MetaMask: Revisiting Dimensional Confounder for Self-Supervised Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

As a successful approach to self-supervised learning, contrastive learning aims to learn invariant information shared among distortions of the input sample. While contrastive learning has yielded continuous advancements in sampling strategy and architecture design, it still remains two persistent defects: the interference of task-irrelevant information and sample inefficiency, which are related to the recurring existence of trivial constant solutions. From the perspective of dimensional analysis, we find out that the *dimensional redundancy* and *dimensional confounder* are the intrinsic issues behind the phenomena, and provide experimental evidence to support our viewpoint. We further propose a simple yet effective approach *MetaMask*, short for the *dimensional **Mask** learned by **Meta**-learning*, to learn representations against dimensional redundancy and confounder. MetaMask adopts the redundancy-reduction technique to tackle the dimensional redundancy issue and innovatively introduces a dimensional mask to reduce the gradient effects of specific dimensions containing the confounder, which is trained by employing a meta-learning paradigm with the objective of improving the performance of masked representations on a typical self-supervised task. We provide solid theoretical analyses to prove MetaMask can obtain tighter risk bounds for downstream classification compared to typical contrastive methods. Empirically, our method achieves state-of-the-art performance on various benchmarks.

1 Introduction

A fundamental idea behind self-supervised learning is to learn discriminative representations from the input data without relying on human annotations. Recent advances in visual self-supervised learning [1, 2, 3, 4] demonstrate that unsupervised approaches can achieve competitive performance over supervised approaches by introducing sophisticated self-supervised tasks. A representative learning paradigm is contrastive learning [5, 6, 7, 8, 9], which aims to learn invariant information from different views (generated by data augmentations) by performing instance-level contrast, i.e., pulling views of the same sample together while pushing views of different samples away. Various[10, 11, 12] techniques have been explored to address the problem of trivial solutions to self-supervised contrastive learning, e.g., constant representations. However, state-of-the-art methods still suffer from two crucial defects, including the interference of task-irrelevant information and sample inefficiency. The trivial solution problem is related to such defects, e.g., a small number of negative samples or a large proportion of task-irrelevant information may lead the model to learn constant representations. We revisit the current self-supervised learning paradigm from the perspective of dimensional analysis

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

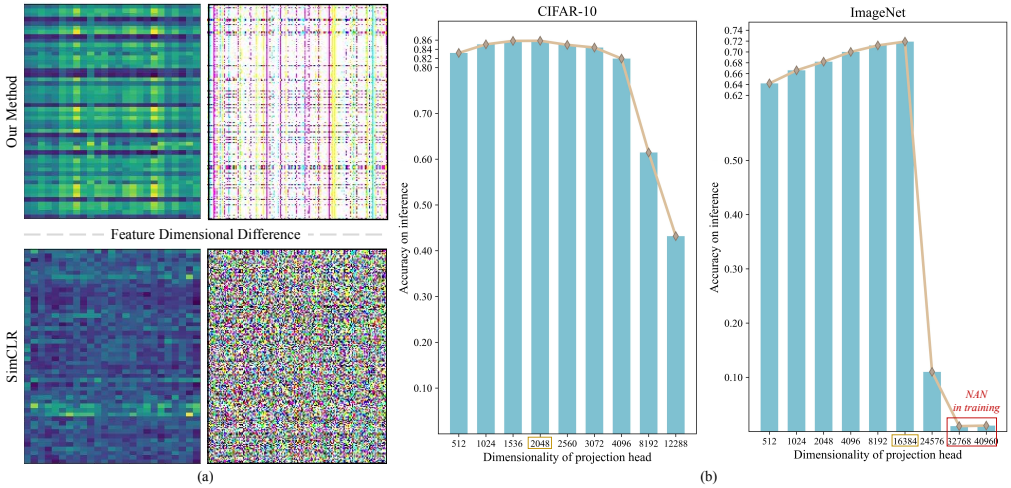


Figure 1: (a) The visualization of the representations learned by SimCLR [6] and our method using the redundancy-reduction technique [14] on CIFAR-10, respectively. The learned features are projected into a color image in RGB format, where different colors represent different types of features. The abscissa axis represents the feature dimensions, and the ordinate axis represents samples of different classes. The greater the color contrast, the lower the dimensional feature similarity. The left plots present the contribution of different dimensions to a specific category classification, and the right plots present the similarity between dimension features within a batch. Compared with SimCLR, our method using redundancy-reduction can indeed learn representations with decoupled dimensions. (b) The experimental results obtained by Barlow Twins on CIFAR-10 and ImageNet datasets. We intuitively increase the dimensionality of the projection head, which leads the dimensions of representations to model more different information. Note that we adopt the official code of Barlow Twins and train on 8 GPUs of NVIDIA Tesla V100.

and argue that the intrinsic issues behind the defects of self-supervised learning are the *dimensional redundancy* and *dimensional confounder* [1].

From the perspective of information theory, each dimension holds a subset of the representation's information entropy. The dimensional redundancy denotes that multiple dimensions hold overlapping information entropy. To tackle this issue, Barlow Twins [14] takes advantage of a neuroscience approach, i.e., redundancy-reduction [15], in self-supervised learning. It makes the cross-correlation matrix of twin embeddings close to the identity matrix, which is conceptually simple yet effective to address the dimensional redundancy issue. Figure 1 (a) supports the superiority of redundancy-reduction in addressing the dimensional redundancy problem. Since Barlow Twins encourages the dimensions of learned representations to model decoupled information, it naturally avoids the collapsed trivial solution that outputs the same vector for all images.

However, only reducing dimensional redundancy still suffers from the problem of dimensional confounder. The dimensional confounder indicates a set of dimensions that solely contain "harmful" information and further degenerates the performance of the model, e.g., the dimensions simply capturing chaotic and worthless background information from the input. To prove the existence of such dimensional confounders, we conduct motivating experiments from two perspectives. It is stated in [14] that Barlow Twins keeps improving as the dimensionality of the projection head increases, since the backbone network acts as a dimensionality bottleneck to constrain the dimensionality of the representation so that the learning paradigm of Barlow Twins can promote the representation capture more task-irrelevant information with the same dimensionality. Counterintuitively, our observation from the motivating experiments is in stark contrast with the conclusion in [14]. As shown in Figure 1 (b), the performance of Barlow twins as a function of the dimensionality presents a convex pattern on either CIFAR-10 or ImageNet. After reaching the peak, the performance gradually decreases as the dimensionality increases until it collapses. We reckon that since the task-relevant semantic information is limited, as the information captured in different dimensions is gradually decoupled,

[1] Throughout this paper, the term *confounder* is used in its idiomatic sense rather than the specific statistical sense in Structural Causal Models [13].

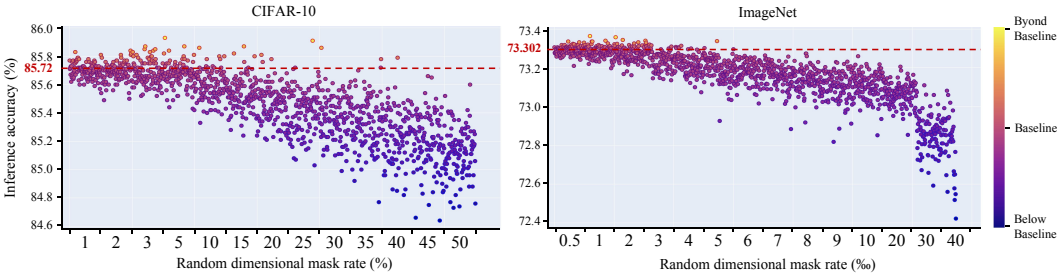


Figure 2: Experimental scatter diagrams obtained by Barlow Twins [14] with randomly masked dimensions on CIFAR-10 and ImageNet datasets, where *Baseline* and the *red dashed lines* denote the performance achieved by the unmasked representation of Barlow Twins. Every single point denotes an independent experimental result achieved by randomly masking the original representation at the dimensional level with a specific mask rate. Note that the original representation is consistently fixed. Following the experimental protocol of [14, 12], we evaluate the performance on CIFAR-10 and ImageNet by adopting KNN prediction and linear probing, respectively, which further proves the general existence of dimensional confounders.

more task-irrelevant noisy information is encoded into the representation, i.e., the dimensional confounder increases, so the performance of the learned representation is getting worse until collapse.

To further validate our inference, we randomly mask several dimensions by frankly setting the masked values to 0, and then we evaluate the performance of the masked representation. The experimental results are demonstrated in Figure 2, where every single point denotes the result of a sampled mask. The representations with some specific masked dimensions can indeed achieve better performance than the unmasked representation, and such an observation is consistent on a small-scaled dataset, e.g., CIFAR-10, and a large-scale dataset, e.g., ImageNet. The dimensional mask rate that can improve the performance of the model on ImageNet is on average lower than that on CIFAR-10. We understand that this phenomenon depends on an ingredient: the representation learned on CIFAR-10 models discriminative information for 10 categories in 512 dimensions, while the representation uses 2048 dimensions to model information for 1000 categories on ImageNet, so the dimension-to-class ratio on CIFAR-10 (51.2) is much higher than that on ImageNet (2.048). Therefore, the presence of dimensional confounders naturally remains at a low level on ImageNet since supporting 1000-category classification requires a large amount of heterogeneous discriminative information. This empirical evidence further supports our assumption of the dimensional confounder.

To this end, we intuitively propose a simple yet effective approach *MetaMask*, short for the ***Meta**-learned dimensional **Mask***, to adjust the impacts of different dimensions in a self-paced manner. In order to learn representations against dimensional redundancy and confounder, we regard the cross-correlation matrix constraint of Barlow Twins as a dimensional regularization term and further introduce a dimensional mask, which simultaneously enhances the gradient effect of the dimensions containing discriminative information and reduces that of the dimensions containing the confounder during training. We employ the meta-learning paradigm to train the learnable dimensional mask with the optimization objective of improving the performance of masked representations on typical self-supervised tasks, e.g., contrastive learning. Consequently, the dimensional mask incessantly adjusts the weights of different dimensions with respect to their gradient contribution to the optimization, so as to make the model focus on the acquisition of discriminative information, which can address the trivial solution caused by treating all dimensions equally in optimization, especially for high-dimensional representations. Empirically, we conduct head-to-head comparisons on various benchmark datasets, which prove the effectiveness of MetaMask. Concretely, the **contributions** of this paper are four-fold:

- We present a timely study on the adaptation of the state-of-the-art self-supervised models on downstream applications and identify two critical issues associated with the learned representation, i.e., dimensional redundancy and dimensional confounder.
- We propose a novel approach to tune the gradient effects of dimensions in a self-paced manner during optimization according to their improvements on a self-supervised task.
- Theoretically, we provide theorems and proofs to support that MetaMask achieves relatively tighter risk bounds for downstream classification compared to typical contrastive methods.
- We conduct extensive experiments to demonstrate the superiority of MetaMask over benchmark methods on downstream tasks.

2 Related Work

Self-supervised learning achieved impressive success in the field of representation learning without human annotation by constructing auxiliary self-supervised tasks. Typically, self-supervised approaches can be categorized as either generative or discriminative [16, 6, 10]. Conventional generative approaches bridge the input data and latent embedding by building a distribution over them, and the learned embeddings are treated as image representations [17, 18, 19, 20, 21, 22]. However, such a learning paradigm learns the distribution at the pixel-level, which is computationally expensive. The detailed information required for image generation may be redundant for representation learning.

For discriminative methods, contrastive learning achieves state-of-the-art performance on various downstream tasks [4, 23, 24, 25, 26, 27, 28, 29, 11]. Deep InfoMax [1] explores to maximize the mutual information between the input and output of an encoder. CPC [3] adopts noise-contrastive estimation [30] as the contrastive loss to train the model by maximizing the mutual information of multiple views, which is deduced by using the Kullback-Leibler divergence [31]. CMC [5] and AMDIM [2] employs contrastive learning on the multi-view data. SimCLR [6] and MoCo [7, 32] use large batch or memory bank to enlarge the amount of available negative features. To sample informative negatives, DebiasedCL[8] and HardCL [9] design sampling strategies by using positive-unlabeled learning methods [33, 34]. Motivated by [35], [36] provides an information theoretical framework for self-supervised learning and proposes an information bottleneck-based approach. However, these methods rely on a large number of negative samples to learn valuable representations while avoiding trivial solutions, i.e., constant representation. SimSiam [12] and BYOL [10] jointly waive negatives and avoid training degradation by using the techniques of asymmetric siamese network and stop-gradient. SSL-HSIC [37] provides a unified framework for negative-free contrastive learning based on Hilbert Schmidt Independence Criterion [38]. The representation learning paradigm employed by current contrastive methods is contrary to the coding criteria proposed by [15] and [39], respectively. Therefore, Barlow Twins [14] proposes to constrain the cross-correlation matrix between the features to be close to the identity matrix, which can effectively reduce the dimensional redundancy. Yet, such an approach still suffers from the dimensional confounder issue, and we thus propose a new meta-learning-based method to attenuate the over-learning of dimensional confounders by learning a mask to filter out irrelevant information.

3 Preliminaries

Formally, we suppose the input multi-view dataset as $\mathcal{X} = \left\{ \boldsymbol{x}_i^j \middle| i \in [\![1, N^*]\!], j \in [\![1, M]\!] \right\}$, where N^* denotes the number of samples, and M denotes the number of views. Therefore, $\boldsymbol{x}_i^j$ represents the j-th view of the i-th sample following $\boldsymbol{x}_i^j \sim \mathcal{X}$. The multiple views are built by performing identically distributed data augmentations [5, 6, 14], i.e., $\{\boldsymbol{t}_j\}_{j=1}^M \sim \mathcal{T}$.

