Self-Supervised Representation Learning with Meta Comprehensive Regularization

Huijie Guo[1,*], **Ying Ba**[4,5,*], **Jie Hu**[6], **Lingyu Si**[2,3], **Wenwen Qiang**[2,3,†], **Lei Shi**[1,‡]

[1]Beihang University
[2]Institute of Software Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences
[4]Gaoling School of Artificial Intelligence, Renmin University of China
[5]Beijing Key Laboratory of Big Data Management and Analysis Methods
[6]Meituan
{guo_hj, leishi}@buaa.edu.cn, yingba@ruc.edu.cn, hujie@ios.ac.cn, {lingyu, qiangwenwen}@iscas.ac.cn

Abstract

Self-Supervised Learning (SSL) methods harness the concept of semantic invariance by utilizing data augmentation strategies to produce similar representations for different deformations of the same input. Essentially, the model captures the shared information among multiple augmented views of samples, while disregarding the non-shared information that may be beneficial for downstream tasks. To address this issue, we introduce a module called CompMod with Meta Comprehensive Regularization (MCR), embedded into existing self-supervised frameworks, to make the learned representations more comprehensive. Specifically, we update our proposed model through a bi-level optimization mechanism, enabling it to capture comprehensive features. Additionally, guided by the constrained extraction of features using maximum entropy coding, the self-supervised learning model learns more comprehensive features on top of learning consistent features. In addition, we provide theoretical support for our proposed method from information theory and causal counterfactual perspective. Experimental results show that our method achieves significant improvement in classification, object detection and instance segmentation tasks on multiple benchmark datasets.

Introduction

Deep learning models have exhibited remarkable capabilities, leading to the widespread adoption of machine learning across diverse fields. Despite the impressive performance of supervised learning methods, their heavy reliance on labeled data for model training poses limitations on their generalization ability and scalability. To address this challenge, Self-Supervised Learning (SSL) has emerged as a promising paradigm that bridges the gap between supervised and unsupervised learning by generating supervised signals directly from the samples without the need for manual annotation. Currently, SSL has achieved remarkable results in computer vision (Tian, Krishnan, and Isola 2020; Chen, Xie, and He 2021; Caron et al. 2020) and natural language processing (Baevski et al. 2020; Akbari et al. 2021; Zhou et al. 2020).

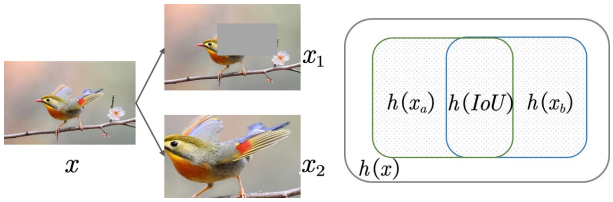

Figure 1: Loss of task-related information caused by data augmentation in SSL methods. (a), the positive sample pair (x_a, x_b) can be obtained from the input x by Random Cropping and Cutout. (b) formally presents the semantics related to label in different augmented views, where $h(\cdot)$ represents the amount of attributes related to the label in sample.

The general framework of self-supervised representation learning consists of two key components: data augmentation and loss function, which try to learn invariance to the transformation generated by data augmentation on the same sample while maintain discrimination to different samples. In practice, data augmentation generates two augmented views of the same image by applying random strategies, such as Cutout (DeVries and Taylor 2017), Coloring (Zhang, Isola, and Efros 2016), Random Cropping (Takahashi, Matsubara, and Uehara 2019), etc. Several studies (Zheng et al. 2021; Shorten and Khoshgoftaar 2019; Zhang and Ma 2022; Tian et al. 2020) also have suggested that not all data augmentations are beneficial for downstream tasks. For instance, rotation invariance may help some flower categories but harm animal recognition (Xiao et al. 2020). Similarly, color invariance may have opposite effects on animal and flower classification tasks. Therefore, recent works have proposed adaptive augmentation strategies to adapt to different data and task environments (Li et al. 2022a; Yang et al. 2022).

