ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240691

CLCE: An Approach to Refining Cross-Entropy and Contrastive Learning for Optimized Learning Fusion

Zijun Long[a,1], Lipeng Zhuang[a], George Killick[a], Zaiqiao Meng[a], Richard Mccreadie[a] and Gerardo Aragon-Camarasa[a]

[a]University of Glasgow

Abstract. State-of-the-art pre-trained image models predominantly adopt a two-stage approach: initial unsupervised pre-training on large-scale datasets followed by task-specific fine-tuning using Cross-Entropy loss (CE). However, it has been demonstrated that CE can compromise model generalization and stability. While recent works employing contrastive learning address some of these limitations by enhancing the quality of embeddings and producing better decision boundaries, they often overlook the importance of hard negative mining and rely on resource intensive and slow training using large sample batches. To counter these issues, we introduce a novel approach named CLCE, which integrates Label-Aware Contrastive Learning with CE. Our approach not only maintains the strengths of both loss functions but also leverages hard negative mining in a synergistic way to enhance performance. Experimental results demonstrate that CLCE significantly outperforms CE in Top-1 accuracy across twelve benchmarks, achieving gains of up to 3.52% in few-shot learning scenarios and 3.41% in transfer learning settings with the BEiT-3 model. Importantly, our proposed CLCE approach effectively mitigates the dependency of contrastive learning on large batch sizes such as 4096 samples per batch, a limitation that has previously constrained the application of contrastive learning in budget-limited hardware environments.

1 Introduction

Approaches for achieving state-of-the-art performance in image classification tasks often employ models initially pre-trained on auxiliary tasks and then fine-tuned on a task-specific labeled dataset with a Cross-Entropy loss (CE) [9, 56, 17, 28, 29, 31, 63]. However, CE's inherent limitations can impact model performance. Specifically, the measure of KL-divergence between one-hot label vectors and model outputs can cause narrow decision margins in the feature space. This hinders generalization [27, 3] and has been shown to be sensitive to noisy labels [39, 27] or adversarial samples [10, 39]. Various techniques have emerged to address these problems, such as knowledge distillation [19], self-training [61], Mixup [65], CutMix [64], and label smoothing [51]. However, in scenarios such as few-shot learning, these issues with CE have not been fully mitigated. Indeed, while techniques such as extended fine-tuning epochs and specialized optimizers [66, 37] can reduce the impact of CE to some extent, they introduce new challenges, such as extended training time and increased model complexity [66, 37, 36].

Amidst these challenges in context of image classification, contrastive learning has emerged as a promising solution [33], particularly in few-shot learning scenarios such as CIFAR-FS [2] and CUB-200-2011 datasets [55]. The effectiveness of contrastive learning lies in its ability to amplify similarities among positive pairs (intra-class data points) and distinguish negative pairs (inter-class data points). SimCLR [4], for instance, has utilized instance-level comparisons unsupervised. However, this unsupervised approach raises concerns regarding its effectiveness , primarily because it limits the positive pairs to be transformed views of an image and treats all other samples in a mini-batch as negatives, potentially overlooking actual positive pairs. We hypothesis incorporating task-specific label information is thus crucial for accurately identifying all positive pairs, especially given the presence of labels in many downstream datasets.

There is a growing trend of using task labels with contrastive learning to replace the standard use of CE [23]. A critical observation here is that many state-of-the-art methods, both in supervised [23, 15] and unsupervised [4, 48, 11, 57, 59] contrastive learning, overlook the strategic selection of negative samples. They fail to differentiate or prioritize these samples during selection or processing, thereby missing the benefits of leveraging “hard" negative samples, as highlighted in numerous studies [49, 25, 50, 5, 69, 26]. While contrastive learning mitigates the limitations of CE, it simultaneously introduces a challenge: a reliance on large batch sizes—such as 2048 or 4096 samples per batch—for superior performance compared to CE. This requirement is often impractical in budget hardware environments, particularly when using GPUs with less than 24 GB of memory. As a consequence, state-of-the-art methods such as SupCon [23] underperform compared to CE when using more commonly employed batch sizes, such as 64 or 128 samples per batch, which limits their application. Motivated by these successes and gaps in research, we pose the question: *How can the performance of contrastive learning be improved to address the shortcomings of cross-entropy loss, while also mitigating the reliance on large batch sizes?*

Building upon the identified research gaps, we propose CLCE, an innovative approach that combines Label-Aware Contrastive Learning with CE. This approach effectively merges the strengths of both loss functions and integrates hard negative mining. This technique refines the selection of positive and negative samples, thereby enabling CLCE to achieve state-of-the-art performance. As our empirical findings illustrate in Fig. 1, CLCE places a greater emphasis on hard negative samples that are visually very similar to positive samples, forcing the encoder to learn how to generate more distinct embeddings and better decision boundaries. The core contributions of our

[1] Email: longkukuhi@gmail.com

work can be summarised as follows:

- Introduction of an innovative approach: We introduce CLCE that boosts model performance without necessitating specialized architectures or additional resources. Our work is the first to successfully integrate explicit hard negative mining into Label-Aware Contrastive Learning, retaining the benefits of CE, while also obviating the dependence on large batch sizes.
- State-of-the-Art Performance in Few-Shot Learning and Transfer Learning settings: CLCE significantly surpasses CE by an average of 2.74% in Top-1 accuracy across four few-shot learning datasets when using the BEiT-3 base model [56], with large gains observed in 1-shot learning scenarios. Additionally, in transfer learning settings, CLCE consistently outperforms other state-of-the-art methods across eight image datasets, setting a new state-of-the-art result for base models (88 million parameters) on ImageNet-1k [7].
- Reduced Contrastive Learning's Dependency on Large Batch Sizes: Empirical evidence shows that CLCE significantly outperforms both CE and previous state-of-the-art contrastive learning methods like SupCon [23] in commonly used batch sizes, such as 64. This is a size at which earlier state-of-the-art contrastive learning methods underperform. This advancement tackles a crucial bottleneck in contrastive learning, particularly in settings with limited resources. It positions CLCE as a viable, efficient alternative to conventional CE.

2 Related Work

2.1 *Limitations of Cross-Entropy loss*

The cross-entropy loss (CE) has long been the default setting for many deep neural models due to its ability to optimize classification tasks effectively. However, recent research has revealed several inherent drawbacks [27, 3, 32]. Specifically, models trained with the CE tend to exhibit poor generalization capabilities. This vulnerability stems from the model having narrow decision margins in the feature space, making it more susceptible to errors introduced by noisy labels [39, 27] and adversarial examples [10, 39]. These deficiencies underscore the need for alternatives that offer better robustness and discrimination capabilities.