3.1 Contrastive Learning

We recap the preliminaries of contrastive learning [6], which aims to learn an embedding that jointly maximizes agreement between views of the same sample and separates the views of different samples in the latent space. Given a minibatch of N examples sampled from the multi-view dataset $\mathcal{X}$, *positives* represent the pairs $\{\boldsymbol{x}_i^j, \boldsymbol{x}_i^{j'}\}$ where $j, j' \in [\![1, M]\!]$, and *negatives* represent $\{\boldsymbol{x}_i^j, \boldsymbol{x}_{i'}^{j'}\}$ where $i \neq i'$. Conventional contrastive methods feed the input $\boldsymbol{x}_i^j$ into a view-shared encoder $f_\theta(\cdot)$ to learn the representation $\boldsymbol{h}_i^j$, which is mapped into a feature $\boldsymbol{z}_i^j$ by a projection head $g_\vartheta(\cdot)$, where θ and ϑ are the network parameters. $f_\theta(\cdot)$ and $g_\vartheta(\cdot)$ are trained by using the contrastive loss [3, 6]:

$$\mathcal{L}_{contrast} = -\sum_{i \in [\![1,N]\!]} \sum_{\substack{j \in [\![1,M-1]\!] \\ j^+ \in [\![j+1,M]\!]}} \log \frac{\exp\left[d\left(\boldsymbol{z}_i^j, \boldsymbol{z}_i^{j^+}\right)/\tau\right]}{\sum_{i' \in [\![1,N]\!]} \sum_{j' \in [\![1,M]\!]} \mathbb{1}_{[i \neq i' \vee j \neq j']} \cdot \exp\left[d\left(\boldsymbol{z}_i^j, \boldsymbol{z}_{i'}^{j'}\right)/\tau\right]} \tag{1}$$

where $d(\cdot)$ is a discriminating function to measure the similarity of the feature pairs, $\mathbb{1}_{[i \neq i' \vee j \neq j']}$ denotes an indicator function equalling to 1 if $i \neq i'$ or $j \neq j'$, and τ is a temperature parameter

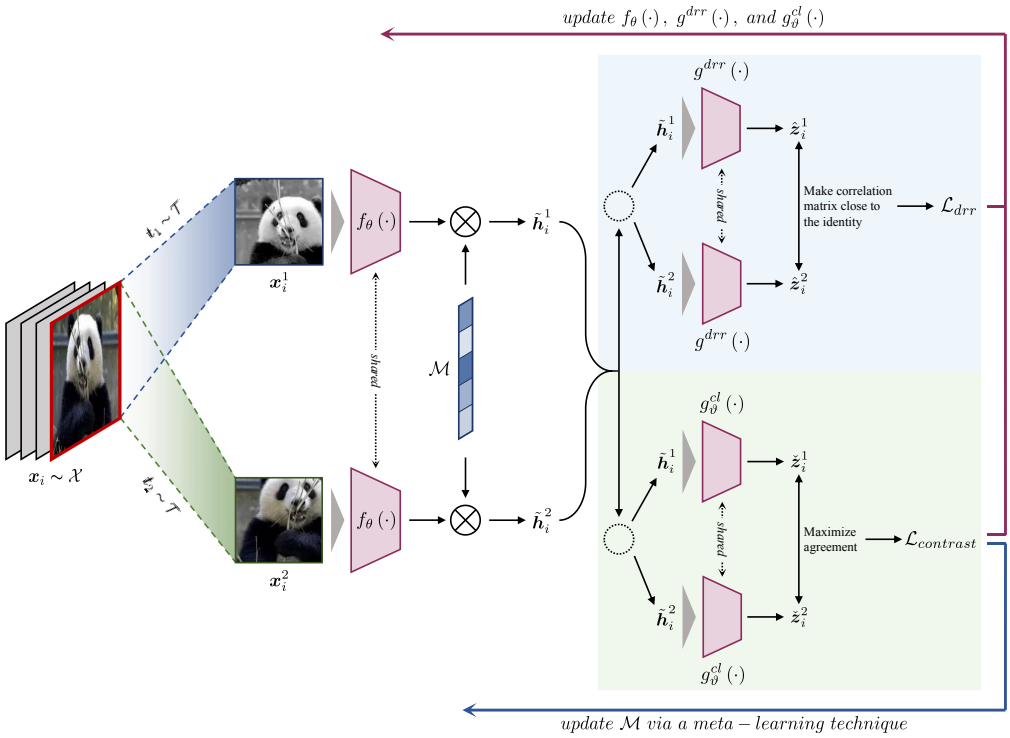


Figure 3: The architecture of MetaMask. The solid red line pointing backwards represents the regular training step, where the encoder and projection heads are trained by jointly back-propagating $\mathcal{L}_{drr}$ and $\mathcal{L}_{contrast}$. The solid blue line pointing backwards represents the meta-learning step, where the dimensional mask $\mathcal{M}$ is trained by performing the second-derivative technique on $\mathcal{L}_{contrast}$. All learnable networks are trained until convergence.

valued by following [6]. In inference, the projection head $g_\vartheta(\cdot)$ is discarded, and the representation $\boldsymbol{h}_i^j$ is directly used for downstream tasks.

3.2 Dimensional Redundant Reduction

Following the principle proposed by [15], [14] introduces a redundancy-reduction objective function:

$$\mathcal{L}_{drr} = \sum_{\substack{j \in [\![1, M-1]\!] \\ j^+ \in [\![j+1, M]\!]}} \left[\sum_{k \in [\![1, D]\!]} \left(1 - \boldsymbol{C}_{kk}^{j,j^+}\right)^2 + \lambda \sum_{\substack{k,k' \in [\![1, D]\!] \\ k \neq k'}} \left(\boldsymbol{C}_{kk'}^{j,j^+}\right)^2 \right] \tag{2}$$

where λ is a positive hyper-parameter valued by following [14], and $\boldsymbol{C}$ is the cross-correlation matrix computed between the outputs of the encoder along the batch dimension:

$$\boldsymbol{C}_{kk'}^{j,j^+} = \frac{\sum_{i \in [\![1, N]\!]} \boldsymbol{z}_{i,k}^{j} \cdot \boldsymbol{z}_{i,k'}^{j^+}}{\sqrt{\sum_{i \in [\![1, N]\!]} \left(\boldsymbol{z}_{i,k}^{j}\right)^2} \cdot \sqrt{\sum_{i \in [\![1, N]\!]} \left(\boldsymbol{z}_{i,k'}^{j^+}\right)^2}} \tag{3}$$

where k and k' index the dimension of features. $\boldsymbol{C}$ is a square matrix with the size of D, which denotes the dimensionality of the projected feature $\boldsymbol{z}$. The value range of the elements in $\boldsymbol{C}$ is $[-1, 1]$, i.e., complete anti-correlation to complete correlation.

4 Methodology

Our goal is to learn discriminative representations against dimensional redundancy and dimensional confounder from multiple views by performing self-supervised learning. To this end, we propose

MetaMask to automatically determine the optimal dimensionality of the representation by decoupling information into different dimensions and filtering out those dimensions with irrelevant information. MetaMask addresses the dimensional redundancy issue by adopting a redundancy-reduction regularization. To tackle the dimensional confounder problem, MetaMask learns a dimensional mask to adjust the gradient weights of dimensions according to their contributions to improving a specific self-supervised task during optimization, which is achieved in a meta-learning manner.

MetaMask's architecture is shown in Figure 3. Specifically, the sample $\boldsymbol{x}_i^j$ is first fed into the encoder to obtain its representation $\boldsymbol{h}_i^j$. Then, we build a learnable dimensional mask $\mathcal{M} = \left\{\boldsymbol{\omega}_k \middle| k \in [\![1, D]\!]\right\}$, which is introduced to assign a *weight* to each dimension of the representation $\boldsymbol{h}_i^j$ by

$$\tilde{\boldsymbol{h}}_i^j = \boldsymbol{h}_i^j \otimes \mathcal{M} \tag{4}$$

where $\tilde{\boldsymbol{h}}_i^j$ denotes the masked representation, and $\cdot \otimes \cdot$ is an element-wise Hadamard product function. $\tilde{\boldsymbol{h}}_i^j$ is then fed into two separate projection heads $g^{drr}(\cdot)$ and $g_\vartheta^{cl}(\cdot)$, respectively. $g_\vartheta^{cl}(\cdot)$ plays the same role as in conventional contrastive learning methods, i.e., the contrastive loss $\mathcal{L}_{contrast}$ in Eq. (1) is applied to the output features of $g_\vartheta^{cl}(\cdot)$. For the projection head $g^{drr}(\cdot)$, the redundancy-reduction loss $\mathcal{L}_{drr}$ in Eq. (2) is performed to its output features to tackle the dimensional redundancy.

In training, the encoder $f_\theta(\cdot)$, the projection heads $g^{drr}(\cdot)$ and $g_\vartheta^{cl}(\cdot)$, as well as the dimensional mask $\mathcal{M}$ are trained in a meta-learning manner consisting of two steps. In the first regular training step, we train $f_\theta(\cdot)$, $g^{drr}(\cdot)$ and $g_\vartheta^{cl}(\cdot)$ by jointly minimizing the redundancy-reduction and contrastive losses, which is formalized by

$$\mathcal{L}_{regular} = \mathcal{L}_{drr} + \alpha \cdot \mathcal{L}_{contrast} \tag{5}$$

where α is a coefficient that controls the balance between $\mathcal{L}_{drr}$ and $\mathcal{L}_{contrast}$, and we evaluate the influence of this hyper-parameter in Section 6.

In the second meta-learning-based step, we update $\mathcal{M}$ by using the second-derivative technique [40] to solve a bi-level optimization problem. We encourage $\mathcal{M}$ to *mask* (reduce the gradient weights of) the specific dimensions containing task-irrelevant information, which are regarded as dimensional confounders, so that the encoders can better explore the discriminative information in training. To this end, $\mathcal{M}$ can be updated by computing its gradients with respect to the performance of $f_\theta(\cdot)$ and $g_\vartheta^{cl}(\cdot)$. The corresponding performance is measured by using the gradients of $f_\theta(\cdot)$ and $g_\vartheta^{cl}(\cdot)$ during the back-propagation of contrastive loss. Formally, we update the dimensional mask $\mathcal{M}$ by

$$\arg\min_{\mathcal{M}} \mathcal{L}_{contrast}\left(g_{\vartheta_{trial}}^{cl}\left(f_{\theta_{trial}}\left(\mathcal{X}'\right) \otimes \mathcal{M}\right)\right) \tag{6}$$

where $\mathcal{X}'$ represents a minibatch sampled from the training dataset $\mathcal{X}$, ϑ_{trial} and θ_{trial} represent the *trial* weights of the encoders and projection heads after one gradient update using the contrastive loss defined in Equation 1, respectively. We formulate the updating of such trial weights as follows:

$$\theta_{trial} = \theta - \ell_\theta \nabla_\theta \mathcal{L}_{contrast}\left(g_\vartheta^{cl}\left(f_\theta\left(\mathcal{X}'\right) \otimes \mathcal{M}\right)\right), \vartheta_{trial} = \vartheta - \ell_\vartheta \nabla_\vartheta \mathcal{L}_{contrast}\left(g_\vartheta^{cl}\left(f_\theta\left(\mathcal{X}'\right) \otimes \mathcal{M}\right)\right) \tag{7}$$

where ℓ_θ and ℓ_ϑ are learning rates. The intuition behind such a behavior is that we perform a derivative over the derivative (Hessian matrix) of the combination $\{\theta, \vartheta\}$ to update $\mathcal{M}$, i.e., the second-derivative trick. Specifically, we compute the derivative with respect to $\mathcal{M}$ by using a retained computational graph of $\{\theta, \vartheta\}$ and then update $\mathcal{M}$ by back-propagating the derivative as Equation 6.

We train $\mathcal{M}$ only based on the performance of contrastive learning for two reasons: 1) the dimensional redundancy reduction is a regularization term with strong constraints and generally stable convergence, and it is independent of downstream tasks; 2) such a meta-learning-based training approach that promotes contrastive learning performance empowers our method to constrain the upper and lower bounds of the cross-entropy loss on downstream tasks, which ensures that $\mathcal{M}$ can partially mask the gradient contributions of dimensions containing task-irrelevant information and further promote the encoder to focus on learning task-relevant information, which is demonstrated in Section 5.

The two steps for updating $f_\theta(\cdot)$, $g^{drr}(\cdot)$, and $g_\vartheta^{cl}(\cdot)$ and updating $\mathcal{M}$ are iteratively repeated until convergence. The training pipeline is detailed by Algorithm 1.