Data augmentation strategies are widely used in SSL to create positive pairs of images that share the same label. However, these strategies may not preserve all the semantic information that is relevant to the label in the augmented views. For example, suppose an image's label is "bird" and

[*]These authors contributed equally.
[†]Corresponding author.
[‡]Co-corresponding author.

it only refers to the foreground object, not the background. Figure 1(a) shows two views of the same image x created by Random Cropping and Cutout, denoted as x_1 and x_2. Note that x_1 contains the bird's beak while x_2 does not, and x_2 contains the bird's wings while x_1 does not. A common assumption in SSL is that the semantic content of an image should be invariant to the applied transformations. However, this assumption can be broken by the transformation methods and may not hold for all label-related attributes, such as the bird's beak and wings. Figure 1(b) illustrates a function $h(\cdot)$ that measures the amount of label-related attributes in an image. The representations learned by SSL methods are based on the shared information between different augmented views, such as the Intersection over Union (IoU). However, this shared information may not capture the entire foreground of the input, and some label-related attributes may be dropped in the model training process. The more label-related information is preserved in the training process, the better the model can learn. Therefore, models trained using traditional SSL methods may exhibit subpar performance in downstream tasks due to the loss of label-related information during the training process.

To address the aforementioned issue, we propose utilizing a more comprehensive representation to guide the training of SSL model, enabling the model to focus on non-shared semantic information that might be beneficial for downstream tasks, thereby enhancing model's generalization capability. We propose a plug-and-play module called CompMod with Meta Comprehensive Regularization to guide the learning of SSL methods by obtaining comprehensive features. Specifically, we employ semantic complementarity to fuse augmented features in a low-dimensional space, utilizing a bi-level optimization mechanism to obtain comprehensive representation that guide the learning of SSL methods. Our contributions are the following:

- From the information theory, we analyze that data augmentation in SSL may lead to the lack of task-related information, which in turn reduces the generalization ability of the model.
- We design a plug-and-play module, called CompMod, to induce existing SSL methods to learn comprehensive feature representations. CompMod ensures comprehensive feature exploration through a bi-level optimization mechanism and constrained extraction of features with maximum entropy coding, guaranteeing complete mining of feature completeness.
- A causal counterfactual analysis provides theoretical support for our proposed method. Empirical evaluations of the proposed method substantiate its superior performance in classification, object detection and instance segmentation tasks.

Related Work

Recently, various frameworks have emerged for self-supervised representation learning, which can be broadly classified into two types (Garrido et al. 2022; Balestriero and LeCun 2022): sample-based and dimension-based contrastive learning methods.

Sample-based contrastive methods learn visual representations by constructing pairs of samples and applying contrastive loss function. These methods encourage the embeddings of augmented views of the same image to be close to each other, while simultaneously pushing away the embeddings of different images. Some notable methods, such as SimCLR (Chen et al. 2020), utilize InfoNCE as the loss function and rely on the quality and quantity of negative samples. However, these methods also necessitate greater computational resources. MoCo (He et al. 2020) tackles this issue by constructing a dynamic dictionary bank that expands the pool of available negative samples. On the other hand, some studies have investigated whether SSL can still work without negative samples. BYOL (Grill et al. 2020) and SimSiam (Chen and He 2021) utilize a distillation-like mechanism to learn representations by computing the similarity between positives, without the need for negative samples. Dimension-based contrastive methods learn visual representations by optimizing the information content of the learned representations and reducing feature redundancy. Barlow Twins (Zbontar et al. 2021) endeavors to make the normalized cross-correlation matrix of the augmented embeddings close to the identity matrix. The loss function of VICReg (Bardes, Ponce, and Lecun 2022) consists of three items: invariance, variance and covariance regularization item. TCR (Li et al. 2022b) employs the Maximum Coding Rate Reduction (MCR^2) objective to learn feature subspaces that are both informative and discriminative. Liu (Liu et al. 2022) proposed using maximum entropy coding for contrastive learning, based on the principle of maximum entropy in information theory, and established a connection between sample-based and dimension-based SSL.

