample contributes two anchors to the loss. The per-anchor contrastive losses are first grouped according to their corresponding class labels.

The average loss for the head and tail groups is computed as

$$\mathcal{L}_{\text{head}} = \frac{1}{|\mathcal{H}|} \sum_{k \in \mathcal{H}} \ell_k, \tag{17}$$

$$\mathcal{L}_{\text{tail}} = \frac{1}{|\mathcal{T}|} \sum_{k \in \mathcal{T}} \ell_k, \tag{18}$$

where ℓ_k denotes the contrastive loss associated with the k-th anchor.

The gradients of the encoder parameters θ are computed independently for each group,

$$g_{\text{head}} = \nabla_\theta \mathcal{L}_{\text{head}}, \tag{19}$$

$$g_{\text{tail}} = \nabla_\theta \mathcal{L}_{\text{tail}}. \tag{20}$$

The gradient magnitude of each group is measured using the Euclidean norm,

$$G_{\text{head}} = \|g_{\text{head}}\|_2 , \tag{21}$$

$$G_{\text{tail}} = \|g_{\text{tail}}\|_2 . \tag{22}$$

Finally, the relative optimization contribution of the head classes is quantified by the normalized gradient ratio,

$$R = \frac{G_{\text{head}}}{G_{\text{head}} + G_{\text{tail}} + \epsilon}, \tag{23}$$

where $\epsilon = 10^{-8}$ is a small constant introduced for numerical stability. A value of $R = 0.5$ indicates equal optimization contributions from head and tail samples, while values greater (or smaller) than 0.5 imply that optimization is dominated by head (or tail) samples.

.10.1 Why Balanced Analysis?

Directly measuring gradient norms on the original long-tailed training distribution introduces an inherent sampling bias because head-class samples occur much more frequently than tail-class samples. Consequently, the accumulated gradients are naturally dominated by head classes, making it difficult to determine whether the observed differences arise from the learned representations or simply from unequal sample frequencies.

To eliminate this confounding factor, we construct a balanced analysis subset by selecting an equal number of samples from the predefined head and tail classes,

$$|\mathcal{H}| = |\mathcal{T}|. \tag{24}$$

The balanced subset is used exclusively for gradient analysis and is never used for network optimization. Since both groups contribute an equal number of samples, any difference observed between G_{head} and G_{tail} reflects the intrinsic optimization behavior learned by the model rather than the class imbalance of the training data. Consequently, the measured gradient statistics provide an unbiased estimate of how the trained representation responds to head and tail samples.

.10.2 Why This Setting is Appropriate for Long-Tailed SSL

Although the model is trained on the original long-tailed dataset, gradient analysis is performed on a separate balanced subset while keeping the network parameters fixed. During analysis, gradients are computed through backpropagation solely to measure their magnitudes, and no optimizer update is performed. Formally, the trained parameters remain unchanged,

$$\theta_{\text{analysis}} = \theta_{\text{trained}}, \tag{25}$$

and the parameter update

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L} \tag{26}$$

is intentionally omitted.

This kind of evaluation offers several advantages. First, it tries to make sure that the head and tail classes contribute equally to the analysis. This is done by removing the unequal class frequencies during gradient measurement. Secondly, since the encoder is already optimized using the original long-tailed dataset, the measured gradients capture the optimization dynamics induced by long-tail learning rather than artifacts of balanced training. The third advantage is that the computed gradients do not alter the learned representations because throughout the analysis the model parameters remain fixed. Lastly, this kind of evaluation provides a fair comparison for analyzing the intrinsic optimization behavior towards the head and tail classes as every method is evaluated under the same balanced setting.

.11 Optimization Perspective of Our proposed MSMV framework

Most self supervised methods are inspired from single image multi-view design. SimCLR, MoCo, Byol, VicReg are the most popular methods among them. In this contribution, we focus primarily on contrastive learning, where the encoder learns the representations by maximizing the agreement between both views. In works like SimCLR, MoCo, Byol and VicReg, given two augmented views $\mathbf{x}_1^{a_1}$ and $\mathbf{x}_1^{a_2}$, the encoder produces

$$\mathbf{z}_1^{a_1} = encoder_x_\theta(\mathbf{x}_1^{a_1}), \qquad \mathbf{z}_1^{a_2} = encoder_x_\theta(\mathbf{x}_1^{a_2}), \tag{27}$$

where $encoder_x_\theta$ denotes the encoder parameterized by θ. The SSL objective is

$$\mathcal{L} = \mathcal{L}\left(\mathbf{z}_1^{a_1}, \mathbf{z}_1^{a_2}\right). \tag{28}$$

The encoder parameters are updated as

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial \mathbf{z}_1^{a_1}} \frac{\partial \mathbf{z}_1^{a_1}}{\partial \theta} + \frac{\partial \mathcal{L}}{\partial \mathbf{z}_1^{a_2}} \frac{\partial \mathbf{z}_1^{a_2}}{\partial \theta}. \tag{29}$$

Since

$$\mathbf{z}_1^{a_1} = encoder_x_\theta(\mathbf{x}_1^{a_1}), \qquad \mathbf{z}_1^{a_1} = encoder_x_\theta(\mathbf{x}_1^{a_2}), \tag{30}$$

each representation depends only on its own augmented view. Consequently, the optimization of each view remains independent until the loss is computed. Although the loss compares the two representations, no feature interaction occurs before gradient propagation.