5 Theoretical Analyses

Current contrastive methods follow the fundamental assumption of multi-view learning (Assumption 1 in [35], Assumption 1 in [36], and Assumption 4.1 in [41]), i.e., views (or distortions) of a same

Algorithm 1 MetaMask

Input: Multi-view dataset $\mathcal{X}$, minibatch size N, and a hyper-parameter α.
Initialize The neural network parameters: θ for $f_\theta(\cdot)$, ϑ for $g_\vartheta^{cl}(\cdot)$, ϑ^{drr} for $g^{drr}(\cdot)$, and $\mathcal{M} = \left\{\boldsymbol{\omega}_k \middle| k \in [\![1, D]\!]\right\}$. The learning rates: ℓ_θ and ℓ_ϑ, etc.
repeat
  for t-th training iteration **do**
    Iteratively sample a minibatch $\mathcal{X}'$ with N examples from $\mathcal{X}$.
    $\#\ regular\ training\ step,\ fix\ \mathcal{M}$
    $\arg\min_{\theta,\vartheta,\vartheta^{drr}} \mathcal{L}_{drr}\left(g_{\vartheta^{drr}}^{drr}\left(f_\theta\left(\mathcal{X}'\right) \otimes \mathcal{M}\right)\right) + \alpha \cdot \mathcal{L}_{contrast}\left(g_\vartheta^{cl}\left(f_\theta\left(\mathcal{X}'\right) \otimes \mathcal{M}\right)\right)$
    $\#\ compute\ trial\ weights\ and\ retain\ computational\ graph\ ,\ fix\ \theta\ and\ \vartheta$
    $\theta_{trial} = \theta - \ell_\theta \nabla_\theta \mathcal{L}_{contrast}\left(g_\vartheta^{cl}\left(f_\theta\left(\mathcal{X}'\right) \otimes \mathcal{M}\right)\right)$
    $\vartheta_{trial} = \vartheta - \ell_\vartheta \nabla_\vartheta \mathcal{L}_{contrast}\left(g_\vartheta^{cl}\left(f_\theta\left(\mathcal{X}'\right) \otimes \mathcal{M}\right)\right)$
    $\#\ meta\ training\ step\ using\ second\ derivative$
    $\arg\min_{\mathcal{M}} \mathcal{L}_{contrast}\left(g_{\vartheta_{trial}}^{cl}\left(f_{\theta_{trial}}\left(\mathcal{X}'\right) \otimes \mathcal{M}\right)\right)$
  end for
until θ, ϑ, ϑ^{drr}, and $\mathcal{M}$ converge.

sample holds the invariant discriminative label-relevant information. Theorem 4.2 in [41] further proposes the guarantees for general encoders in the learning paradigm of regular contrastive learning, which proves that the contrastive loss in the self-supervised learning stage can constrain the upper and lower bounds of the cross-entropy loss in the supervised learning stage for downstream tasks. To elaborate the behavior of our MetaMask, we extend this Theorem to build a connection between the proposed masked representation and downstream performance by

Theorem 5.1. *(Connecting Masked Representations to Downstream Cross-Entropy Loss). Under the general assumption of self-supervised multi-view learning, when $\mathcal{M}$ is optimal, for any coupled $\boldsymbol{h}, \tilde{\boldsymbol{h}} \in \mathbb{R}$, $\mathcal{L}_{crossentropy}(\boldsymbol{h})$ for downstream classification can be bounded by $\mathcal{L}_{contrast}\left(g_\vartheta^{cl}\left(\tilde{\boldsymbol{h}}\right)\right)$ in self-supervised learning:*

$$\mathcal{L}_{contrast}\left(g_\vartheta^{cl}\left(\tilde{\boldsymbol{h}}\right)\right) - \sqrt{\Phi\left(g_\vartheta^{cl}\left(\tilde{\boldsymbol{h}}\right)\middle|\boldsymbol{y}\right)} - \frac{1}{2} \cdot \sum_{k \in [\![1,D]\!]} \sqrt{\Phi\left(\lfloor g_\vartheta^{cl}\left(\tilde{\boldsymbol{h}}\right)\rceil^k \middle|\boldsymbol{y}\right)} - \mathcal{O}\left(M^{-\frac{1}{2}}\right)$$
$$\leq \mathcal{L}_{crossentropy}(\boldsymbol{h}) + \log\left(\frac{M}{D}\right) \leq \mathcal{L}_{contrast}\left(g_\vartheta^{cl}\left(\tilde{\boldsymbol{h}}\right)\right) + \sqrt{\Phi\left(g_\vartheta^{cl}\left(\tilde{\boldsymbol{h}}\right)\middle|\boldsymbol{y}\right)} + \mathcal{O}\left(M^{-\frac{1}{2}}\right) \tag{8}$$

where M denotes the quantity of negative samples, D denotes the dimensionality of the representation, $\boldsymbol{h}$ denotes the unmasked representation, $\tilde{\boldsymbol{h}}$ denotes the masked representation obtained by Equation 4, $\boldsymbol{y}$ denotes the target label, $\Phi(\cdot|\boldsymbol{y}) = \mathbb{E}_{\mathcal{P}(y)}\left[\mathbb{E}_{\mathcal{P}(\cdot|y)}\delta\left(\cdot, \mathbb{E}_{\mathcal{P}(\cdot|y)}\right)\right]$ represents the conditional variance function where $\delta(\cdot,\cdot)$ is a discriminating function to measure the difference between the terms, e.g., $\delta(\cdot,\cdot) = \|[\cdot] - [\cdot]\|^2$ in low-dimensional embedding space or $\delta(\cdot,\cdot) = -\log(\cos(\cdot,\cdot)) = -\log\frac{[\cdot]\times[\cdot]}{||\cdot||\times||\cdot||}$ in high-dimensional embedding space, $\lfloor\cdot\rceil^k$ is a function acquiring k-th dimension vector, $\log\left(\frac{M}{D}\right)$ is a constant, and $\mathcal{O}\left(M^{-\frac{1}{2}}\right)$ is the approximation error's order.

Theorem 5.1 states that reducing the risk of contrastive learning can improve the performance on downstream tasks, which further supports our intuition to make the model focus on the acquisition of discriminative information by learning a dimensional mask with the objective of improving the performance of masked representations on contrastive learning is theoretically sound. Refer to Appendix A.1 for the elaboration of MetaMask's behavior to mask the gradient contribution of specific dimensions containing confounder information. Hence, we derive

Theorem 5.2. *(Guarantees for Reduced Conditional Variance of Masked Representations). When $\mathcal{M}$ is optimal, for any coupled $\boldsymbol{h}, \tilde{\boldsymbol{h}} \in \mathbb{R}$, given label $\boldsymbol{y}$, the conditional variance of $\tilde{\boldsymbol{h}}$ is reduced:*

$$\Phi\left(g_\vartheta^{cl}\left(\tilde{\boldsymbol{h}}\right)\middle|\boldsymbol{y}\right) \leq \Phi\left(g_\vartheta^{cl}(\boldsymbol{h})\middle|\boldsymbol{y}\right),\ yet\ \ \Phi\left(\lfloor g_\vartheta^{cl}\left(\tilde{\boldsymbol{h}}\right)\rceil^k \middle|\boldsymbol{y}\right) \cong \Phi\left(\lfloor g_\vartheta^{cl}(\boldsymbol{h})\rceil^k \middle|\boldsymbol{y}\right). \tag{9}$$

Theorem 5.2 states that given the label y, the masked representation $\tilde{\boldsymbol{h}}$ has smaller conditional variance than $\boldsymbol{h}$ in contrastive learning. We bring Theorem 5.2 into Theorem 5.1 to derive a conclusion: our

Table 1: Comparison of different methods on classification accuracy (top 1). We follow the experimental setting of [5] and adopt *conv* and *fc* as the backbones in the experiments.

Model	IN-200 [42]		STL-10 [43]		CIFAR-10 [42]		CIFAR-100 [42]	
	conv	fc	conv	fc	conv	fc	conv	fc
Barlow Twins [14]	39.81	40.34	80.97	81.43	76.63	78.49	52.80	52.95
+ MetaMask$_{+4.95}$	42.91$_{+3.10}$	42.06$_{+1.72}$	84.96$_{+3.99}$	86.13$_{+4.70}$	85.24$_{+8.61}$	86.90$_{+8.41}$	58.02$_{+5.22}$	56.78$_{+3.83}$
SwAV [4]	39.67	39.02	74.65	75.30	72.19	72.36	46.58	46.75
+ MetaMask$_{+5.00}$	40.28$_{+0.61}$	40.05$_{+1.03}$	81.34$_{+6.69}$	82.19$_{+6.89}$	80.37$_{+8.18}$	77.22$_{+4.86}$	52.07$_{+5.49}$	52.98$_{+6.23}$
SimCLR [6]	36.24	39.83	75.57	77.15	80.58	80.07	50.03	49.82
+ MetaMask$_{+3.60}$	40.56$_{+4.32}$	40.82$_{+0.99}$	81.37$_{+5.80}$	82.90$_{+5.75}$	82.85$_{+2.27}$	81.64$_{+1.57}$	53.70$_{+3.67}$	54.24$_{+4.42}$
CMC [5]	41.58	40.11	83.03	85.06	81.31	83.28	58.13	56.72
+ MetaMask$_{+1.73}$	42.91$_{+1.33}$	42.06$_{+1.95}$	84.96$_{+1.93}$	86.13$_{+1.07}$	85.24$_{+3.93}$	86.90$_{+3.62}$	58.02$_{-0.09}$	56.78$_{+0.06}$
BYOL [10]	41.59	41.90	81.73	81.57	77.18	80.01	53.64	53.78
+ MetaMask$_{+1.68}$	42.72$_{+1.13}$	43.29$_{+1.39}$	83.01$_{+1.28}$	83.44$_{+1.87}$	79.26$_{+2.08}$	82.73$_{+2.72}$	55.68$_{+2.04}$	54.72$_{+0.94}$
SimSiam [12]	41.03	41.27	80.91	81.88	78.14	81.13	52.55	53.52
+ MetaMask$_{+1.44}$	41.82$_{+0.79}$	42.67$_{+1.40}$	82.35$_{+1.44}$	83.96$_{+2.08}$	80.03$_{+1.89}$	81.39$_{+0.26}$	54.70$_{+2.15}$	55.02$_{+1.50}$
DCL [8]	38.79	40.26	77.09	78.39	80.89	80.93	51.38	51.09
+ MetaMask$_{+1.83}$	40.28$_{+1.49}$	41.69$_{+1.43}$	80.94$_{+3.85}$	81.06$_{+2.67}$	81.33$_{+0.44}$	81.34$_{+0.41}$	53.05$_{+1.67}$	53.74$_{+2.65}$
HCL [9]	40.05	41.23	79.86	80.20	82.13	82.76	52.69	53.13
+ MetaMask$_{+1.14}$	40.43$_{+0.38}$	41.01$_{-0.22}$	81.52$_{+1.66}$	82.09$_{+1.89}$	83.65$_{+1.52}$	83.95$_{+1.19}$	54.03$_{+1.34}$	54.26$_{+1.13}$

approach can better bound the downstream classification risk, i.e., the upper and lower bounds of supervised cross-entropy loss obtained by MetaMask are tighter than typical contrastive learning methods. Please refer to Appendix A.2 for the corresponding proofs.

6 Experiments

6.1 Benchmarking MetaMask with Various Backbones

Implementations. For the comparisons demonstrated in Table 1, we uniformly set the batch size as 64, and we adopt a network with the 5 convolutional layers in AlexNet [44] as *conv* and a network with 2 additional fully connected layers as *fc*. We straightforwardly adopt the same data augmentations in [5] and [45], and the memory bank [46] is adopted to facilitate calculations by retrieving 4,096 past negative features during training. Following the experimental principle of linear probing, we remove the projection head and re-train an MLP as the classification head for the learned representation.

For the experiments in Table 2, the batch size is valued by 512, and ResNet-18 [47] is used as the backbone encoder. We adopt the data augmentation and other experimental settings following [14]. Note that the KNN classifier is adopted for evaluation.

Discussion of the comparisons using conv and fc encoders. Table 1 shows the comparisons on four benchmark datasets, and the corresponding positive or negative margins are reported. We observe that MetaMask beats the best prior methods by a significant margin on most downstream tasks, which demonstrates that MetaMask can efficiently improve the performance of the learned representation when the self-supervision is insufficient, i.e., the limited batch size and weak encoder.

Discussion of the comparisons using the ResNet encoder. We further perform classification comparisons using a strong encoder, i.e., ResNet. Table 2 shows that MetaMask can consistently improve the state-of-the-art methods, which indicates that the proposed method has strong adaptability to different encoders. The experimental evidence proves our assumption of the dimensional confounder. Thus, given a fixed dimensionality of the projection head, MetaMask can indeed improve the discriminativeness of the learned representation by *masking* the dimensional confounder.

6.2 Validation of the Robustness of MetaMask against Dimensional Confounder

We conduct experiments to prove that MetaMask is able to mitigate the negative impact of the dimensional confounder on the discriminativeness of the learned representations, and the results are shown in Figure 4. The intriguing outcome demonstrates two merits of the proposed method: 1)

Table 2: Comparison of different methods on classification accuracy (top 1). We adopt *ResNet-18* as the backbones.