These works mentioned above are based on the invariance of semantic among augmented views, while ignoring the partial loss of label-related information in each view after augmentation, leading to imperfect consistency in semantic information across views. By leveraging the comprehensive information between views, our work allows the feature extractor to gather more abundant information, thereby inducing the learned sample representations to be more generalizable. Our proposed Meta Comprehensive Regularization can be integrated into existing SSL framework.

Methodology

Figure 2 shows the overview of our proposed method. We design a new module, CompMod, to improve existing self-supervised method. Next, we first theoretically analyze the lack of partial semantic information caused by data augmentation is not conducive to downstream tasks in SSL from an information-theoretic perspective, and then introduce our proposed method and the training process of the model.

Contrastive Learning

Let $D = \{x_i\}_{i=1}^n$ denote the unlabeled training set, where x_i is an input image. Two augmented views x_i^1 and x_i^2 of the sample x_i are generated by different augmentation strategies t^1 and t^2 sampled from a augmentation distribution A. The augmented views are fed into a shared encoder f_θ to obtain their representations $h_i^1 = f_\theta(x_i^1)$ and

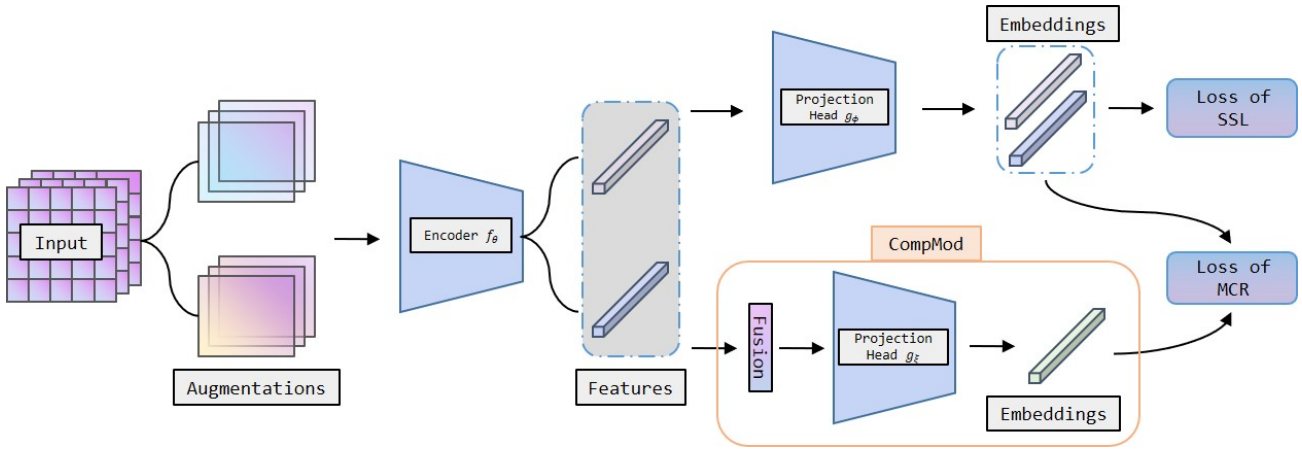

Figure 2: Illustration of self-supervised representation learning framework with Meta Comprehensive Regularization.