2.2 *Contrastive Learning and Negative Mining*

The exploration of negative samples, particularly hard negatives, in contrastive learning has emerged as a critical yet relatively underexplored area. While the significance of positive sample identification is well-established [60, 12, 68], recent studies have begun to unravel the intricate role of hard negatives. The potential of hard negative mining in latent spaces has been validated in numerous studies [50, 5, 69, 26, 58, 62, 13, 30, 34**?** , 44]. These studies highlight the pivotal role of hard negatives in enhancing the discriminative capability of embeddings. In the contrastive learning domain, [6] tackled the challenge of discerning true negatives from a vast pool of candidates by approximating the true negative distribution. Later, [47] applied hard negative mining to unsupervised contrastive learning, resulting in a framework where only a single positive pair is utilized in each iteration of the loss calculation. However, these approaches still presents limitations, such as inaccurately identifying positive and negative samples and only using one positive pair, which harms the performance of contrastive learning. H-SCL [22] expand upon the concept of hard negative mining within a supervised framework. Although both our work and H-SCL utilize hard negative sampling, the methodologies for implementing sampling significantly differ between the two. Their approach employs a consistent threshold-based dot product for identifying "hard" samples. However, determining an appropriate threshold remains challenging, as it varies significantly across different datasets and even within individual mini-batches. In contrast, our CLCE method dynamically determines the weighting of each sample, proving to be significantly more effective than H-SCL. Moreover, their methodology does not tackle the dependency on large batch sizes, which is a critical limitation on performance and applicability.

Our work builds on these foundational insights, aiming to synergize the strengths of contrastive learning with CE, particularly by employing hard negative mining guided by label information. CLCE employs a dynamic and adaptive strategy to assign weights to "hard" samples in each minibatch, offering a more refined approach compared to previous studies. Additionally, CLCE achieves superior performance to CE without relying on large batch sizes.

3 APPROACH

In this paper, we propose an enhanced approach named CLCE for image models that integrates our propose Label-Aware Contrastive Learning with the Hard Negative Mining (LACLN) and the Cross-Entropy (CE). CLCE harnesses the potential of contrastive learning to mitigate the limitations inherent in CE while preserving its advantages. Specifically, LACLN enhances similarities between instances of the same class (i.e. positive samples) using label information and contrasts them against instances from other classes (i.e. negative samples), with particular emphasis on hard negative samples. Thus, LACLN reshapes pretrained embeddings into a more distinct and discriminative space, enhancing performance on target tasks. Moreover, CLCE's foundation draws from the premise that the training efficacy of negative samples varies between soft and hard samples. We argue that weighting negative samples based on their dissimilarity to positive samples is more effective than treating them equally. This allows the model to prioritize distinguishing between positive samples and those negative samples that the embedding deems similar to the positive ones, ultimately enhancing overall performance.

3.1 *CLCE*

The overall proposed CLCE approach is a weighted combination of LACLN and standard CE, as expressed in Eq. 1:

$$\mathcal{L}_{\mathrm{CLCE}} = (1-\lambda)\mathcal{L}_{\mathrm{CE}} + \lambda\mathcal{L}_{\mathrm{LACLN}} \tag{1}$$

In Eq. 1, the term $\mathcal{L}_{\mathrm{CE}}$ represents the CE loss, while $\mathcal{L}_{\mathrm{LACLN}}$ symbolizes our proposed LACLN loss. λ represents a scalar weighting hyperparameter. λ determines the relative importance of each of the two losses. To provide context for $\mathcal{L}_{\mathrm{CE}}$, we refer to the standard definition of the multi-class CE loss, detailed in Eq. 2:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} z_{i,c}\log(\hat{z}_{i,c}) \tag{2}$$

In Eq. 2, $z_{i,c}$ and $\hat{z}_{i,c}$ represent the label and the model's output probability for the ith instance belonging to class c, respectively.

We present the formal definition of our LACLN in Eq. 3. This loss introduces a weighting factor for each negative sample, calculated based on the dot product (indicating similarity) between the sample embeddings and the anchor, and normalized by a temperature

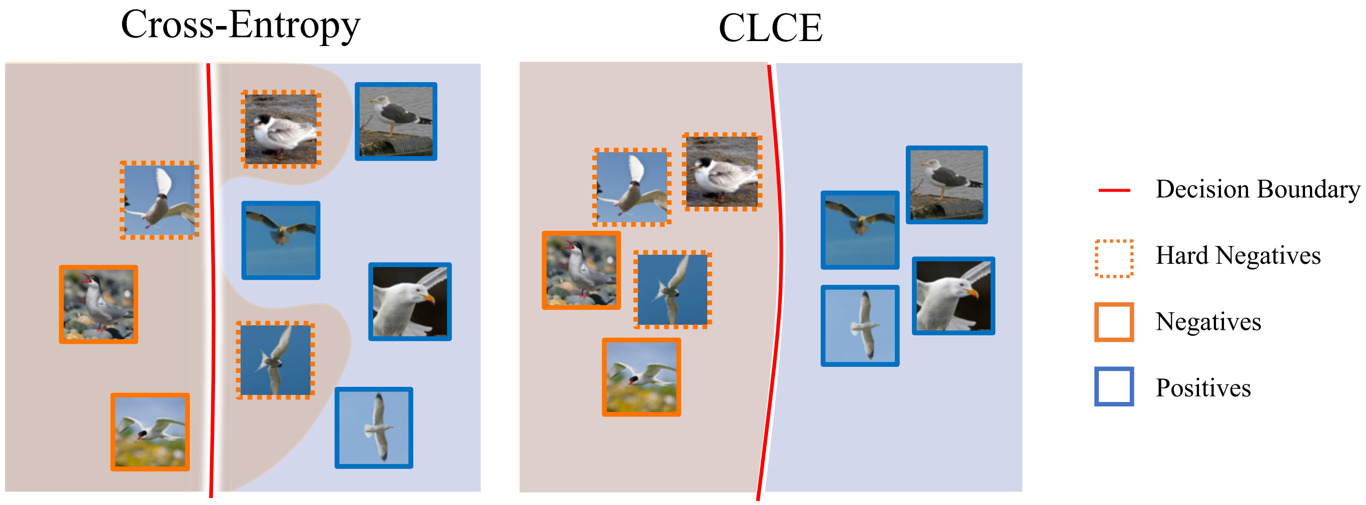


Figure 1. CLCE, our proposed approach, integrates a Label-Aware Contrastive Learning with the Hard Negative Mining (LACLN) term and a CE term. Illustrated with CUB-200-2011 dataset, it emphasizes hard negatives (thick dashed borders) for better class separation. This underscores their marked visual similarity to their positive counterparts. Blue indicates positive examples and orange denotes negatives. On the right, CLCE visibly separates class embeddings more effectively and results a better decision boundary than traditional CE.