Model	CIFAR-10	CIFAR-100	STL-10	IN-200
Barlow Twins [14]	85.72	60.31	73.62	45.10
+ MetaMask$_{+1.25}$	87.53$_{+1.81}$	61.42$_{+1.11}$	73.97$_{+0.35}$	46.81$_{+1.71}$
SimCLR [6]	81.73	58.27	70.09	44.46
+ MetaMask$_{+3.32}$	86.01$_{+4.28}$	61.03$_{+2.76}$	74.90$_{+4.81}$	45.87$_{+1.41}$
BYOL [10]	88.05	60.94	72.04	46.72
+ MetaMask$_{+0.28}$	87.53$_{-0.52}$	61.42$_{+0.48}$	73.97$_{+1.93}$	46.81$_{+0.11}$
DCL [8]	85.61	59.29	71.18	45.77
+ MetaMask$_{+1.65}$	86.47$_{+0.86}$	60.82$_{+1.53}$	74.79$_{+3.61}$	46.35$_{+0.58}$
HCL [9]	85.27	61.21	71.92	46.90
+ MetaMask$_{+0.95}$	85.97$_{+0.70}$	61.63$_{+0.42}$	74.03$_{+2.11}$	46.06$_{+0.55}$
NNCLR [50]	81.44	61.64	71.89	47.10
+ MetaMask$_{+1.63}$	85.78$_{+4.34}$	61.89$_{+0.25}$	74.02$_{+2.13}$	46.90$_{-0.20}$

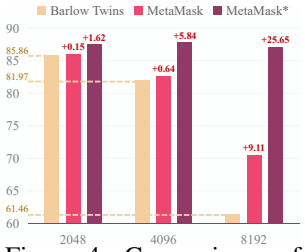


Figure 4: Comparisons of varying the dimensionality of the projection head on CIFAR-10 by using ResNet-18. $*$ denotes MetaMask using a trick of fixed learning rate instead of the cosine annealing strategy.

even if the dimensionality of the projection head changes, MetaMask is consistently robust to the dimensional confounder, i.e., the model maintains relatively consistent performance; 2) MetaMask can alleviate the negative impact of *curse of dimensionality* [48, 49], e.g., given a fixed number of examples, as the dimensionality of the representation increases during training, the performance of the model on downstream tasks instead degenerates. Concretely, Figure 2 demonstrates that regardless of whether representation learning suffers from the curse of dimensionality, dimensionality confounding factors are widely present in learned representations, and hence, the defined dimensional confounder is a more generalized issue of the curse of dimensionality. Refer Appendix A.3 for the further discussion. Moreover, we observe a surprising result that MetaMask* achieves better performance. We provide an understandable explanation: the essence of the curse of dimensionality is that a network with a large number of parameters is prone to over-fitting to the training data. Therefore, compared with the annealing strategy, a fixed and relatively large learning rate can prevent the model from falling into a local optimum (over-fitting) during optimization. Our further experiments show that MetaMask* collapses abruptly at larger dimensional settings, e.g., 16384, which demonstrates that the trick of using a fixed learning rate can only unsteadily improve the performance of MetaMask, but this behavior is *not* essential. Figure 4 shows that under the different experimental settings, vanilla MetaMask consistently counteracts the negative effects of the dimensional confounder.

7 Conclusion and Future Discussion

From the dimensional perspective, we first clarify that the dimensional redundancy and confounder issues are the essence of the current self-supervised learning paradigm's defects. Multiple motivating experiments are conducted to support our viewpoint, and the intriguing results are opposite to the traditional conclusions. To jointly tackle the dimensional redundancy and confounder issues, we propose a heuristic learning approach, called MetaMask, which leverages a second-derivative technique to learn a dimensional mask in order to adjust the gradient weights of different dimensions during training. The proposed theoretical analyses demonstrate that MetaMask obtains relatively tighter risk bounds for downstream classification compared to conventional contrastive methods. From the experiments, we observe a consistent improvement in the performance of our method over state-of-the-art methods on benchmark datasets.

One of our critical contributions is proposing a heuristic learning paradigm for self-supervised learning: regarding the dimensional redundancy reduction as a regularization term and further exploring to mask dimensions containing the confounder. In the current methodology, we train the learnable dimensional mask with the objective of improving the performance of masked representations on a specific *self-supervised task*, i.e., contrastive learning, and provide the theoretical analysis to prove that such a learning approach can indeed prompt the model focus on acquiring discriminative information. However, contrastive learning is immutable for the *self-supervised task* in MetaMask. It is commonly acknowledged that masked image modeling (MIM) methods [51, 52, 53] achieve impressive performance in self-supervised learning. Although the theoretical contribution bridging the gap between MIM loss and cross-entropy loss on downstream tasks has not yet emerged, we still deem combining MIM approaches with our proposed paradigm is attractive to the community.

References

[1] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[2] Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *NeurIPS 2019*, 2019.

[3] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. 2020.

[5] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research. PMLR, 2020.

[7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020.

[8] C. Y. Chuang, J. Robinson, Y. C. Lin, A. Torralba, and S. Jegelka. Debiased contrastive learning. 2020.

[9] J. Robinson, C. Y. Chuang, S. Sra, and S. Jegelka. Contrastive learning with hard negative samples. 2020.

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[11] A. Ermolov, A. Siarohin, E. Sangineto, and N. Sebe. Whitening for self-supervised representation learning. 2020.

[12] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021.

[13] M. Glymour J. Pearl and N.P. Jewell. Causal inference in statistics: A primer. Wiley, 2016.

[14] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research. PMLR, 2021.

[15] H. B Barlow. Possible principles underlying the transformations of sensory messages. In *Sensory Communication*. The MIT Press, 1961.

[16] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015.

[17] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*. ACM, 2008.

[18] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[19] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2014.

[20] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019.

[21] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martín Arjovsky, Olivier Mastropietro, and Aaron C. Courville. Adversarially learned inference. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[22] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[23] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[24] OJ Hénaff, A. Srinivas, J De Fauw, A. Razavi, C. Doersch, Sma Eslami, and Avd Oord. Data-efficient image recognition with contrastive predictive coding. 2019.

[25] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020.

[26] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[27] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[28] Vikas Verma, Thang Luong, Kenji Kawaguchi, Hieu Pham, and Quoc V. Le. Towards domain-agnostic contrastive learning. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research. PMLR, 2021.

[29] Tete Xiao, Xiaolong Wang, Alexei A. Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[30] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. 2010.

[31] J. Goldberger, S. Gordon, and H. Greenspan. An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures. In *IEEE International Conference on Computer Vision*, 2003.

[32] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, 2020.

[33] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*. ACM, 2008.

[34] Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2014.

[35] K. Sridharan and S. M. Kakade. An information theoretic framework for multi-view learning. *Conference on Learning Theory*, 2008.

[36] Yhh Tsai, Y. Wu, R. Salakhutdinov, and L. P. Morency. Self-supervised learning from a multi-view perspective. 2020.

[37] Yazhe Li, Roman Pogodin, Danica J. Sutherland, and Arthur Gretton. Self-supervised learning with kernel dependence maximization. *CoRR*, 2021.

[38] A. Smola A. Gretton, O. Bousquet and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. *ALT, Springer*, 2005.

[39] Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[40] S. Liu, Andrew J Davison, and E. Johns. Self-supervised generalisation with meta auxiliary learning. 2019.

[41] Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. *arXiv preprint arXiv:2203.13457*, 2022.

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

[43] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011.

[44] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[45] Jiangmeng Li, Wenwen Qiang, Changwen Zheng, Bing Su, and Hui Xiong. Metaug: Contrastive learning via meta feature augmentation. *arXiv preprint arXiv:2203.05119*, 2022.

[46] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. *Unsupervised feature learning via non-parametric instance discrimination*. 2018.

[47] Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[48] R. Bellman and R. Karush. Dynamic programming: A bibliography of theory and application,. *Dynamic Programming A Bibliography of Theory and Application*, 1964.

[49] Vincent Spruyt. The curse of dimensionality in classification. 2014.

[50] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. 2021.

[51] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[52] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[53] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. *arXiv preprint arXiv:2111.09886*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [No] Since this work presents a general technique to tackle the dimensional confounder for self-supervised learning, we did not see any particular foreseeable ethical consequence.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 5.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix A.1 and Appendix A.2.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the supplementary files.
   (b) Did you specify all the training details (e.g., data splits, hyper-parameters, how they were chosen)? [Yes] See Section 6, Appendix A.4, and the supplementary files.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] The experimental results provided by benchmark methods do not contain such error bars. We also follow this principle, and we reckon the reason behind such a behavior is that the comparisons are conducted in balanced datasets, so the difference between results is not excessively large.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 1 and Appendix A.4.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 1, Section 2, Section 5, and Section 6.
   (b) Did you mention the license of the assets? [No] The assets we adopted are totally established and public to the community.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See the supplementary files.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] The datasets are all public to the community.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The datasets have been discussed, validated, and used by many researchers.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

A.1 Elaborating MetaMask's Behavior of Masking Gradient Contribution

For better understanding the intuition behind the behavior of the proposed MetaMask, we elaborate the procedure of our method masking gradient contributions of dimensions. We restate Equation 1 of contrastive loss by converting the original projected features $\boldsymbol{z}$ into the masked feature $\tilde{\boldsymbol{z}}$:

$$\begin{aligned}\mathcal{L}_{maskedcl} &= -\sum\sum \log \frac{\exp\left[d\left(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{z}}^{+}\right)/\tau\right]}{\sum\sum \mathbb{1}\cdot\exp\left[d\left(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{z}}'\right)/\tau\right]} \\ &= -\sum\sum \log \frac{\exp\left[d\left(g_{\vartheta}^{cl}\left(\tilde{\boldsymbol{h}}\right), g_{\vartheta}^{cl}\left(\tilde{\boldsymbol{h}}^{+}\right)\right)/\tau\right]}{\sum\sum \mathbb{1}\cdot\exp\left[d\left(g_{\vartheta}^{cl}\left(\tilde{\boldsymbol{h}}\right), g_{\vartheta}^{cl}\left(\tilde{\boldsymbol{h}}'\right)\right)/\tau\right]} \\ &= -\sum\sum \log \frac{\exp\left[d\left(g_{\vartheta}^{cl}\left(\boldsymbol{h}\otimes\mathcal{M}\right), g_{\vartheta}^{cl}\left(\boldsymbol{h}^{+}\otimes\mathcal{M}\right)\right)/\tau\right]}{\sum\sum \mathbb{1}\cdot\exp\left[d\left(g_{\vartheta}^{cl}\left(\boldsymbol{h}\otimes\mathcal{M}\right), g_{\vartheta}^{cl}\left(\boldsymbol{h}'\otimes\mathcal{M}\right)\right)/\tau\right]},\end{aligned} \tag{10}$$

which is derived by considering Equation 4. To simplify the equation, we hold

$$\mathcal{L}_{maskedcl} = \sum_{i\in[\![1,N]\!]} \sum_{\substack{j\in[\![1,M-1]\!] \\ j^{+}\in[\![j+1,M]\!]}} \mathcal{L}_{i,j,j^{+}}. \tag{11}$$

Therefore, the indicator function $\mathbb{1}_{[i\neq i' \vee j\neq j']}$ can be substituted as follows:

$$\mathcal{L}_{i,j,j^{+}} = \log \frac{-\exp\left[d\left(g\left(\boldsymbol{h}\otimes\mathcal{M}\right), g\left(\boldsymbol{h}^{+}\otimes\mathcal{M}\right)\right)/\tau\right]}{\exp\left[d\left(g\left(\boldsymbol{h}\otimes\mathcal{M}\right), g\left(\boldsymbol{h}^{+}\otimes\mathcal{M}\right)\right)/\tau\right] + \sum\limits_{\boldsymbol{h}^{-}}\exp\left[d\left(g\left(\boldsymbol{h}\otimes\mathcal{M}\right), g\left(\boldsymbol{h}^{-}\otimes\mathcal{M}\right)\right)/\tau\right]}, \tag{12}$$

where $g\left(\cdot\right)$ denotes $g_{\vartheta}^{cl}\left(\cdot\right)$, and $\boldsymbol{h}^{-}$ is neither equal to $\boldsymbol{h}$ nor equal to $\boldsymbol{h}^{+}$.

Proposition A.1. *The mask $\mathcal{M}$ exists on each partial differential equation of the gradient of $\mathcal{L}_{i,j,j^{+}}$:*

$$\begin{aligned}\nabla_{\boldsymbol{h}}\mathcal{L}_{i,j,j^{+}} =& \frac{\sum\limits_{\boldsymbol{h}^{-}}\exp\left[d\left(g\left(\boldsymbol{h}\otimes\mathcal{M}\right), g\left(\boldsymbol{h}^{-}\otimes\mathcal{M}\right)\right)/\tau\right] - \exp\left[d\left(g\left(\boldsymbol{h}\otimes\mathcal{M}\right), g\left(\boldsymbol{h}^{+}\otimes\mathcal{M}\right)\right)/\tau\right]}{\sum\limits_{\boldsymbol{h}^{-}}\exp\left[d\left(g\left(\boldsymbol{h}\otimes\mathcal{M}\right), g\left(\boldsymbol{h}^{-}\otimes\mathcal{M}\right)\right)/\tau\right]} \\ &\times\left[\sum_{\boldsymbol{h}^{-\prime}}\frac{\exp\left[d\left(g\left(\boldsymbol{h}\otimes\mathcal{M}\right), g\left(\boldsymbol{h}^{-\prime}\otimes\mathcal{M}\right)\right)/\tau\right]}{\sum\limits_{\boldsymbol{h}^{-}}\exp\left[d\left(g\left(\boldsymbol{h}\otimes\mathcal{M}\right), g\left(\boldsymbol{h}^{-}\otimes\mathcal{M}\right)\right)/\tau\right]}\times g\left(\boldsymbol{h}^{-\prime}\otimes\mathcal{M}\right) - g\left(\boldsymbol{h}^{+}\otimes\mathcal{M}\right)\right] \\ &\times\prod_{k\in[\![1,N^{g}]\!]}\boldsymbol{w}_{k}\times\nabla_{\boldsymbol{h}}\boldsymbol{h}\otimes\mathcal{M}\end{aligned} \tag{13}$$

$$\begin{aligned}\nabla_{\boldsymbol{h}^{+}}\mathcal{L}_{i,j,j^{+}} =& \frac{\exp\left[d\left(g\left(\boldsymbol{h}\otimes\mathcal{M}\right), g\left(\boldsymbol{h}^{+}\otimes\mathcal{M}\right)\right)/\tau\right] - \sum\limits_{\boldsymbol{h}^{-}}\exp\left[d\left(g\left(\boldsymbol{h}\otimes\mathcal{M}\right), g\left(\boldsymbol{h}^{-}\otimes\mathcal{M}\right)\right)/\tau\right]}{\sum\limits_{\boldsymbol{h}^{-}}\exp\left[d\left(g\left(\boldsymbol{h}\otimes\mathcal{M}\right), g\left(\boldsymbol{h}^{-}\otimes\mathcal{M}\right)\right)/\tau\right]} \\ &\times g\left(\boldsymbol{h}\otimes\mathcal{M}\right)\times\prod_{k\in[\![1,N^{g}]\!]}\boldsymbol{w}_{k}\times\nabla_{\boldsymbol{h}^{+}}\boldsymbol{h}^{+}\otimes\mathcal{M}\end{aligned} \tag{14}$$