$h_i^2 = f_\theta(x_i^2)$, which are then mapped via a projector g_ϕ onto the embedding space, $z_i^1 = g_\phi(h_i^1)$ and $z_i^2 = g_\phi(h_i^2)$. We denote the embedding matrix of augmentation view 1 as $Z_1 = [z_1^1, ..., z_i^1, ..., z_n^1]^T \in R^{n \times d}$, where d is the dimension of the embedding space, so does matrix Z_2. Represented by SimCLR, the objective function of the Contrastive Learning (CL) employs the Noise Contrastive Estimation (NCE) loss (Gutmann and Hyvärinen 2010):

$$\mathcal{L}_{ssl} = \mathbb{E}_{z_i^1, z_i^2}[-\log \frac{e^{s(z_i^1, z_i^2)/\tau}}{e^{s(z_i^1, z_i^2)/\tau} + \sum_{z_j} e^{s(z_i^1, z_j)/\tau}}] \quad (1)$$

where $s(\cdot, \cdot)$ denotes the cosine similarity and τ represents the temperature hyper-parameter, z_j is the negative sample, $z_j \in \mathcal{Z}_1 \cup \mathcal{Z}_2 / \{z_i^1, z_i^2\}$, $\mathcal{Z}_1$ and $\mathcal{Z}_2$ represent the sets of augmented views in the embedding space, respectively.

Analysis Based on Information Theory

We assume that the original input images inherently encompass all relevant task-related information, e.g., $I(x; T) = H(T)$, where $x \sim D$ is a random variable, I denotes the mutual information, H represents the information entropy, and T refers to a random variable for the downstream task.

As evident from Figure 1, data augmentation on the input sample results in loss of task-relevant information within the data. Consequently, we deduce: $I(x_1; T), I(x_2; T) \leq H(T)$, where $(x_1, x_2) \sim \{(x_i^1, x_i^2)\}_{i=1}^n$. Also, we can obtain: $I(x_1; x_2; T) \leq H(T)$. A general explanation for CL is to maximize the mutual information between two augmented views(Wang et al. 2022):

$$\max_{f,g} \quad I(z_1; z_2) \quad (2)$$

where z_1 and z_2 are random variables, $(z_1, z_2) \sim (\mathcal{Z}_1, \mathcal{Z}_2)$. Applying Data Processing Inequality (Klir and Wierman 1999) to the Markov chain $x \to x_1(x_2) \to z_1(z_2)$, we have:

$$\begin{aligned} H(x) \geq I(x_1; x_2) \geq I(z_1; z_2) \\ I(x; T) \geq I(x_1; x_2; T) \geq I(z_1; z_2; T) \end{aligned} \quad (3)$$

Based on Eq. 2 and Eq. 3, we can draw the conclusion that due to the disruption caused by data augmentation to the semantic information of input samples, contrastive learning is constrained to extract only a subset of task-related information. In order to elucidate this conclusion, we begin by providing a definition for comprehensive representation, followed by deriving the following theorem.

Definition 1. *(Comprehensive representation) For a random variable z defined in encoder space. z is a comprehensive representation if and only if $I(z; T) = H(T)$.*

Theorem 1. *(Task-Relevant information in representations) In contrastive learning, given a random variable x representing the original sample space, two random variables x_1 and x_2 characterizing the sample space after augmentation, and two random variable z_1 and z_2 denoting the augmented samples within the feature space, we have:*

$$\begin{aligned} H(T) \geq \{I(x_1; T), I(x_2; T)\} \geq I(x_1; x_2; T) \\ H(T) \geq \{I(z_1; T), I(z_2; T)\} \geq I(z_1; z_2; T) \end{aligned} \quad (4)$$

From Theorem 1, we deduce the following: (1) Data augmentation leads to a reduction in the amount of task-related information present within the data. (2) The task-related information contained in the commonalities between z_1 and z_2 is individually less than the task-related information within z_1 and z_2. Obviously, if the representation of the augmented view is close to the comprehensive representation, it is more beneficial to the downstream tasks. Next, we propose to use Meta Comprehensive Regularization to force the augmented representation to be a comprehensive representation.

Meta Comprehensive Regularization

To address the issue of semantic loss resulting from data augmentation, we propose a new module called CompMod, which learns a more comprehensive representation and helps facilitate model learning, as shown in Figure 2. This module can be directly incorporated into the traditional SSL framework and can complement existing SSL methods. Then, we introduce the module CompMod.