$$\mathcal{L}_{\text{LACLN}} = \sum_{x_i \in \mathcal{D}^*} \log \frac{-1}{|\mathcal{D}^{*+}_{-x_i}|} \frac{\displaystyle\sum_{x_p \in \mathcal{D}^{*+}_{-x_i}} \exp(x_i \cdot x_p/\tau)}{\displaystyle\sum_{x_p \in \mathcal{D}^{*+}_{-x_i}} \exp(x_i \cdot x_p/\tau) + \sum_{x_k \in \mathcal{D}^{*-}_{-x_i}} \frac{|\mathcal{D}^{*-}_{-x_i}|}{\displaystyle\sum_{x_k \in \mathcal{D}^{*-}_{-x_i}} \exp(x_i \cdot x_k/\tau)} \exp(x_i \cdot x_k/\tau)^2} \tag{3}$$

parameter τ. This formulation strategically emphasizes “hard” negative samples — those closely associated with the positive samples by the model’s current embeddings. Specifically, the weighting factor for negative samples is determined by calculating their relative proportion based on the average similarity (dot product) observed within each mini-batch. The essence of Eq. 3 is to minimize the distance between positive pair embeddings and maximize the separation between the anchor and negative samples, particularly the hard negatives. This objective is achieved through two components: the numerator, focusing on bringing positive sample embeddings closer to the anchor, and the denominator, containing both positive and weighted negative samples to ensure the anchor’s embedding is distant from negative samples, with a special focus on the more challenging ones. The integration of hard negative mining into contrastive learning is critical as it sharpens the model’s ability to differentiate between closely related samples, thus enhancing feature extraction and overall model performance.

Specifically, $\mathcal{D}^*$ represents the entire mini-batch composed of an embedding x for each image view (or anchor) i. Therefore, $x_i \in \mathcal{D}^*$ is a set of embeddings within the mini-batch. The superscripts $+$ and $-$, e.g. $\mathcal{D}^{*+}$, denote sets of embeddings consisting only of positive and negative examples, respectively, for the current anchor within the mini-batch. The term $|\mathcal{D}^{*+}_{-x_i}|$ represents the cardinality of the positive set for the current anchor, while the subscript $-x_i$ denotes that this set excludes the embedding x_i. The symbol $\cdot$ represents the dot product. τ is a scalar temperature parameter controlling class separation. A lower value for τ encourages the model to differentiate positive and negative instances more distinctly.

3.2 Analysis of CLCE

Notably, our proposed CLCE has the following desirable properties:

- Robust Positive/Negative Differentiation: We ensure a clear distinction between true positive and true negative samples by leveraging explicit label information, as encapsulated in Eq. 3. This not only prevents the model from being misled by incorrectly contrasting of samples but also reinforces the core philosophy of contrastive learning. The aim is two-fold: to reduce the distance between the embeddings of positive pairs and to increase the distance for negative pairs, ensuring robust class separation.
- Discriminating Fine Detail with Hard Negatives: Our loss adjusts the weighting of negative samples based on their similarities to positive instances, as defined in Eq. 3. This nuanced approach ensures that the model not only differentiates between glaringly distinct samples but also adeptly distinguishes more challenging, closely related negative samples. Such an approach paves the way for a robust model that discerns real-world scenarios where differences between classes might be minimal.

3.3 Representation Learning Framework

We use a representation learning framework comprised of three main components, designed specifically to optimize our CLCE approach:

- **Data Augmentation module,** $Aug(\cdot)$: This component creates two different views of each sample r, denoted $\tilde{r} = Aug(r)$. This means that every sample will have at least one similar sample (positive pair) in a batch during training.
- **Encoder Network,** $Enc(\cdot)$: This network encodes the input data, r, into a representation vector, $x = Enc(r)$. Each of the two different views of the data is fed into the encoder separately.
- **Classification head,** $Head(\cdot)$: This maps the representation vector, x, to probabilities of classes in the target task. The mapping primarily consists of a linear layer, and we utilize its output to calculate the cross-entropy loss.

Our CLCE approach (Eq. 3) can be applied using a wide range of encoders, such as BEiT-3 [56] or the ResNets [16] for image classification. Following the method in [4], every image in a batch is altered to produce two separate views (anchors). Views with the same label as the anchor are considered positive, while the rest are viewed as negative. The encoder output, represented by $x_i = Enc(r_i)$, is used to calculate the contrastive loss. In contrast, the output from the classification head, denoted as $z_i = Head(Enc(r_i))$, is used for the CE. We have incorporated L2 normalization on encoder outputs, a strategy demonstrated to enhance performance significantly [52].

4 Evaluation

We evaluate our proposed approach, CLCE, on image classification in two settings: few-shot learning and transfer learning. We also conduct several analytical experiments. For CLCE experiments, a grid-based hyperparameter search is conducted on the validation set. Optimal settings ($\tau = 0.5$ and $\lambda = 0.9$) are employed because they consistently yield the highest validation accuracies. For all experiments, we use the official train/test splits and report the mean Top-1 test accuracy across at least three distinct initializations.

We employ representative models from two categories of architectures – BEiT-3/MAE/ViT base [56, 17, 9] (transformers based models), and ResNet-101 [16] (convolutional neural network). While new state-of-the-art models are continuously emerging (e.g. DINOv2 [41]), our focus is not on the specific choice of architecture. Instead, we aim to show that CLCE is model-agnostic by demonstrating performance gains with two very different and widely used architectures, as well as show it can be trained and deployed in hardware-constrained settings. Further implementation details and the complete code for all experiments are publicly available at https://github.com/longkukuhi/CLCE.

4.1 Few-shot Learning

We evaluate our proposed CLCE in the few-shot learning setting. The experiments on few-shot learning aim to assess the quality of the learned representations. Specifically, each test run comprises 3,000 randomly sampled tasks, and we report median Top-1 accuracy with a 95% confidence interval across three runs, maintaining a consistent query shot count of 15. Four prominent benchmarks are used for evaluation: CIFAR-FS [2], FC100 [42], miniImageNet [54], and tieredImageNet [46]. We follow established splitting protocols for a fair comparison [2, 42, 45].