$$\begin{aligned}\nabla_{\boldsymbol{h}^{-}}\mathcal{L}_{i,j,j^{+}} =& \frac{\sum\limits_{\boldsymbol{h}^{-}}\exp\left[d\left(g\left(\boldsymbol{h}\otimes\mathcal{M}\right), g\left(\boldsymbol{h}^{-}\otimes\mathcal{M}\right)\right)/\tau\right] - \exp\left[d\left(g\left(\boldsymbol{h}\otimes\mathcal{M}\right), g\left(\boldsymbol{h}^{+}\otimes\mathcal{M}\right)\right)/\tau\right]}{\sum\limits_{\boldsymbol{h}^{-}}\exp\left[d\left(g\left(\boldsymbol{h}\otimes\mathcal{M}\right), g\left(\boldsymbol{h}^{-}\otimes\mathcal{M}\right)\right)/\tau\right]} \\ &\times\frac{\exp\left[d\left(g\left(\boldsymbol{h}\otimes\mathcal{M}\right), g\left(\boldsymbol{h}^{-}\otimes\mathcal{M}\right)\right)/\tau\right]}{\sum\limits_{\boldsymbol{h}^{-\prime}}\exp\left[d\left(g\left(\boldsymbol{h}\otimes\mathcal{M}\right), g\left(\boldsymbol{h}^{-\prime}\otimes\mathcal{M}\right)\right)/\tau\right]}\times g\left(\boldsymbol{h}\otimes\mathcal{M}\right)\times\prod_{k\in[\![1,N^{g}]\!]}\boldsymbol{w}_{k} \\ &\times\nabla_{\boldsymbol{h}^{-}}\boldsymbol{h}^{-}\otimes\mathcal{M}\end{aligned} \tag{15}$$

where $\left\{\boldsymbol{w}_k \middle| k \in [\![1, N^g]\!]\right\}$ denotes the weights of the projection head $g(\cdot)$, N^g denotes the number of corresponding projection layers, and $\boldsymbol{h}^{-'}$ is neither equal to $\boldsymbol{h}$ nor equal to $\boldsymbol{h}^+$ (note that $\boldsymbol{h}^-$ and $\boldsymbol{h}^{-'}$ are two independent variables). $\mathcal{M}$ is the dimensional mask.

Proposition A.1 states that the dimensional mask $\mathcal{M}$ consistently exists on each partial differential equation of the gradient of the contrastive loss $\mathcal{L}_{i,j,j^+}$. We further observe that there exists 7 types of irreducible terms containing $\mathcal{M}$: $g(\boldsymbol{h} \otimes \mathcal{M})$, $g(\boldsymbol{h}^+ \otimes \mathcal{M})$, $g(\boldsymbol{h}^- \otimes \mathcal{M})$, $g\left(\boldsymbol{h}^{-'} \otimes \mathcal{M}\right)$, $\nabla_{\boldsymbol{h}} \boldsymbol{h} \otimes \mathcal{M}$, $\nabla_{\boldsymbol{h}^+} \boldsymbol{h} \otimes \mathcal{M}$, $\nabla_{\boldsymbol{h}^-} \boldsymbol{h} \otimes \mathcal{M}$, where the first 4 types are in the regular computing manner so that $\mathcal{M}$ can directly change the values of different dimensions by assigning dimension-specific weights, and the last 3 types are the first-order derivatives. Note that the scalar derivation, e.g., the loss, of a matrix can be thought of as derivation of each element of the matrix, and such a derivation is derived and then putting the element-wise results in the order of the matrix to get a gradient matrix of the same dimension.

Following such a law, we utilize Chain Rules to derive the derivative of the loss $\mathcal{L}_{i,j,j^+}$ with respect to each element in the batch matrix containing $\boldsymbol{h}$, $\boldsymbol{h}^+$, $\boldsymbol{h}^-$ and further obtain the derivative matrix:

$$\frac{\partial \mathcal{L}_{i,j,j^+}}{\partial \mathcal{X}'_{p,q}} = \frac{\partial \mathcal{L}_{i,j,j^+}}{\partial \mathcal{B}_{p,q}} \frac{\partial \mathcal{B}_{p,q}}{\partial \mathcal{X}'_{p,q}} = \frac{\partial \mathcal{L}_{i,j,j^+}}{\partial \left[\left(\mathcal{B} \otimes \hat{\mathcal{M}}\right)_{p,q}\right]} \frac{\partial \left[\left(\mathcal{B} \otimes \hat{\mathcal{M}}\right)_{p,q}\right]}{\partial \mathcal{B}_{p,q}} \frac{\partial \mathcal{B}_{p,q}}{\partial \mathcal{X}'_{p,q}}$$

$$= \begin{bmatrix} \boldsymbol{\omega_1} \frac{\partial \lfloor f_\theta(\boldsymbol{x}) \rceil^1}{\partial \boldsymbol{x}} & \boldsymbol{\omega_2} \frac{\partial \lfloor f_\theta(\boldsymbol{x}) \rceil^2}{\partial \boldsymbol{x}} & \cdots & \boldsymbol{\omega_{D-1}} \frac{\partial \lfloor f_\theta(\boldsymbol{x}) \rceil^{D-1}}{\partial \boldsymbol{x}} & \boldsymbol{\omega_D} \frac{\partial \lfloor f_\theta(\boldsymbol{x}) \rceil^D}{\partial \boldsymbol{x}} \\ \boldsymbol{\omega_1} \frac{\partial \lfloor f_\theta(\boldsymbol{x}^+) \rceil^1}{\partial \boldsymbol{x}^+} & \boldsymbol{\omega_2} \frac{\partial \lfloor f_\theta(\boldsymbol{x}^+) \rceil^2}{\partial \boldsymbol{x}^+} & \cdots & \boldsymbol{\omega_{D-1}} \frac{\partial \lfloor f_\theta(\boldsymbol{x}^+) \rceil^{D-1}}{\partial \boldsymbol{x}^+} & \boldsymbol{\omega_D} \frac{\partial \lfloor f_\theta(\boldsymbol{x}^+) \rceil^D}{\partial \boldsymbol{x}^+} \\ \boldsymbol{\omega_1} \frac{\partial \lfloor f_\theta(\boldsymbol{x}_1^-) \rceil^1}{\partial \boldsymbol{x}_1^-} & \boldsymbol{\omega_2} \frac{\partial \lfloor f_\theta(\boldsymbol{x}_1^-) \rceil^2}{\partial \boldsymbol{x}_1^-} & \cdots & \boldsymbol{\omega_{D-1}} \frac{\partial \lfloor f_\theta(\boldsymbol{x}_1^-) \rceil^{D-1}}{\partial \boldsymbol{x}_1^-} & \boldsymbol{\omega_D} \frac{\partial \lfloor f_\theta(\boldsymbol{x}_1^-) \rceil^D}{\partial \boldsymbol{x}_1^-} \\ \boldsymbol{\omega_1} \frac{\partial \lfloor f_\theta(\boldsymbol{x}_2^-) \rceil^1}{\partial \boldsymbol{x}_2^-} & \boldsymbol{\omega_2} \frac{\partial \lfloor f_\theta(\boldsymbol{x}_2^-) \rceil^2}{\partial \boldsymbol{x}_2^-} & \cdots & \boldsymbol{\omega_{D-1}} \frac{\partial \lfloor f_\theta(\boldsymbol{x}_2^-) \rceil^{D-1}}{\partial \boldsymbol{x}_2^-} & \boldsymbol{\omega_D} \frac{\partial \lfloor f_\theta(\boldsymbol{x}_2^-) \rceil^D}{\partial \boldsymbol{x}_2^-} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \boldsymbol{\omega_1} \frac{\partial \lfloor f_\theta(\boldsymbol{x}_{N-3}^-) \rceil^1}{\partial \boldsymbol{x}_{N-3}^-} & \boldsymbol{\omega_2} \frac{\partial \lfloor f_\theta(\boldsymbol{x}_{N-3}^-) \rceil^2}{\partial \boldsymbol{x}_{N-3}^-} & \cdots & \boldsymbol{\omega_{D-1}} \frac{\partial \lfloor f_\theta(\boldsymbol{x}_{N-3}^-) \rceil^{D-1}}{\partial \boldsymbol{x}_{N-3}^-} & \boldsymbol{\omega_D} \frac{\partial \lfloor f_\theta(\boldsymbol{x}_{N-3}^-) \rceil^D}{\partial \boldsymbol{x}_{N-3}^-} \\ \boldsymbol{\omega_1} \frac{\partial \lfloor f_\theta(\boldsymbol{x}_{N-2}^-) \rceil^1}{\partial \boldsymbol{x}_{N-2}^-} & \boldsymbol{\omega_2} \frac{\partial \lfloor f_\theta(\boldsymbol{x}_{N-2}^-) \rceil^2}{\partial \boldsymbol{x}_{N-2}^-} & \cdots & \boldsymbol{\omega_{D-1}} \frac{\partial \lfloor f_\theta(\boldsymbol{x}_{N-2}^-) \rceil^{D-1}}{\partial \boldsymbol{x}_{N-2}^-} & \boldsymbol{\omega_D} \frac{\partial \lfloor f_\theta(\boldsymbol{x}_{N-2}^-) \rceil^D}{\partial \boldsymbol{x}_{N-2}^-} \end{bmatrix} \tag{16}$$

where $\mathcal{B} = \left[\boldsymbol{h}, \boldsymbol{h}^+, \boldsymbol{h}_1^-, \boldsymbol{h}_2^-, \ldots, \boldsymbol{h}_{N-3}^-, \boldsymbol{h}_{N-2}^-\right]$ is a batch of features with N examples, and $\mathcal{X}'$ is the corresponding input data matrix. $\boldsymbol{h}$, $\boldsymbol{h}^+$, and $\boldsymbol{h}^-$ are feature vectors containing D dimensions. $\hat{\mathcal{M}} = [\mathcal{M}, \mathcal{M}, \ldots, \mathcal{M}, \mathcal{M}]$ is an extended matrix containing N copies of $\mathcal{M}$. Intriguingly, we observe that $\mathcal{M}$ generally assign the dimension-specific weights for each samples, which empowers MetaMask to naturally reduce the gradient effect of specific dimensions containing the confounder by employing the meta-learning paradigm with the objective of improving the performance of masked representations on a typical self-supervised task. After clarifying the dimension-specific weighting mechanism from the perspective of gradient analysis, we further demonstrate the effectiveness of MetaMask's training paradigm as follows.

A.2 Proofs

In order to prove Theorem 5.2 and the conclusion that the bounds of supervised cross-entropy loss obtained by MetaMask are tighter than typical contrastive learning methods, we provide separated proofs for the corresponding two parts of Theorem 5.2 and the tighter bounds on downstream cross-entropy loss.

A.2.1 Proof for the Equality Part

To prove

$$\Phi\left(\lfloor g_\vartheta^{cl}\left(\tilde{\boldsymbol{h}}\right)\rceil^k \middle| \boldsymbol{y}\right) \cong \Phi\left(\lfloor g_\vartheta^{cl}\left(\boldsymbol{h}\right)\rceil^k \middle| \boldsymbol{y}\right), \tag{17}$$

we observe that the only variable that differs on both sides of the Equation 17 is the representation, which is masked on the left hand side of the equation while unmasked on the right hand side of the

equation. Then, we substitute the definition of $\Phi(\cdot)$ into Equation 17 and derive

$$\begin{aligned}&\mathbb{E}_{\mathcal{P}(y)}\left[\mathbb{E}_{\mathcal{P}\left(\lfloor g_{\vartheta}^{cl}(\tilde{\boldsymbol{h}})\rceil^{k}|y\right)}\delta\left(\lfloor g_{\vartheta}^{cl}\left(\tilde{\boldsymbol{h}}\right)\rceil^{k},\mathbb{E}_{\mathcal{P}\left(\lfloor g_{\vartheta}^{cl}(\tilde{\boldsymbol{h}})\rceil^{k}|y\right)}\right)\right.\\ \cong\ &\mathbb{E}_{\mathcal{P}(y)}\left[\mathbb{E}_{\mathcal{P}\left(\lfloor g_{\vartheta}^{cl}(\boldsymbol{h})\rceil^{k}|y\right)}\delta\left(\lfloor g_{\vartheta}^{cl}(\boldsymbol{h})\rceil^{k},\mathbb{E}_{\mathcal{P}\left(\lfloor g_{\vartheta}^{cl}(\boldsymbol{h})\rceil^{k}|y\right)}\right).\end{aligned} \tag{18}$$