Different augmentations of the same sample are typically derived through various data augmentation techniques, empirically implying that each augmentation results in distinct

semantic information loss. As a result, a method that yields a comprehensive feature representation without semantic information loss, distinct from the original input data representation, involves integrating features from different perspectives of the same sample. Thus, a more comprehensive representation of x_i can be obtained by the following formula:

$$\hat{h}_i := h_i^1 \oplus h_i^2 \tag{5}$$

where $\oplus$ represents the fusion strategy, such as the concatenation of vectors: $\hat{h}_i = \left[h_i^1\ h_i^2\right] \in R^{2d'}$, where d' is the output dimension of a backbone network. $\hat{h}_i$ fuses the semantic information from all augmented views, so as to solve the problem of partial semantic missing. Next, a projection head g_ξ parameterized by ξ maps $\hat{h}_i$ onto the same embedding space as z_i^1 and z_i^2. And then, we obtain the so-called more comprehensive embedding of sample x_i, denoted as $\hat{z}_i = g_\xi(\hat{h}_i) \in R^d$. The comprehensive embedding matrix is defined as $\hat{Z} = [\hat{z}_1, ..., \hat{z}_i, ..., \hat{z}_n]^T \in R^{n \times d}$.

Simple fusion alone does not guarantee that the learned features encompass all semantics. Inspired by the maximum entropy principle in information theory, a generalizable representation should be the one with the maximum entropy among all possible representations, corresponding to the maximization of the semantic information associated with the true label. Here, we use the code length of the lossy data coding (Cover 1999) to calculate the entropy of the embedding matrix $\hat{Z}$, which apply the Taylor series expansion:

$$\mathcal{L}_{comp}(\hat{Z}) = -\mathrm{Tr}(\mu \sum_{k=1}^{m} \frac{(-1)^{k+1}}{k} (\lambda \hat{Z}\hat{Z}^T)^k) \tag{6}$$

where k is the order of Taylor expansion, μ and λ are hyperparameters. Therefore, to further ensure the comprehensiveness of the obtained $\hat{Z}$, we propose that the obtained $\hat{Z}$ should minimize $\mathcal{L}_{comp}(\hat{Z})$.

Next, we use the comprehensive representation to guide the learning of the backbone network. To enhance the semantic richness of the representation in each augmented view, we constrain the information contained in the augmented embeddings Z_1, Z_2 to equal the information contained in the comprehensive embedding $\hat{Z}$. We propose to minimize the following loss:

$$\mathcal{L}_{mcr}(Z_1, Z_2) = -\sum_{t=1}^{2} \mathrm{Tr}(\mu \sum_{k=1}^{m} \frac{(-1)^{k+1}}{k} (\lambda \hat{Z} Z_t^T)^k) \tag{7}$$

where Z_t is the embedding matrix of augmented view, $t = 1, 2$. As we can see, when $\hat{Z}$ is predetermined and carries maximal information content, in order to minimize Eq. 7, it is necessary for Z_t to be equal to $\hat{Z}$. Thus, minimizing Eq. 7 can be considered as a conduit for transferring comprehensive information from $\hat{Z}$ to both Z_1 and Z_2, enabling them to compensate for the semantic loss incurred by data augmentation. Consequently, while extracting consistent semantic information from Z_1 and Z_2, minimizing Eq. 7 facilitates the extraction of comprehensive semantic information.