Tab. 1 shows the performance of BEiT-3 and ResNet-101 models under various methods, including CE, H-SCL [22], and the same weighted combination of CE and state-of-the-art supervised contrastive learning loss (SupCon) [23] as CLCE. The results reveal that our CLCE approach consistently improves classification accuracy over other methods, demonstrating superior generalization with limited training data for each class. Our CLCE enhances models' performance on few-shot datasets, significantly outperforming both CE and CE+SuperCon (paired t-test, p < 0.01). In the 1-shot learning context when compared to BEiT-3 trained with CE (BEiT-3-CE), the most remarkable improvement is seen on the FC100 dataset, with accuracy rising by 3.52% through the use of CLCE (BEiT-3-CLCE). Indeed, across all datasets, BEiT-3-CLCE shows an average accuracy improvement of 2.7%. For 5-shot learning, the average improvements across the datasets are 1.4% in accuracy for BEiT-3-CLCE, demonstrating CLCE's effectiveness in scenarios with fewer positive samples per class and its ability to yield consistent and reliable results, evident in the tighter confidence intervals for Top-1 accuracy. As for ResNet-101, CLCE (ResNet-101-CLCE) demonstrates even more significant improvements over both CE and CE+SupCon. The enhancement is especially remarkable in the case of tieredImagenet, where ResNet-101-CLCE achieves increases of 16.68% over ResNet-101-CE and 14.17% over ResNet-101-CE+SupCon in 1-shot learning. For 5-shot learning, the improvements are 16.9% and 13.36%, respectively. On average, ResNet-101-CLCE achieves a 9.82% improvement in 1-shot and an 8.71% improvement in 5-shot settings over the ResNet-101-CE. Lastly, H-SCL [22] underperforms compared to CE at a batch size of 128. This highlights contrastive learning's limitation of needing very large batch sizes for better performance than CE, evident in ResNet-101 and BEiT-3 models.

Overall, the enhancement of our CLCE is particularly effective for few-shot scenarios, where limited labelled data requires the model to rely more on high-quality, discriminative representations. These outcomes underline the efficacy of our proposed CLCE approach and CLCE's broad applicability across different model architectures for few-shot learning tasks.

4.2 Transfer Learning

We now assess the transfer learning performance of our proposed CLCE. Here, adhering to the widely accepted paradigm for achieving state-of-the-art results, models are initialized with publicly-available weights from pretraining on ImageNet-21k [7] since they are state-of-the-art, and are fine-tuned on smaller datasets using our new loss function. We leverage 8 datasets: CIFAR-100 [24], CUB-200-2011 [55], Caltech-256 [14], Oxford 102 Flowers [40], Oxford-IIIT Pets [43], iNaturalist 2017 [20], Places365 [70], and ImageNet-1k [7]. We adhere to official train/test splits and report mean Top-1 test accuracy over three different initializations.

Tab. 2 presents the results of transfer learning, which offers further evidence of the effectiveness of our proposed CLCE approach beyond few-shot scenarios. When applied to four state-of-the-art image models, including BEiT-3, ResNet-101, ViT-B and MAE, our proposed CLCE approach consistently surpasses other methods, including the standard CE, H-SCL [22] and the same weighted combination of CE and SupCon loss as CLCE. A paired t-test confirms these improvements as statistically significant ($p < 0.05$). While the increase in performance with BEiT-3-CLCE over the BEiT-3-CE baseline is modest in some cases, such as the rise from 98.00% (BEiT-3-CE) to 98.93% (BEiT-3-CLCE) on CUB-200, it shows significant enhancements in challenging datasets with a higher level of

		CIFAR-FS		FC100		miniImageNet		tieredImageNet	
Model	Loss	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
[8]	Transductive	76.58±0.68	85.79±0.50	43.16±0.59	57.57±0.55	65.73±0.68	78.40±0.52	73.34±0.71	85.50±0.50
[67]	Meta-QDA	75.83±0.88	88.79±0.75	-	-	67.83±0.64	84.28±0.69	74.33±0.65	89.56±0.79
[18]	FewTRUE-ViT	76.10±0.88	86.14±0.64	46.20±0.79	63.14±0.73	68.02±0.88	84.51±0.53	72.96±0.92	87.79±0.67
[18]	FewTRUE-Swin	77.76±0.81	88.90±0.59	47.68±0.78	63.81±0.75	72.40±0.78	86.38±0.49	76.32±0.87	89.96±0.55
[21]	BAVARDAGE	82.68±0.25	89.97±0.18	52.60±0.32	65.35±0.25	77.85±0.28	88.02±0.14	79.38±0.29	88.04±0.18
ResNet-101	CE	69.80±0.84	85.20±0.62	43.71 ±0.73	58.65±0.74	55.73±0.85	73.86±0.65	46.93±0.85	62.93±0.76
ResNet-101	H-SCL [22]	67.25±0.86	84.51±0.65	41.34±0.72	57.02±0.70	53.38±0.79	70.29±0.63	44.43±0.82	60.83±0.71
ResNet-101	CE+SupCon	73.61±0.80	86.15±0.53	45.30±0.62	60.18±0.72	57.49±0.82	75.63±0.61	49.44±0.79	66.47±0.60
ResNet-101	CLCE (this work)	76.14±0.75	87.93±0.48	49.48±0.57	64.31±0.70	66.20±0.74	83.41±0.55	63.61±0.72	79.83±0.51
BEiT-3	CE	83.68±0.80	93.01±0.38	66.35±0.95	84.33±0.54	90.62±0.60	95.77±0.28	84.84±0.70	94.81±0.34
BEiT-3	H-SCL [22]	82.21±0.80	91.49±0.37	65.27±0.98	82.61±0.52	88.57±0.62	93.03±0.29	81.37±0.73	93.26±0.33
BEiT-3	CE+SupCon	84.93±0.74	93.36±0.34	67.58±0.86	86.10±0.57	91.04±0.55	95.97±0.24	85.72±0.64	95.33±0.29
BEiT-3	CLCE (this work)	**87.00±0.70**	**93.77±0.36**	**69.87±0.91**	**87.06±0.52**	**92.35±0.53**	**96.78±0.23**	**87.24±0.62**	**96.09±0.29**

Table 1. Comparison to baselines on the few-shot learning setting. Average few-shot classification accuracies (%) with 95% confidence intervals on test splits of four few-shot learning datasets.