In our approach, the learned representation generally exists in high-dimensional embedding space, e.g., 1024, 2048, 8192, etc. Therefore, in high-dimensional embedding space, the left hand side of the equation can be extended to

$$\begin{aligned}&\mathbb{E}_{\mathcal{P}(y)}\left[\mathbb{E}_{\mathcal{P}\left(\lfloor g_{\vartheta}^{cl}(\tilde{\boldsymbol{h}})\rceil^{k}|y\right)}\delta\left(\lfloor g_{\vartheta}^{cl}\left(\tilde{\boldsymbol{h}}\right)\rceil^{k},\mathbb{E}_{\mathcal{P}\left(\lfloor g_{\vartheta}^{cl}(\tilde{\boldsymbol{h}})\rceil^{k}|y\right)}\right)\right]\\ =\ &\mathbb{E}_{\mathcal{P}(y)}\left[\mathbb{E}_{\mathcal{P}\left(\lfloor g_{\vartheta}^{cl}(\tilde{\boldsymbol{h}})\rceil^{k}|y\right)}-\log\left(\cos\left(\lfloor g_{\vartheta}^{cl}\left(\tilde{\boldsymbol{h}}\right)\rceil^{k},\mathbb{E}_{\mathcal{P}\left(\lfloor g_{\vartheta}^{cl}(\tilde{\boldsymbol{h}})\rceil^{k}|y\right)}\right)\right)\right]\\ =\ &\mathbb{E}_{\mathcal{P}(y)}\left[\mathbb{E}_{\mathcal{P}\left(\lfloor g_{\vartheta}^{cl}(\tilde{\boldsymbol{h}})\rceil^{k}|y\right)}-\log\frac{\lfloor g_{\vartheta}^{cl}\left(\tilde{\boldsymbol{h}}\right)\rceil^{k}\times\mathbb{E}_{\mathcal{P}\left(\lfloor g_{\vartheta}^{cl}(\tilde{\boldsymbol{h}})\rceil^{k}|y\right)}}{||\lfloor g_{\vartheta}^{cl}\left(\tilde{\boldsymbol{h}}\right)\rceil^{k}||\times||\mathbb{E}_{\mathcal{P}\left(\lfloor g_{\vartheta}^{cl}(\tilde{\boldsymbol{h}})\rceil^{k}|y\right)}||}\right]\\ =\ &\mathbb{E}_{\mathcal{P}\left(\lfloor g_{\vartheta}^{cl}(\tilde{\boldsymbol{h}})\rceil^{k},y\right)}-\log\frac{\lfloor g_{\vartheta}^{cl}\left(\tilde{\boldsymbol{h}}\right)\rceil^{k}\times\mu_{\boldsymbol{y}}}{||\lfloor g_{\vartheta}^{cl}\left(\tilde{\boldsymbol{h}}\right)\rceil^{k}||\times||\mu_{\boldsymbol{y}}||}\\ \approx\ &\mathbb{E}_{\mathcal{P}\left(\lfloor\tilde{\boldsymbol{h}}\rceil^{k},y\right)}-\log\frac{\lfloor\tilde{\boldsymbol{h}}\rceil^{k}\times\mu_{\boldsymbol{y}}}{||\lfloor\tilde{\boldsymbol{h}}\rceil^{k}||\times||\mu_{\boldsymbol{y}}||}\\ =\ &-\mathbb{E}_{\mathcal{P}\left(\lfloor\tilde{\boldsymbol{h}}\rceil^{k},y\right)}\log\frac{\lfloor\tilde{\boldsymbol{h}}\rceil^{k}\times\mu_{\boldsymbol{y}}}{||\lfloor\tilde{\boldsymbol{h}}\rceil^{k}||\times||\mu_{\boldsymbol{y}}||},\end{aligned} \tag{19}$$

which is derived by following the Bayes Rule, the definition of $\delta(\cdot,\cdot)$ in Theorem 5.1, and the fact that $g_{\vartheta}^{cl}(\cdot)$ is an MLP without amounts of deep layers so that the embedding space shift is limited. Note that $\mu_{\boldsymbol{y}}$ denotes the center of categories' feature or the corresponding classification weight of the classifier for a specific category, and $\boldsymbol{y}=\left\{\boldsymbol{y}_l\middle|l\in[\![1,N^L]\!]\right\}$, where N^L is the number of categories. Then, we bring Equation 19 into Equation 18 and get

$$\begin{aligned}-\mathbb{E}_{\mathcal{P}\left(\lfloor\tilde{\boldsymbol{h}}\rceil^{k},y\right)}\log\frac{\lfloor\tilde{\boldsymbol{h}}\rceil^{k}\times\mu_{\boldsymbol{y}}}{||\lfloor\tilde{\boldsymbol{h}}\rceil^{k}||\times||\mu_{\boldsymbol{y}}||}&\cong-\mathbb{E}_{\mathcal{P}(\lfloor\boldsymbol{h}\rceil^{k},y)}\log\frac{\lfloor\boldsymbol{h}\rceil^{k}\times\mu_{\boldsymbol{y}}}{||\lfloor\boldsymbol{h}\rceil^{k}||\times||\mu_{\boldsymbol{y}}||}\\ \mathbb{E}_{\mathcal{P}\left(\lfloor\tilde{\boldsymbol{h}}\rceil^{k},y\right)}\log\frac{\lfloor\tilde{\boldsymbol{h}}\rceil^{k}\times\mu_{\boldsymbol{y}}}{||\lfloor\tilde{\boldsymbol{h}}\rceil^{k}||\times||\mu_{\boldsymbol{y}}||}&\cong\mathbb{E}_{\mathcal{P}(\lfloor\boldsymbol{h}\rceil^{k},y)}\log\frac{\lfloor\boldsymbol{h}\rceil^{k}\times\mu_{\boldsymbol{y}}}{||\lfloor\boldsymbol{h}\rceil^{k}||\times||\mu_{\boldsymbol{y}}||}.\end{aligned} \tag{20}$$

To prove Equation 20, we demonstrate an evidence example in Figure 5. As shown in the subfigure (a), the cosine error risk of the target vector $\mu_{\boldsymbol{y}}$ and the candidate vector, e.g., $\lfloor\boldsymbol{h}\rceil^{k}$, $\lfloor\tilde{\boldsymbol{h}}\rceil^{k}$, etc, is insensitive to the specific value of vectors, i.e., the cosine error risk keeps generally consistent as the values of vectors fall or vice versa. Such an assumption is supported by comparing plot (1 with plot (2 or comparing plot 3) with plot 4). Therefore, we derive

$$\log\frac{\lfloor\tilde{\boldsymbol{h}}\rceil^{k}\times\mu_{\boldsymbol{y}}}{||\lfloor\tilde{\boldsymbol{h}}\rceil^{k}||\times||\mu_{\boldsymbol{y}}||}=\log\frac{\lfloor\boldsymbol{h}\rceil^{k}\times\mu_{\boldsymbol{y}}}{||\lfloor\boldsymbol{h}\rceil^{k}||\times||\mu_{\boldsymbol{y}}||} \tag{21}$$

so that

$$\mathbb{E}_{\mathcal{P}\left(\lfloor\tilde{\boldsymbol{h}}\rceil^{k},y\right)}\log\frac{\lfloor\tilde{\boldsymbol{h}}\rceil^{k}\times\mu_{\boldsymbol{y}}}{||\lfloor\tilde{\boldsymbol{h}}\rceil^{k}||\times||\mu_{\boldsymbol{y}}||}\cong\mathbb{E}_{\mathcal{P}(\lfloor\boldsymbol{h}\rceil^{k},y)}\log\frac{\lfloor\boldsymbol{h}\rceil^{k}\times\mu_{\boldsymbol{y}}}{||\lfloor\boldsymbol{h}\rceil^{k}||\times||\mu_{\boldsymbol{y}}||}. \tag{22}$$

A.2.2 Proof for the Inequality Part

To prove

$$\Phi\left(g_{\vartheta}^{cl}\left(\tilde{\boldsymbol{h}}\right)\middle|\boldsymbol{y}\right)\leq\Phi\left(g_{\vartheta}^{cl}(\boldsymbol{h})\middle|\boldsymbol{y}\right) \tag{23}$$

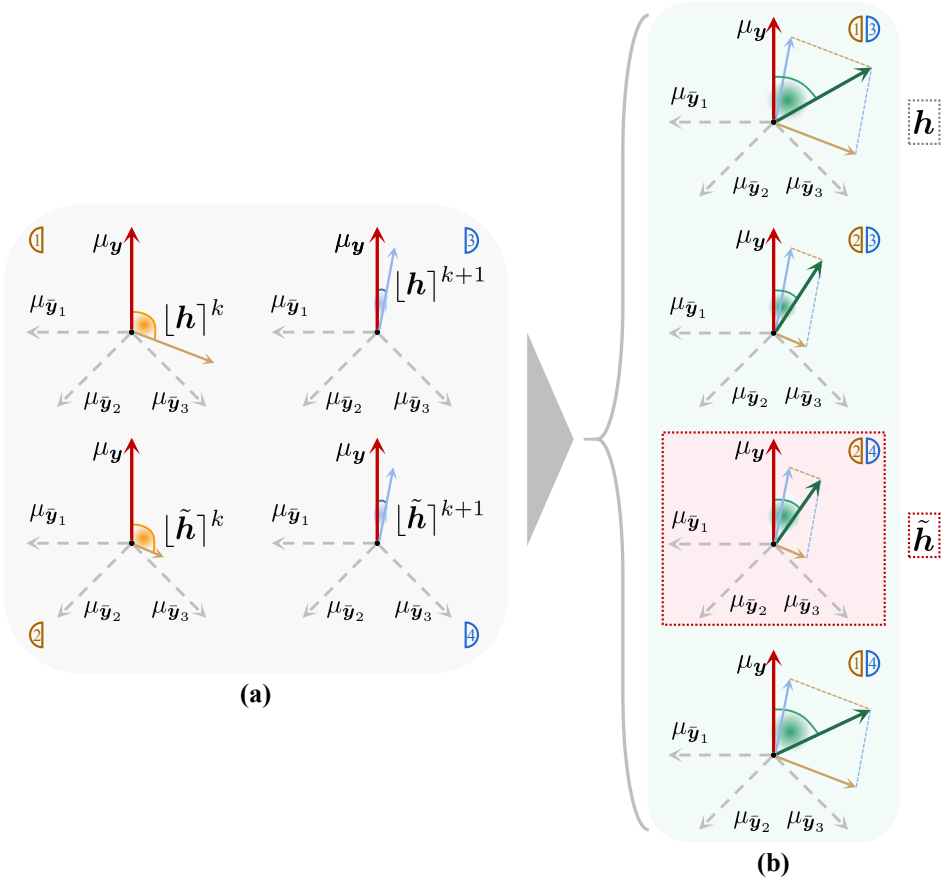


Figure 5: Evidence examples of the proofs for Theorem 5.2. (a) the visual examples of the k-th dimension of the unmasked representation $\boldsymbol{h}$ or the masked representation $\tilde{\boldsymbol{h}}$. (b) the real and hypothetical examples of $\boldsymbol{h}$ or $\tilde{\boldsymbol{h}}$. $\mu_{\boldsymbol{y}}$ denotes the target category's high-dimensional feature, and $\{\mu_{\bar{\boldsymbol{y}}_1}, \mu_{\bar{\boldsymbol{y}}_2}, \mu_{\bar{\boldsymbol{y}}_3}\}$ denote the features of other categories.

we follow the formula transformation of Equation 19 and Equation 20 to derive

$$\begin{aligned}
\mathbb{E}_{\mathcal{P}(y)}\left[\mathbb{E}_{\mathcal{P}\left(g_{\vartheta}^{cl}(\tilde{\boldsymbol{h}})|y\right)}\delta\left(g_{\vartheta}^{cl}\left(\tilde{\boldsymbol{h}}\right), \mathbb{E}_{\mathcal{P}\left(g_{\vartheta}^{cl}(\tilde{\boldsymbol{h}})|y\right)}\right)\right] &\leq \mathbb{E}_{\mathcal{P}(y)}\left[\mathbb{E}_{\mathcal{P}\left(g_{\vartheta}^{cl}(\boldsymbol{h})|y\right)}\delta\left(g_{\vartheta}^{cl}\left(\boldsymbol{h}\right), \mathbb{E}_{\mathcal{P}\left(g_{\vartheta}^{cl}(\boldsymbol{h})|y\right)}\right)\right] \\
-\mathbb{E}_{\mathcal{P}(\tilde{\boldsymbol{h}},y)} \log \frac{\tilde{\boldsymbol{h}} \times \mu_{\boldsymbol{y}}}{||\tilde{\boldsymbol{h}}|| \times ||\mu_{\boldsymbol{y}}||} &\leq -\mathbb{E}_{\mathcal{P}(\boldsymbol{h},y)} \log \frac{\boldsymbol{h} \times \mu_{\boldsymbol{y}}}{||\boldsymbol{h}|| \times ||\mu_{\boldsymbol{y}}||} \\
\mathbb{E}_{\mathcal{P}(\tilde{\boldsymbol{h}},y)} \log \frac{\tilde{\boldsymbol{h}} \times \mu_{\boldsymbol{y}}}{||\tilde{\boldsymbol{h}}|| \times ||\mu_{\boldsymbol{y}}||} &\geq \mathbb{E}_{\mathcal{P}(\boldsymbol{h},y)} \log \frac{\boldsymbol{h} \times \mu_{\boldsymbol{y}}}{||\boldsymbol{h}|| \times ||\mu_{\boldsymbol{y}}||}.
\end{aligned} \tag{24}$$

As shown in the subfigure (a) of Figure 5, all of $\lfloor \boldsymbol{h} \rceil^k$, $\lfloor \tilde{\boldsymbol{h}} \rceil^k$, $\lfloor \boldsymbol{h} \rceil^{k+1}$, and $\lfloor \tilde{\boldsymbol{h}} \rceil^{k+1}$ denote a specific dimension of the corresponding representations. Then, we reconstruct the complete representations from the dimensions, which is achieved in the subfigure (b) of Figure 5. The combinations of (2, 3) and (1, 4) are hypothetical and fake, the combination (1, 3) denotes the original and unmasked representation $\boldsymbol{h}$, and (2, 4) denotes the masked representation $\tilde{\boldsymbol{h}}$. We observe that although the specific value of the vector of each dimension has a thin effect on the corresponding cosine error risk, the combined representation vector is largely sensitive to such values. As demonstrated in (1, 3) and (2, 4) in the subfigure (b), the cosine error risk of $\tilde{\boldsymbol{h}}$ and $\mu_{\boldsymbol{y}}$ is apparently less than $\boldsymbol{h}$ and $\mu_{\boldsymbol{y}}$, which supports that, given a target $\boldsymbol{y}$, the masked representation can indeed derive smaller conditional variance than the unmasked representation. The reason behind such a phenomenon is that, following Theorem 5.1, the self-paced dimensional mask jointly enhances the gradient effect of the dimensions containing discriminative information and reduces that of the dimensions containing the confounder by adjusting the weights of different dimensions with respect to their gradient contribution to the optimization of a specific self-supervised task, e.g., contrastive learning,

during training. Therefore, the dimensions containing discriminative information have relatively large weights, while the dimensions containing confounder information are partially *masked*.