Suppose now that another image $\mathbf{x}_2$ is available as in our proposed work. In our proposed design, we call them an anchor image and it's random counterpart. In our framework, we represent $\mathbf{x_1}$ and $\mathbf{x_2}$ as $\mathbf{x_A}$ (anchor) $\mathbf{x_C}$ (counterpart). These two images give four views, two normalized $\mathbf{x_1^n}, \mathbf{x_2^n}$ and two augmented views $\mathbf{x_1^a}, \mathbf{x_2^a}$. The encoder produces

$$\begin{aligned} \mathbf{z}_A^n &= encoder_x_\theta(\mathbf{x}_A^n), \qquad \mathbf{z}_A^a = encoder_x_\theta(\mathbf{x}_A^a) \\ \mathbf{z}_C^n &= encoder_c_\eta(\mathbf{x}_C^n), \qquad \mathbf{z}_C^a = encoder_c_\eta(\mathbf{x}_C^a). \end{aligned} \tag{31}$$

where $encoder_x_\theta$ and $encoder_c_\eta$ denotes the encoder parameterized by θ and η. Instead of directly optimizing $\mathbf{z}_A$, we first perform fusion

$$\tilde{\mathbf{z}}_{encoder_x} = f(\mathbf{z}_A^n, \mathbf{z}_C^a), \qquad \tilde{\mathbf{z}}_{encoder_c} = f(\mathbf{z}_A^a, \mathbf{z}_C^n), \tag{32}$$

where $f()$ denotes the proposed differentiable fusion operation $\circledast$.

The SSL objective becomes

$$\mathcal{L} = \mathcal{L}(\tilde{\mathbf{z}}_{encoder_x}, \tilde{\mathbf{z}}_{encoder_c}). \tag{33}$$

Applying the chain rule, (We analyze gradients only for the encoder , as the encoder is updated by EMA rather than gradient descent.)

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{z}}_{encoder_x}} \frac{\partial \tilde{\mathbf{z}}_{encoder_x}}{\partial \theta}. \tag{34}$$

Since

$$\tilde{\mathbf{z}}_{encoder_x} = f(\mathbf{z}_A^n, \mathbf{z}_C^a), \tag{35}$$

its derivative expands as

$$\frac{\partial \tilde{\mathbf{z}}_{encoder_x}}{\partial \theta} = \frac{\partial \tilde{\mathbf{z}}_{encoder_x}}{\partial \mathbf{z}_A^n} \frac{\partial \mathbf{z}_A^n}{\partial \theta} + \frac{\partial \tilde{\mathbf{z}}_{encoder_x}}{\partial \mathbf{z}_C^a} \frac{\partial \mathbf{z}_C^a}{\partial \theta}. \tag{36}$$

Substituting the above expression yields

$$\boxed{\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{z}}_{encoder_x}} \left(\frac{\partial \tilde{\mathbf{z}}_{encoder_x}}{\partial \mathbf{z}_A^n} \frac{\partial \mathbf{z}_A^n}{\partial \theta} + \frac{\partial \tilde{\mathbf{z}}_{encoder_x}}{\partial \mathbf{z}_C^a} \frac{\partial \mathbf{z}_C^a}{\partial \theta} \right).} \tag{37}$$

Equation 37 highlights the fundamental difference between conventional SSL and our proposed framework. Without fusion,

$$\frac{\partial \tilde{\mathbf{z}}_{encoder_x}}{\partial \mathbf{z}_C^a} = 0, \tag{38}$$

and therefore

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial \mathbf{z}_\mathbf{A}^\mathbf{n}} \frac{\partial \mathbf{z}_\mathbf{A}^\mathbf{n}}{\partial \theta}, \tag{39}$$

meaning that each sample contributes only its own optimization signal.

In contrast, with fusion,

$$\frac{\partial \tilde{\mathbf{z}}_{encoder_x}}{\partial \mathbf{z}_C^a} \neq 0, \tag{40}$$

which introduces an additional optimization pathway through the interacting sample. Consequently, the parameter update is jointly influenced by both representations,

$$\underbrace{\frac{\partial \tilde{\mathbf{z}}_{encoder_x}}{\partial \mathbf{z}_A^n}\frac{\partial \mathbf{z}_A^n}{\partial \theta}}_{\text{Self Contribution}} + \underbrace{\frac{\partial \tilde{\mathbf{z}}_{encoder_x}}{\partial \mathbf{z}_C^a}\frac{\partial \mathbf{z}_C^a}{\partial \theta}}_{\text{Fusion Contribution for counterpart image}} . \tag{41}$$

Therefore, unlike conventional multi-view SSL where each image is optimized independently until the loss computation, fusion couples the optimization dynamics of samples. Each sample is updated using not only its own visual evidence, but also complementary semantic information extracted from another image.

For long-tailed self-supervised learning, this distinction is particularly important. Tail classes often contain only a few training instances, limiting the semantic diversity available to each individual sample. In conventional SSL, each tail sample therefore receives an optimization signal derived solely from its own representation. Our proposed framework enriches this optimization process by allowing semantically related samples to exchange complementary information before the loss is computed. As a result, the encoder receives a richer optimization signal, improving semantic coverage, reducing the effect of sample scarcity, and producing more informative representations for tail classes.

.11.1 Why FCL becomes the natural option to go with MSMV?

Our design naturally provides two scenarios

- **Both images are almost similar**
- **Both images are almost dissimilar**

If both images are almost similar/dissimilar. Both scenarios are not useful for learning a better representation. The significance of such examples is almost negligible, because they have almost no value to add in learning. Apart from these extreme features, all other features are somewhat related to each other which are more semantically meaningful. In a larger picture, our model is trying to learn the relation between two images by using inter- and intra pairs. To provide meaningful and semantically related examples it is necessary to eliminate such examples during training.