Model	Loss	CIFAR-100	CUB-200	Caltech-256	Oxford-Flowers	Oxford-Pets	iNat2017	Places365	ImageNet-1k
ResNet-101	CE	96.27	84.62	81.38	95.71	93.24	66.11	54.73	78.70
ResNet-101	H-SCL [22]	92.78	77.14	78.64	92.34	92.58	63.14	52.02	77.10
ResNet-101	CE+SupCon	96.31	84.70	81.61	95.73	93.49	66.90	55.41	79.03
ResNet-101	CLCE (this work)	**96.92**	87.48	85.05	**96.33**	94.21	67.93	57.30	80.16
ViT-B	CE	87.13	76.93	90.92	90.86	93.81	65.26	54.06	77.91
ViT-B	CLCE (this work)	88.53	78.21	92.10	92.04	94.01	71.25	58.70	83.94
MAE	CE	87.67	78.46	91.82	91.67	94.05	70.50	57.90	83.60
MAE	CLCE (this work)	90.29	81.30	93.11	92.82	94.88	71.62	58.40	84.02
BEiT-3	CE	92.96	98.00	98.53	94.94	94.49	72.31	59.81	85.40
BEiT-3	H-SCL [22]	89.50	95.70	96.24	92.60	93.28	68.51	56.66	82.25
BEiT-3	CE+SupCon	92.74	98.06	98.65	94.92	94.77	73.58	60.52	85.70
BEiT-3	CLCE (this work)	93.56	**98.93**	**99.41**	95.43	**95.62**	**75.72**	**62.22**	**86.14**

Table 2. Comparison to baselines on transfer learning setting. The results are Top-1 classification accuracies across eight diverse datasets.

class diversity. A notable example is iNaturalist2017, which has 5089 different classes, where CLCE leads to a marked improvement in accuracy from 72.31% to 75.72%. This substantial increase suggests that CLCE's benefits are more pronounced in more varied datasets. In the case of ImageNet-1k, accuracy increased from 85.40% (BEiT-3-CE) to 86.14% (BEiT-3-CLCE), setting a new state-of-the-art for base models (88 million parameters) [2]. We observe similar improvements in other transformer-based models, such as ViT and MAE. The use of CLCE in fine-tuning ResNet-101 also resulted in significant performance gains, particularly in the Caltech-256 dataset. Here, the model's accuracy increases from 81.38% (ResNet-101-CE) to 85.05% (ResNet-101-CLCE). Compared to ResNet-101-CE, there has been an average increase in accuracy of 1.83% for ResNet-101-CLCE. Furthermore, H-SCL [22] yields inferior results compared to CE, mirroring the result observed in few-shot scenarios. Overall, the consistent achievement of high accuracies across diverse datasets using models fine-tuned with CLCE, especially ResNet-101 and BEiT-3, underscores the effectiveness of CLCE in improving model performance. Remarkably, this is achieved without resorting to specialized architectures, extra data, or heightened computational requirements, thereby establishing CLCE as a powerful alternative to traditional CE.

4.3 Reducing Batch Size Dependency

We evaluate the effect of batch size on the performance, specifically comparing our CLCE approach with CE and SupCon [23]. The results, as detailed in Tab. 3, indicate that SupCon's performance is sensitive to batch size variations, a limitation not observed with CE. Particularly, SupCon shows inferior performance compared to CE with the commonly used batch size of 64 on both tested datasets. Even when the batch size is increased to 128, SupCon continues to underperform relative to CE. In our experiments, SupCon generally needs a batch size exceeding 512 to outperform CE, a requirement that is impractical for most single-GPU setups. This scenario mirrors the results of H-SCL [22] in the context of few-shot and transfer learning. In contrast, CLCE not only surpasses CE performance on the iNat2017 dataset with a 1.41% accuracy improvement with batch size of 64 but also demonstrates an even more performance gain of 3.52% in accuracy with batch size of 128. Thus, our CLCE approach significantly mitigates the dependency on large batch sizes typically associated with contrastive learning approaches like SupCon and H-SCL. The reduction in dependency on large batch sizes greatly enhances the adaptability and effectiveness of CLCE in diverse computational settings, such as environments with budget GPUs equipped with 12 GB of memory.

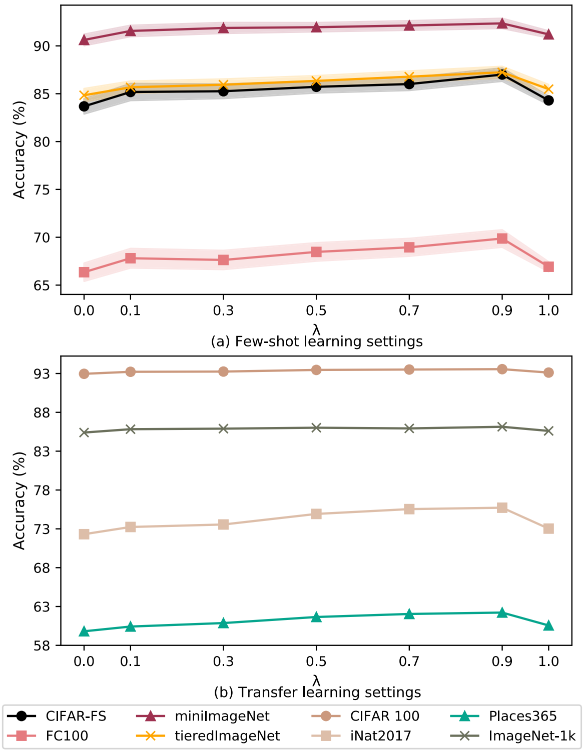
Moreover, gradient accumulation is commonly used in cross-entropy loss to achieve a similar effect when requiring large batch sizes. However, gradient accumulation is very challenging in contrastive learning due to the need to ensure that the accumulated gradients accurately reflect the contrastive nature of the task, particularly in maintaining the integrity of positive and negative pair distributions. This also increases the complexity of maintaining effective sampling strategies which could vary among datasets, in pairs or triplets across accumulation steps. Thus, gradient accumulation is an inadequate method for overcoming the dependency on large batch sizes. CLCE, on the other hand, offers a more efficient

[2] https://paperswithcode.com/sota/image-classification-on-imagenet

Loss	Batch Size	CIFAR-FS	iNat2017
CE	64	83.68	72.31
CE	128	83.39	72.20
SupCon [23]	64	80.31	69.05
SupCon [23]	128	82.17	69.93
CLCE (this work)	64	84.59	73.72
CLCE (this work)	128	87.00	75.72

Table 3. Impact of different batch size. Performance of BEiT-3 base model when trained on CIFAR-FS and iNat2017 datasets. “CE" denotes cross-entropy loss. “SupCon" denotes supervised contrastive learning loss. “CLCE" denotes our proposed joint loss.

and effective solution.