A.2.3 Proof for the Tighter Bounds

Being aware of proofs in Section A.2.1 and Section A.2.2, we confirm the validation of Theorem 5.2. Then, we bring Theorem 5.2 into Theorem 5.1 to derive the comparison of the lower bounds of supervised cross-entropy loss that are separately obtained by the masked representation $\tilde{\boldsymbol{h}}$ and the unmasked representation $\boldsymbol{h}$:

$$\begin{aligned}
&\mathcal{L}_{contrast}\left(g_{\vartheta}^{cl}\left(\tilde{\boldsymbol{h}}\right)\right)-\sqrt{\Phi\left(g_{\vartheta}^{cl}\left(\tilde{\boldsymbol{h}}\right)\Big|\boldsymbol{y}\right)}-\frac{1}{2}\cdot\sum_{k\in[\![1,D]\!]}\sqrt{\Phi\left(\lfloor g_{\vartheta}^{cl}\left(\tilde{\boldsymbol{h}}\right)\rceil^{k}\Big|\boldsymbol{y}\right)}-\mathcal{O}\left(M^{-\frac{1}{2}}\right)\\
&\geq\mathcal{L}_{contrast}\left(g_{\vartheta}^{cl}\left(\boldsymbol{h}\right)\right)-\sqrt{\Phi\left(g_{\vartheta}^{cl}\left(\boldsymbol{h}\right)\Big|\boldsymbol{y}\right)}-\frac{1}{2}\cdot\sum_{k\in[\![1,D]\!]}\sqrt{\Phi\left(\lfloor g_{\vartheta}^{cl}\left(\boldsymbol{h}\right)\rceil^{k}\Big|\boldsymbol{y}\right)}-\mathcal{O}\left(M^{-\frac{1}{2}}\right).
\end{aligned}\tag{25}$$

Therefore, the lower bound obtained by the masked representation, i.e., MetaMask, is larger than the unmasked representation, i.e., typical contrastive learning methods. We further analyze the upper bound of supervised cross-entropy loss on downstream tasks and derive

$$\begin{aligned}
&\mathcal{L}_{contrast}\left(g_{\vartheta}^{cl}\left(\tilde{\boldsymbol{h}}\right)\right)+\sqrt{\Phi\left(g_{\vartheta}^{cl}\left(\tilde{\boldsymbol{h}}\right)\Big|\boldsymbol{y}\right)}+\mathcal{O}\left(M^{-\frac{1}{2}}\right)\\
&\leq\mathcal{L}_{contrast}\left(g_{\vartheta}^{cl}\left(\boldsymbol{h}\right)\right)+\sqrt{\Phi\left(g_{\vartheta}^{cl}\left(\boldsymbol{h}\right)\Big|\boldsymbol{y}\right)}+\mathcal{O}\left(M^{-\frac{1}{2}}\right),
\end{aligned}\tag{26}$$

where the upper bound obtained by MetaMask is smaller than the typical contrastive learning methods. Concretely, we conclude that our approach can better bound the downstream classification risk, i.e., the upper and lower bounds of supervised cross-entropy loss obtained by MetaMask are tighter than typical contrastive learning methods.

A.3 Discussion on the Difference between Dimensional Confounder and Curse of Dimensionality

The restricted definition of the curse of dimensionality is that given a fixed number of examples, as the dimensionality of the representation increases during training, the performance of the model on downstream tasks instead degenerates, e.g., the *Hughes phenomenon*. Recent researchers state a generalized definition of the curse of dimensionality from the perspective of over-fitting: given a fixed number of examples, the performance of the model degrades as the amount of *computation* in the network increases, e.g., the increase of representations' dimensionality or the increase of the layers of a deep network. Concretely, the curse of dimensionality is a theory to explain a specific phenomenon of over-fitting from the perspective of sample size.

However, our *dimensional confounder* is defined as a negative factor that may lead to model degradation, which is proposed from the dimensional perspective. The motivating example in Figure 2 demonstrates that even if the deep network architecture and the dimension of the representation are consistent, the dimensional confounder still causes the performance of the model to degenerate. This experimental analysis proves that the proposed dimensional confounder and the curse of dimensionality are two different terminologies. Further, to conduct the experiments in Figure 1, we fix the dimensionality of the representation and only increase the dimensionality of the features generated by the projection head, which is under the restriction of the redundancy-reduction technique [14]. The intuition behind such a behavior is that as the dimensions of the projected features increase, the information entropy contained in each dimension of the representation becomes more decoupled, and the total amount of task-relevant information entropy is limited, such that the dimensional confounder (i.e., the dimensions only containing task-irrelevant information) is gradually increasing. The experimental results demonstrate that as the dimensional confounder in the learned representation increases, the performance of the model is weakened. Such a phenomenon is consistent with the generalized definition of the curse of dimensionality, since the dimensionality of the representation is constant while the parameter (computation) of the deep network is increasing. Therefore, we conclude that the dimensional confounder can lead to the curse of dimensionality, and compared with the curse

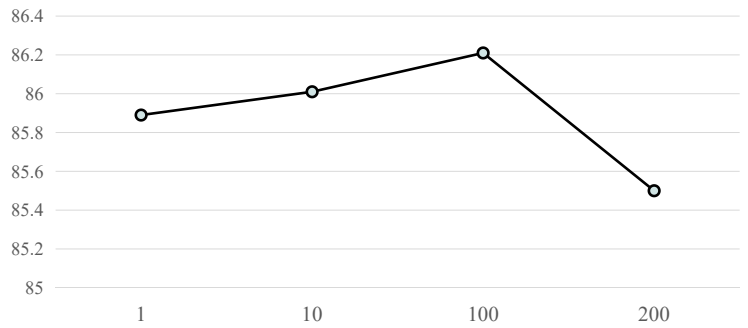


Figure 6: Comparisons of MetaMask using different settings of the hyper-parameter α.

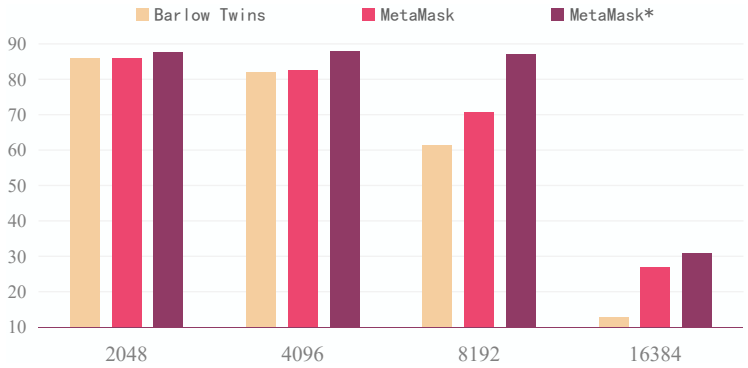


Figure 7: Further comparisons of varying the projection head's dimensionality from the range of $\{2048, 4096, 8192, 16384\}$, which are conducted on CIFAR-10 by using ResNet-18. $*$ denotes MetaMask using a trick of fixed learning rate instead of the cosine annealing strategy.

of dimensionality, the proposed dimensional confounder is a generalized issue. The comparisons in Table 1 and Table 2 prove that our MetaMask can alleviate the dimensional confounder issue, and the experiments in Figure 4 further support the superiority of MetaMask over current self-supervised methods to alleviate the curse of dimensionality issue.

A.4 Extended Experiments

We conduct several experiments to study the intrinsic property of the proposed MetaMask.

A.4.1 The Impact of the Hyper-Parameter α

As demonstrated in Figure 6, we conduct comparisons of MetaMask using ResNet-18 on CIFAR-10. We follow the experimental principle in Section 6 and use KNN prediction as the evaluation approach. The comparison results show that an elaborate assignment of α can indeed improve the performance of MetaMask, and when $\alpha = 100$, the accuracy of MetaMask achieves the peak value. The reasons behind this curve are as follows: 1) $\mathcal{L}_{drr}$ is naturally excessively larger than $\mathcal{L}_{contrast}$ so that when α is not large enough, $\mathcal{L}_{contrast}$'s impact with respect to the gradient is diminished; 2) when α is over-large, the impact of $\mathcal{L}_{drr}$ is weakened, and the dimensional redundancy issue cannot be sufficiently addressed so that the representation may be collapsed to a trivial constant.

A.4.2 Further Exploration of the Robustness of MetaMask

The experimental analysis demonstrated in Section 6.2 proves that MetaMask can indeed alleviate the negative impact caused by the dimensional confounder. However, MetaMask* achieves better performance than compared methods, including vanilla MetaMask. We provide a further exploration shown in Figure 7, and we observe that MetaMask* collapses abruptly at larger dimensional settings, e.g., 16384, which demonstrates our explanation of such a phenomenon in Section 6.2, i.e., the fixed learning rate trick can temporarily improve the performance of MetaMask, but this improvement is inconsistent. The results in Figure 7 further show that with or without this trick, the proposed MetaMask is robust against the dimensional confounder.

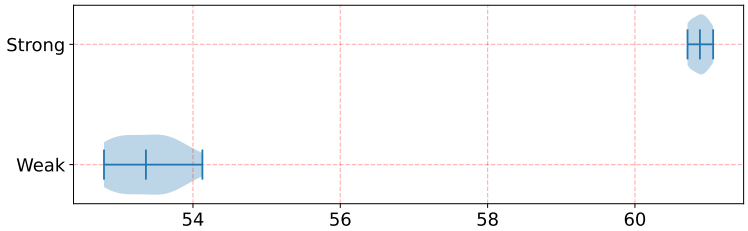


Figure 8: Comparisons of BYOL using different backbones on CIFAR100 with 5% random dimensional mask. "Weak" denotes the weak encoder *conv*, and "Strong" denotes the strong encoder *ResNet-18*.

Table 3: The complexity comparisons between MetaMask and benchmark methods on CIFAR10. Note that for fair comparisons, this experiment is based on 1 GPU of NVIDIA Tesla V100.

Methods	Parameters	Training time cost for an epoch
ResNet-18	11.2M	-
SimCLR	13M	70s
Barlow Twins	22.7M	80s
SimCLR + Barlow Twins	24.6M	85s
MetaMask	24.6M + 512	210s

A.4.3 Impacts of Encoder's Ability towards MetaMask

From the experimental results reported in Section 6, we observe that MetaMask improves the benchmark methods with weak encoders by excessively significant margins, but the improvement with strong encoders is relatively limited. Our consideration behind this phenomenon is that although representations learned by both weak and strong encoders (without our method) may contain dimensional confounders, the strong encoder can better capture semantic information so that the dimensional confounder of the learned representation is naturally less, and the useful discriminative information is much more than the representation learned by the weak encoder. [41] provide the theorem and corresponding proof to demonstrate that the contrastive loss can bound the cross-entropy loss in downstream tasks. Strong encoders better minimize the contrastive loss, and thus the representations learned by strong encoders contain more semantic information, and accordingly, fewer dimensional confounders. However, from the results reported in Table 2, our method can still improve the benchmark methods.

We further conduct experiments by imposing random dimensional masks on the learned representations for weak and strong encoders. The results are reported in Figure 8. The comparisons demonstrate that under the consistent setting of 5% random dimensional mask rate, the results of the weak encoder (conv) range from 52.79 to 54.13, and the results of the strong encoder (ResNet-18) range from 60.71 to 61.06. Note that we conduct 10 trials for each experiment to achieve unbiased results. The quality of the strong encoder's representation is apparently better, and the performance of the strong encoder is more stable (the variance of the strong encoder's results is smaller), which proves that the representation learned by the strong encoder contains fewer dimensional confounders, and each dimension contains more discriminative information. Therefore, our method improves weak encoders more than strong encoders.

A.4.4 Training Complexity of MetaMask

To compare the training complexity of MetaMask and benchmark methods, we conduct experiments on CIFAR10 by using ResNet-18 as the backbone network. The results are demonstrated in Table 3, which shows that the parameter number used by MetaMask is close to the ablation model, i.e., SimCLR + Barlow Twins, and Barlow Twins. Compared with Barlow Twins, SimCLR + Barlow Twins only adds a MLP with several layers as the projection head for the contrasting learning of SimCLR. For MetaMask, we only add a learnable dimensional mask $\mathcal{M}$ in the network of MetaMask. Note that to decrease the time complexity of MetaMask, we create an additional parameter space to save the temporary parameters for computing second-derivatives, but such cloned parameters are not included in the computation of the network during training.

Table 4: The comparisons between MetaMask and benchmark methods on the CIFAR10 dataset by using the same total time costs. Note that for fair comparisons, this experiment is based on 1 GPU of NVIDIA Tesla V100.