(a) Few-shot learning settings

(b) Transfer learning settings

Figure 2. Evaluation of the impact of the λ hyperparameter. Results on eight tested datasets with λ values ranging from $\{0, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$. The numerical details for these figures are provided in the supplementary material [35].

4.4 Optimizing λ: Bridging CE and LACLN

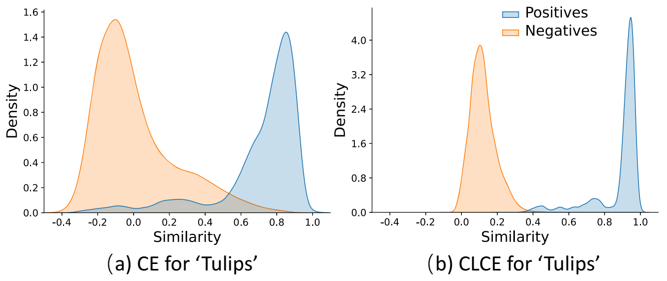
Our proposed CLCE incorporates a hyperparameter, λ, to control the contributions of the CE term and the proposed LACLN term, as shown in Eq. 1. To understand the influence of λ, we evaluate its effect on classification accuracy in few-shot learning and transfer learning. Fig. 2 presents the test accuracy for varying values of λ. Our experiments reveal a consistent trend: as the weight assigned to the LACLN term (λ) increases, performance progressively improves across all tested datasets, peaking at $\lambda = 0.9$. For instance, this optimal setting yields an average performance boost of 2.14% and 2.74% over the exclusive use of either the LACLN or CE term on four few-shot datasets. This trend also manifests in transfer learning settings, highlighting the complementary nature of CE and LACLN. Thus, optimizing this balance is crucial for maximizing performance with CLCE.

CE	CL	HNM	CIFAR-FS	iNat2017
✓			83.68	72.31
✓	✓		84.85	73.53
✓	✓	✓	87.00	75.72

Table 4. Results on CIFAR-FS and iNat2017 when training BEiT-3 base model using ablated versions of our CLCE. “CE" denotes cross-entropy loss. “CL" refers to our proposed label-aware contrastive learning, and “HNM" refers to hard negative mining.

4.5 Ablation Study

We conducted an ablation study on the CIFAR-FS and iNat2017 datasets to evaluate the contributions of two key components in our proposed loss: the proposed label-aware contrastive learning loss without hard negative mining (CL), and the proposed hard negative mining strategy (HNM), as presented in Tab. 4. Across both tested datasets, integrating CL with CE is essential for achieving better performance than the CE—e.g. on the CIFAR-FS dataset, there is a notable performance increase of 1.17%. Meanwhile, the integration of our proposed HNM is critical for CLCE’s enhanced performance, representing one of the main contributions of this paper. For example, it yields a gain of 2.19% accuracy on the iNat2017 dataset compared to the variant of CLCE without HNM. Hence, we conclude that both components are important and complementary.



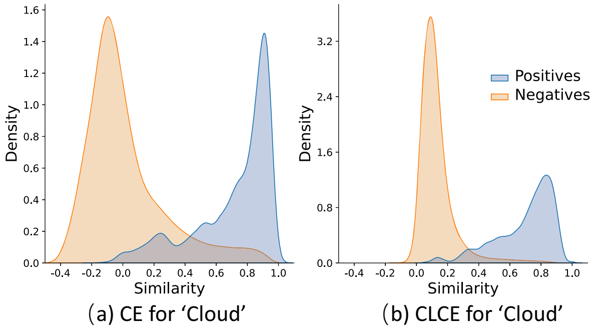
(a) CE for ‘Tulips’

(b) CLCE for ‘Tulips’

Figure 3. Plot of cosine similarity distribution across the “tulips” class from CIFAR-100. Blue represents similarities of positive samples, while orange represents similarities of negative samples.

4.6 Embedding Quality Analysis

To validate the enhancements brought by the proposed approach, CLCE, we perform a thorough evaluation focusing on the geometric characteristics of the generated representation spaces. We hypothesize that our CLCE enhances the quality of embeddings, thereby sharpening class distinction and improving performance. To elaborate, we examine the CE embeddings and CLCE embeddings produced by the BEiT-3 base model. Specifically, we evaluate two key aspects: (1) Distributions of cosine similarities between image pairs. This assessment provides insights into how well the model differentiates between classes in the embedding space. (2) Visualization of



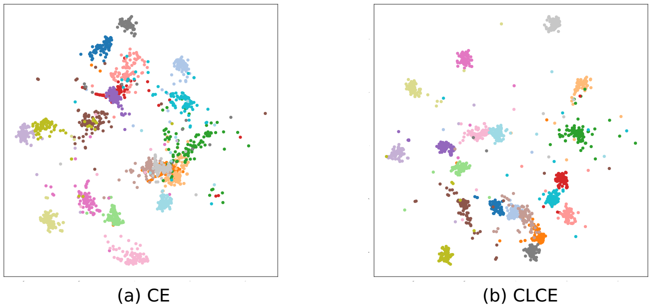
(a) CE for 'Cloud'

(b) CLCE for 'Cloud'

Figure 4. Plot of cosine similarity distribution across the "cloud" class from CIFAR-100. Blue represents similarities of positive samples, while orange represents similarities of negative samples.

the embedding space using the t-SNE algorithm [53]. This visualization allows us to observe the separation or clustering of data points belonging to different classes. (3) We employ the Isotropy Score as defined by [38] to evaluate the quality of produced embeddings. The Isotropy Score measures the distribution of data in the embedding space and serves as a metric for the quality of the produced embeddings. Historically, isotropy has served as an evaluation metric for representation quality [1]. This is based on the premise that widely distributed representations across different classes in the embedding space facilitate better distinction between them.

We present the pairwise cosine similarity distributions of CE and CLCE embeddings in Figs. 3 and 4. Specifically, we randomly select the "tulips" and "cloud" classes from CIFAR-100 to compute cosine similarities for positive (same class) and negative pairs (different classes). Observations from these plots reveal that the CLCE embeddings demonstrate superior separation between classes and less overlap between positive and negative samples compared to CE.



(a) CE

(b) CLCE

Figure 5. Embedding Space Visualization for CE vs. CLCE, over twenty CIFAR-100 test set classes using t-SNE. Each dot represents a sample, with distinct colors indicating different label classes.

In Fig. 5, the t-SNE visualization of the embedding space for CE and CLCE across twenty CIFAR-100 classes. The CE embeddings (Fig. 5a) display instances where the same class nodes are relatively closely packed but also reveal many outliers. This suggests a reduced discriminative capability. On the contrary, CLCE embeddings (Fig. 5b) display more separated and compact class clusters, suggesting improved discriminative capabilities.

Formally, we calculate the quantitative Isotropy Score (IS) [38], which is defined as follows:

$$IS(\mathcal{V}) = \frac{max_{c \subset C} \sum_{v \subset V} \exp{(C^T V)}}{min_{c \subset C} \sum_{v \subset V} \exp{(C^T V)}} \tag{4}$$

where V is a set of vectors, C is the set of all possible unit vectors (i.e., any c so that $||c|| = 1$) in the embedding space. In practice, C is approximated by the eigenvector set of $V^T V$ (V are the stacked embeddings of v). The larger the IS value, the more isotropic an embedding space is (i.e., a perfectly isotropic space obtains an IS score of 1).

Model	iNaturalist2017	Imagenet-1k	Places365
BEiT3-CE	0.32	0.27	0.34
BEiT3-CLCE	0.98	0.92	0.93

Table 5. Comparison of Isotropy Score across three datasets for BEiT-3-CE and BEiT-3-CLCE. A higher value is better. A higher Isotropy Score indicates better isotropy and generalizability.

Tab. 5 demonstrates that the IS score for BEiT-3-CLCE is superior to that of BEiT-3-CE, confirming that CLCE produces a more isotropic semantic space. The BEiT-3-CE embeddings are more anisotropic, implying that BEiT-3-CLCE embeddings more distinctly separate the different classes.

These observations indicate that the proposed CLCE approach restructures the embedding space to enhance class distinction, addressing the generalization limitation of the CE. This enhancement is particularly effective for few-shot scenarios, where limited labelled data requires the model to rely more on high-quality, discriminative representations.

5 Discussion and Conclusion

Limitations. While our CLCE approach advances the state-of-the-art, it still has certain limitations. Firstly, CLCE shows increased performance with larger batch sizes. As Table 3 illustrates, CLCE surpasses CE in accuracy in few-shot and transfer learning scenarios at a batch size of 64, with further improvements observed at larger batch sizes. Secondly, our approach applies hard negative mining solely to the contrastive learning component and not to the CE component. This is due to differing implementations of hard negative mining in each loss. In cross-entropy, hard negatives are identified based on loss values, necessitating a unique strategy that might interfere with the existing sampling process in contrastive learning and potentially cause conflicting outcomes. Additionally, the divergent goals of cross-entropy and contrastive learning, where the former focuses on minimizing the discrepancy between predicted and true distributions and the latter emphasizes embedding similarities, complicate the use of a unified hard negative mining approach.

Conclusion. In this work, we proposed a approach for training image models, denoted CLCE. CLCE combines label-aware contrastive learning with hard negative mining and CE, to address the shortcomings of CE and existing contrastive learning methods. Our empirical results demonstrate that CLCE consistently outperforms traditional CE and prior contrastive learning approaches, both in few-shot learning and transfer learning settings. Furthermore, CLCE offers an effective solution for researchers and developers who can only access commodity GPU hardware, as CLCE maintains its effectiveness when working with smaller batch sizes that can be loaded onto cheaper GPU cards with less on-board memory. To summarize, our comprehensive investigations and robust empirical evidence compellingly substantiate our methodological decisions, underscoring that CLCE serves as a superior alternative to CE for augmenting the performance of image models for image classification.

References

[1] S. Arora, Y. Li, Y. Liang, et al. A latent variable model approach to pmi-based word embeddings. *Trans. Assoc. Comput. Linguistics*, 2016.

[2] L. Bertinetto, J. F. Henriques, et al. Meta-learning with differentiable closed-form solvers. In *Proc. ICLR 2019*.

[3] K. Cao, C. Wei, et al. Learning imbalanced datasets with label-distribution-aware margin loss. *CoRR*, abs/1906.07413, 2019.

[4] T. Chen, S. Kornblith, et al. A simple framework for contrastive learning of visual representations. In *Proc.ICML*, Proceedings of Machine Learning Research, 2020.

[5] G. Chu, X. Wang, et al. Cuco: Graph representation with curriculum contrastive learning. In *Proc. IJCAI*.

[6] C. Chuang, J. Robinson, et al. Debiased contrastive learning. *CoRR*, abs/2007.00224, 2020. URL https://arxiv.org/abs/2007.00224.

[7] J. Deng, W. Dong, et al. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR, 2009*.

[8] G. S. Dhillon et al. A baseline for few-shot image classification. In *Proc. ICLR 2020*.

[9] A. Dosovitskiy, L. Beyer, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. ICLR, 2021*.

[10] G. F. Elsayed et al. Large margin deep networks for classification. In *Proc. NeurIPS*, 2018.

[11] W. Ge, W. Huang, D. Dong, and M. R. Scott. Deep metric learning with hierarchical triplet loss. *CoRR*, abs/1810.06951, 2018. URL http://arxiv.org/abs/1810.06951.

[12] X. Ge et al. Structured multi-modal feature embedding and alignment for image-sentence retrieval. In *Proceedings of the 29th ACM international conference on multimedia*, pages 5185–5193, 2021.

[13] X. Ge et al. Algrnet: Multi-relational adaptive facial action unit modelling for face representation and relevant recognitions. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2023.

[14] G. Griffin et al. *Caltech-256 Object Category Dataset*. Mar 2007.

[15] B. Gunel, J. Du, A. Conneau, et al. Supervised contrastive learning for pre-trained language model fine-tuning. *CoRR*, abs/2011.01403, 2020.

[16] K. He, X. Zhang, et al. Deep residual learning for image recognition. In *Proc. CVPR*, 2016.

[17] K. He, X. Chen, et al. Masked autoencoders are scalable vision learners. In *Proc. CVPR*, 2022.

[18] M. Hiller, R. Ma, et al. Rethinking generalization in few-shot classification. In *NeurIPS*, 2022.

[19] G. E. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

[20] G. V. Horn, O. M. Aodha, et al. The inaturalist species classification and detection dataset. In *Proc. CVPR*, 2018.

[21] Y. Hu, S. Pateux, et al. Adaptive dimension reduction and variational inference for transductive few-shot classification. In *International Conference on Artificial Intelligence and Statistics*, 2023.

[22] R. Jiang, T. Nguyen, et al. Supervised contrastive learning with hard negative samples. *CoRR*, abs/2209.00078, 2022.

[23] P. Khosla, P. Teterwak, et al. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 2020.

[24] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[25] B. G. V. Kumar, B. Harwood, G. Carneiro, I. D. Reid, and T. Drummond. Smart mining for deep metric learning. *CoRR*, 2017.