Methods	Epoch	Training time cost	Accuracy
SimCLR	2400	46h	81.75
Barlow Twins	2100	46h	85.71
SimCLR + Barlow Twins	2000	47h	85.79
MetaMask	800	46h	86.01

Table 5: The comparisons of Barlow Twins using different dropout ratios on CIFAR10 with ResNet-18.

Dropout ratio	Dropout shared between views	Accuracy
0	No	85.72
0.01	No	85.76
0.05	No	86.11
0.1	No	85.72
0.2	No	85.13
0.5	No	81.54
0.05	Yes	31.17
0.1	Yes	37.15
0.5	Yes	31.62

For the time complexity, due to the learning paradigm of meta-learning (second-derivatives technique), MetaMask has larger time complexity than benchmark methods (including the ablation model SimCLR + Barlow Twins). We further conduct experiments to evaluate the performance of the compared methods using similar total training time costs. The results are reported in Table 4, which demonstrates that MetaMask still achieves the best performance, and MetaMask can also beat the ablation model SimCLR + Barlow Twins. This proves that although the time complexity of MetaMask is a little bit high, the improvement of MetaMask is consistent and solid. We also include these results and the corresponding analysis in Appendix of the rebuttal revised version.

A.4.5 Exploration of Barlow Twins using the Dropout Trick

We apply dropout to randomly set the features generated by the backbone network to 0 with a given probability. We use two different settings: 1) features from different views are randomly set to 0 independently with 6 probabilities, including 0.01, 0.05, 0.1, 0.2, and 0.5, which is shown as the setting of dropout shared between views is "No"; 2) for the same sample, we set the same channels of the features from different views to 0. In detail, we let the features from the first view pass the dropout layer, and then record the channels which are set to 0. Finally, we set the same feature channels of the second view to 0 and multiply the features with 1/(1-p), where "p" refers to the probability of dropout. In this case, we use three probabilities: 0.05, 0.1, and 0.5. This setting is shown as "Yes" for the dropout shared between views. The results are reported in Table 5. We observe that models trained by following the second setting are collapsed, and these models far underperform MetaMask. For the first setting, the performance trend is in accordance with our expectation, using several well-chosen dropout ratios can improve the performance of Barlow Twins to a certain extent, but the improvement is inconsistent. Furthermore, even the best model of Barlow Twins using the dropout trick cannot achieve comparable performance to MetaMask, since as shown in Table 2, MetaMask achieves 87.53 on CIFAR10 with ResNet-18.

A.4.6 Performing MetaMask on ImageNet with ResNet-50

Training one trial of MetaMask on ImageNet is excessively time-consuming for the adopted server. It is hard to impose sufficient hyperparameter tuning experiments, because tuning parameters on the validation set and then retraining is time-consuming. For the current version, we principally adopt the hyperparameter of MetaMask on ImageNet-200 with ResNet-18 for the experiments on ImageNet with ResNet-50. We report the results on Table 6, which demonstrates that MetaMask achieves the

Table 6: Top-1 and top-5 accuracies (%) under linear evaluation on ImageNet with ResNet-50 encoder. MetaMask is performed based on Barlow Twins and SimCLR. Top-3 best methods are underlined.

Method	Accuracy	
	Top-1	Top-5
Supervised	76.5	-
MoCo	60.6	-
SimCLR	69.3	89.0
SimSiam	71.3	-
BYOL	74.3	91.6
Barlow Twins	73.2	91.0
MetaMask	73.9	91.4

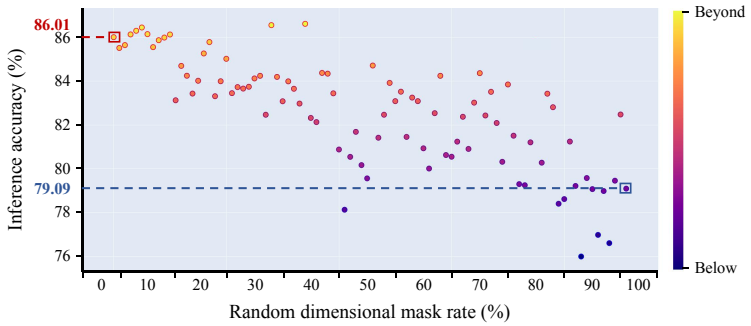


Figure 9: Experimental scatter diagrams obtained by MetaMask with randomly masked dimensions on CIFAR-10 with ResNet-18. We collect the final dimensional weight matrix $\mathcal{M}$ and then choose dimensions with weights below average as the masked dimensions. These dimensions are considered to be associated with the dimensional confounders, and we impose random dimensional masks on these dimensions. For fair comparisons, 10 trials are conducted per mask rate, except for 0% and 100% mask rates, and every single point denotes a trial. The trial boxed by *red lines* presents the performance achieved by the unmasked representation of MetaMask, and the trial boxed by *blue lines* presents the performance achieved by MetaMask's representation where all dimensions containing confounders are masked.

top-3 best performance. Note that the major results are reported by [14], and the implementation of MetaMask is based on SimCLR and Barlow Twins. For comparisons with the main baselines, MetaMask beats SimCLR by 4.6 on top-1 accuracy and 2.4 on top-5 accuracy, and MetaMask beats Barlow Twins by 0.7 on top-1 accuracy and 0.4 on top-5 accuracy. The improvement of MetaMask is in accordance with our observation in Appendix A.4.3. Furthermore, ImageNet-200 is a truncated dataset of ImageNet, so the domain shift between these datasets is slight. The comparisons in Table 1, and Table 2 demonstrate that the proposed MetaMask can still improve the benchmark methods in large-scale datasets.

A.4.7 Exploration of Confounders Contained by Masked Dimensions

MetaMask trains $\mathcal{M}$ by adopting a meta-learning-based training approach, which ensures that $\mathcal{M}$ can partially mask the "gradient contributions" of dimensions containing task-irrelevant information and further promote the encoder to focus on learning task-relevant information. So, MetaMask only performs the gradient mask during training instead physically masking dimensions in the test. We provide theoretical explanation and proofs in Appendix A.1 and Appendix A.2. The reason behind our behavior is that even dimensions that contain dimensional confounders are also possible to contain discriminative information so that lowering the gradient contribution of such dimensions can not only prevent the over-interference of the dimensional confounders on the representation learning but also preserve the acquisition of the information of these dimensions. Accordingly, the foundational idea behind self-supervised learning is to learn "general" representation that can be generalized to various tasks. MetaMask introduces a meta-learning-based approach to train the dimensional mask $\mathcal{M}$ with respect to improving the performance of contrastive learning. However, the theorems,

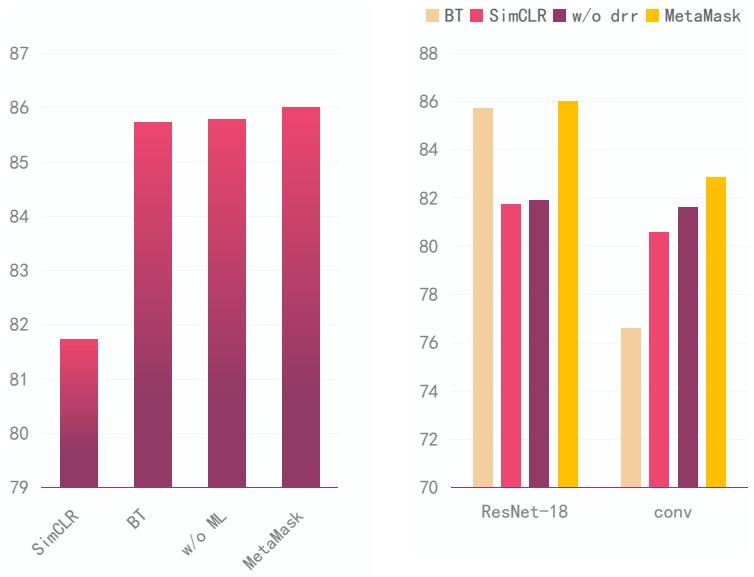
proposed by [41], only prove that the contrastive learning objective is associated with the downstream classification task, while there is no theoretical evidence to demonstrate the connection between contrastive learning objective and other downstream tasks. Therefore, we consider not directly masking the dimensions containing dimensional confounders in the test.

Furthermore, we conduct experiments to explore the performance of the variant that directly masks these dimensions in the test, which is demonstrated in Figure 9. For the exploration of our masking scheme and its variants, we conduct experiments as follows: after training, we collect the final dimensional weight matrix $\mathcal{M}$ and then choose dimensions with weights below average as the masked dimensions. These dimensions are considered to be associated with dimensional confounders. To prove whether these dimensions have confounders, we perform random dimensional masking to these dimensions, and when the masking rate is 100%, the model turns to the variant that directly masks all these dimensions in the test. The experiments are based on SimCLR + MetaMask. Note that we conduct 10 trials per mask rate (except for 0% and 100% mask rates) for fair comparisons. We observe that the original MetaMask (i.e., mask rate is 0%) achieves the best performance on average, and MetaMask outperforms the variant masking the dimensions with confounders by a significant margin, which proves that our proposed approach, i.e., masking the "gradient contributions" of dimensions in the training is more effective than the compared approach, i.e., directly masking dimensions in the test. While, several trials with specific mask rates demonstrate better performance than MetaMask, which proves that the dimensions filtered by MetaMask indeed contain dimensional confounders. Additionally, observing the results reported in Figure 2 and the results in Figure 9, we find that the results achieved by the proposed variants are better than Barlow Twins with random dimensional masks on average, which can further prove the filtered dimensions containing confounders and that MetaMask indeed assigns lower gradient weights to the dimensions containing confounders.

Additionally, we conduct experiments to directly impose a linear probe on the whole masked or unmasked features. Note that the definition of masked features is mentioned above, and the experiments are conducted on CIFAR10 with ResNet-18. As shown in Figure 9, a linear probe is solely imposed on the whole "unmasked" features (without "masked" features), which is the same as the 100% mask rate variant for the "masked" features above, and the result is 79.09. We provide the corresponding reasons and analyses above. Additionally, such results may be due to the masking scheme for these experiments, i.e., collecting the final dimensional weight matrix $\mathcal{M}$ and then masking the dimensions with weights below average, which is dramatically different from MetaMask's behavior, and this scheme is only used for this exploration. The result achieved by a linear probe on the whole "masked" features is 56.30, which demonstrates our statement that dimensions containing dimensional confounders are also possible to contain discriminative information, because the achieved accuracy is not under 10. The result also proves that MetaMask can indeed assign lower gradient weights to the dimensions containing confounders, since such a model far underperforms MetaMask (86.01).

A.4.8 Further Ablation Study of MetaMask

Barlow Twins handles dimensional redundancy but suffers dimensional confounders. MetaMask mitigates dimensional confounders by learning and applying a dimensional mask. We conduct experiments to demonstrate our statement, and the results are reported in Figure 10. For the experiments shown in Figure 10 (a), we demonstrate the effectiveness of the proposed dimensional mask $\mathcal{M}$ and the corresponding meta-learning-based training paradigm by directly removing such approaches. The results show that the sole w/o ML only improves Barlow Twins by 0.04, while MetaMask can improve BT by 0.29, which proves the proposed $\mathcal{M}$ is pivotal to the performance promotion. In Figure 10 (b), to verify whether there would be performance gain only from alleviating dimensional confounders without $\mathcal{L}_{drr}$, we evaluate the performance of w/o drr and MetaMask. We observe that for the experiments with ResNet-18, w/o drr (without Barlow Twins) improves SimCLR by 0.2 but cannot reach the performance of BT, and MetaMask can improve both SimCLR and BT by 4.28 and 0.29, respectively. For the experiments with conv, w/o drr improves SimCLR by 1.03 and also outperforms BT by 4.08, and MetaMask improves both SimCLR and BT by 2.27 and 6.22, respectively. Concretely, the performance of w/o drr is related to the performance of SimCLR, and it can always improve SimCLR but underperform the complete MetaMask. MetaMask has consistent best performance by using different encoders.



(a) (b)

Figure 10: Ablation studies obtained by MetaMask on CIFAR-10 with ResNet-18 and conv encoders, where "BT" denotes Barlow Twins, "w/o ML" denotes the ablation model without the proposed $\mathcal{M}$ and the corresponding meta-learning-based learning paradigm, i.e., SimCLR + Barlow Twins, and "w/o drr" denotes the ablation model without the dimensional redundancy reduction loss $\mathcal{L}_{drr}$. (a) We conduct experiments to prove the effectiveness of the proposed dimensional mask $\mathcal{M}$ on CIFAR10 with ResNet-18, and the results show that only w/o ML (SimCLR + Barlow Twins) can not improve the performance of the model by a significant margin, and the proposed approach is crucial to the performance improvement. (b) We conduct experiments on CIFAR10 with ResNet-18 and conv, respectively. The results further prove the effectiveness of the proposed $\mathcal{M}$ and $\mathcal{L}_{drr}$.

We consider that $\mathcal{L}_{drr}$ (Barlow Twins loss) could exacerbate dimensional confounders, but as our discussion in Appendix A.4.7, i.e., dimensions containing confounders are also possible to contain discriminative information, more dimensions with confounders due to $\mathcal{L}_{drr}$ may also carry more discriminative information. Likewise, the model without $\mathcal{L}_{drr}$ may generate representations with over-redundant dimensions so that the total amount of available discriminative information will decrease. However, roughly using the representations with complex dimensional information (without $\mathcal{L}_{drr}$) may result in insufficient discriminative information mining, e.g., w/o ML can only outperform SimCLR by a limited margin. Our proposed MetaMask effectively avoids the appearance of such an undesired phenomenon by leveraging $\mathcal{M}$ and the corresponding meta-learning-based training paradigm, which is supported by the empirical results.