[26] D. Li, Z. Wang, et al. Self-guided hard negative generation for unsupervised person re-identification. In *Proc. IJCAI*, 2022.

[27] W. Liu, Y. Wen, Z. Yu, et al. Large-margin softmax loss for convolutional neural networks. In *Proc. ICML*, 2016.

[28] Z. Long and R. McCreadie. Automated crisis content categorization for covid-19 tweet streams. In *Proc. ISCRAM*, 2021.

[29] Z. Long and R. McCreadie. Is multi-modal data key for crisis content categorization on social media? In *Proc. ISCRAM*, 2022.

[30] Z. Long, G. Killick, et al. When hard negative sampling meets supervised contrastive learning. *arXiv preprint arXiv:2308.14893*, 2023.

[31] Z. Long, R. McCreadie, et al. Crisisvit: A robust vision transformer for crisis image classification. In *Proc. ISCRAM*, 2023.

[32] Z. Long, R. McCreadie, et al. Lacvit: A label-aware contrastive fine-tuning framework for vision transformers. In *Proc. ICASSP*, 2023.

[33] Z. Long, G. Killick, et al. Robollm: Robotic vision tasks grounded on multimodal large language models. In *Proc. ICRA*, 2024.

[34] Z. Long, G. Killick, et al. Elucidating and overcoming the challenges of label noise in supervised contrastive learning. *Proc. ECCV*, 2024.

[35] Z. Long, G. Killick, L. Zhuang, G. Aragon-Camarasa, Z. Meng, and R. Mccreadie. Clce: An approach to refining cross-entropy and contrastive learning for optimized learning fusion. *arXiv:2402.14551*, 2024. Full version of this paper.

[36] Z. Long et al. Multiway-adapater: Adapting large-scale multi-modal models for scalable image-text retrieval. In *Proc. ICASSP, 2024*.

[37] M. Mosbach et al. On the stability of fine-tuning BERT: misconceptions, explanations, and strong baselines. *CoRR*, abs/2006.04884, 2020.

[38] J. Mu and P. Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *Proc. ICLR*, 2018.

[39] K. Nar, O. Ocal, et al. Cross-entropy loss and low-rank features have responsibility for adversarial examples. *CoRR*, abs/1901.08360, 2019.

[40] M. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proc. ICVGIP, 2008*.

[41] M. Oquab, T. Darcet, et al. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023.

[42] B. N. Oreshkin, P. R. López, et al. TADAM: task dependent adaptive metric for improved few-shot learning. In *Proc. NeurIPS*, 2018.

[43] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *Proc. CVPR*, pages 3498–3505. IEEE Computer Society, 2012.

[44] A. V. Petrov et al. gsasrec: reducing overconfidence in sequential recommendation trained with negative sampling. In *Proceedings of the 17th ACM Conference on Recommender Systems*, 2023.

[45] S. Ravi and H. Larochelle. Optimization as a model for few-shot learning. In *Proc.ICLR*, 2017.

[46] M. Ren, E. Triantafillou, S. Ravi, et al. Meta-learning for semi-supervised few-shot classification. In *Proc. ICLR*, 2018.

[47] J. D. Robinson, C. Chuang, et al. Contrastive learning with hard negative samples. In *Proc. ICLR*. OpenReview.net, 2021.

[48] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, 2016.

[49] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. *CoRR*, abs/1511.06452, 2015.

[50] Y. Suh, B. Han, W. Kim, and K. M. Lee. Stochastic class-based hard example mining for deep metric learning. In *Proc. CVPR*, 2019.

[51] C. Szegedy, V. Vanhoucke, S. Ioffe, et al. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[52] Y. Tian, Y. Wang, D. Krishnan, et al. Rethinking few-shot image classification: A good embedding is all you need? In *Proc. ECCV*, 2020.

[53] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[54] O. Vinyals, C. Blundell, T. Lillicrap, et al. Matching networks for one shot learning. In *Proc. NeurIPS*, pages 3630–3638, 2016.

[55] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[56] W. Wang, H. Bao, L. Dong, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *CoRR*, abs/2208.10442, 2022. doi: 10.48550/arXiv.2208.10442.

[57] X. Wang, X. Li, et al. Using self-supervised dual constraint contrastive learning for cross-modal retrieval. In K. Gal, A. Nowé, G. J. Nalepa, R. Fairstein, and R. Radulescu, editors, *Proc. ECAI*, 2023.

[58] C. Wu, F. Wu, and Y. Huang. Rethinking infonce: How many negative samples do you need? In *Proc. IJCAI*, pages 2509–2515. ijcai.org, 2022.

[59] L. Wu and J. Liu. Contrastive learning with diverse samples. In *Proc. ECAI*, 2023.

[60] Q. Xiao et al. Identical and fraternal twins: Fine-grained semantic contrastive learning of sentence representations. In *Proc. ECAI*, 2023.

[61] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan. Billion-scale semi-supervised learning for image classification. *CoRR*, abs/1905.00546, 2019. URL http://arxiv.org/abs/1905.00546.

[62] C. Yang et al. Trading hard negatives and true negatives: A debiased contrastive collaborative filtering approach. In *Proc. IJCAI*, 2022.

[63] Z. Yi, Z. Long, et al. Large multi-modal encoders for recommendation. *arXiv preprint arXiv:2310.20343*, 2023.

[64] S. Yun, D. Han, S. J. Oh, et al. Cutmix: Regularization strategy to train strong classifiers with localizable features. *CoRR*, abs/1905.04899, 2019. URL http://arxiv.org/abs/1905.04899.

[65] H. Zhang, M. Cissé, et al. mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412, 2017. URL http://arxiv.org/abs/1710.09412.

[66] T. Zhang, F. Wu, A. Katiyar, K. Q. Weinberger, and Y. Artzi. Revisiting few-sample BERT fine-tuning. *CoRR*, abs/2006.05987, 2020.

[67] X. Zhang, D. Meng, H. Gouk, et al. Shallow bayesian meta learning for real-world few-shot recognition. In *Proc. ICCV*. IEEE, 2021.

[68] X. o. Zhang. Gla-AI4BioMed at RRG24: Visual instruction-tuned adaptation for radiology report generation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, 2024.

[69] H. Zhao, X. Yang, Z. Wang, et al. Graph debiased contrastive learning with joint representation clustering. In *Proc. IJCAI*, 2021.

[70] B. Zhou, À. Lapedriza, et al. Learning deep features for scene recognition using places database. In *Proc. NeurIPS*, pages 487–495, 2014.