Algorithm 1: The main algorithm

Input: Training set D; Batch Size n; Encoder function f_θ; Projection Head g_ϕ; Multi-layer Network g_ξ.
Parameter: Regularization Parameter: λ_1, λ_2
Output: The optimal encoder: f_{θ^*}

for sample batch X from D **do**
  # generate two augmented views
  $x_i^1, x_i^2 = t_1(x_i), t_2(x_i), t_1, t_2 \in A, x_i \in X$
  # obtain the augmented embeddings
  $z_i^1 = g_\phi(f_\theta(x_i^1))$
  $z_i^2 = g_\phi(f_\theta(x_i^2))$
  # obtain the more comprehensive embedding
  $\hat{z}_i = g_\xi(\hat{h}_i), where\ \hat{h}_i = h_i^1 \oplus h_i^2$
  # Under the fixed ξ, update $\{\theta_\xi, \phi_\xi\}$
  $\{\theta_\xi, \phi_\xi\} = \{\theta, \phi\} - r \cdot \nabla_{\theta,\phi}(\mathcal{L}_{ssl} + \lambda_1 \mathcal{L}_{mcr})$
  # Under the fixed $\{\theta_\xi, \phi_\xi\}$, update ξ
  $\xi = \xi - r \cdot \nabla_\xi(\mathcal{L}_{ssl}(\theta_\xi, \phi_\xi) + \lambda_2 \mathcal{L}_{comp})$
end for

Model Objective

Finally, we present the objective during the training phase, which can be divided into two steps. The first step is to learn f_θ and g_ϕ that can extract feature representation. The second step is to learn g_ξ that can obtain comprehensive representation by a bi-level optimization mechanism. The training process is shown in Algorithm 1.

Specifically, in the first step of each epoch, we fix g_ξ and update f_θ and g_ϕ through the following formulation:

$$\{\theta, \phi\} = \{\theta, \phi\} - r \cdot \nabla_{\theta,\phi}(\mathcal{L}_{ssl} + \lambda_1 \mathcal{L}_{mcr}) \tag{8}$$

where r is the learning rate and λ_1 is the hyperparameter. The purpose of learning Eq. 8 is to extract consistency semantic information between Z_1 and Z_2. However, $\mathcal{L}_{ssl} + \lambda_1 \mathcal{L}_{mcr}$ in Eq. 8 enables both Z_1 and Z_2 to simultaneously possess comprehensive semantic information. Thus, learning Eq. 8 can result in the consistency information between Z_1 and Z_2 being comprehensive information.

In the second step of each epoch, we fix f_θ and g_ϕ and update g_ξ through the following formulation:

$$\begin{aligned} &\xi = \xi - r \cdot \nabla_\xi(\mathcal{L}_{ssl}(\theta_\xi, \phi_\xi) + \lambda_2 \mathcal{L}_{comp}) \\ &s.t. \{\theta_\xi, \phi_\xi\} = \{\theta, \phi\} - r \cdot \nabla_{\theta,\phi}(\mathcal{L}_{ssl} + \lambda_1 \mathcal{L}_{mcr}) \end{aligned} \tag{9}$$

where $\mathcal{L}_{ssl}(\theta_\xi, \phi_\xi)$ represents that the loss $\mathcal{L}_{ssl}$ is calculated based on f_{θ_ξ} and g_{ϕ_ξ}, and λ_2 is the hyperparameter. It is important to note that during the computation of $\mathcal{L}_{ssl}$, g_ξ is not involved, hence direct differentiation of $\mathcal{L}_{ssl}$ with respect to ξ is not possible. However, when computing matrix $\mathcal{L}_{ssl}(\theta_\xi, \phi_\xi)$, as indicated by Eq. 9, θ_ξ and ϕ_ξ can be treated as functions of ξ, allowing for direct differentiation of $\mathcal{L}_{ssl}(\theta_\xi, \phi_\xi)$ with respect to ξ. Simultaneously, optimizing $\nabla_\xi \mathcal{L}_{ssl}(\theta_\xi, \phi_\xi)$ can be conceptualized as follows: by manipulating ξ to induce changes in $\mathcal{L}_{ssl}(\theta_\xi, \phi_\xi)$, constrained by the conditions outlined in Eq. 9, where $\mathcal{L}_{ssl}(\theta_\xi, \phi_\xi)$ is consistently optimized under these ξ-induced circumstances. Subsequently, among all optimal states of $\mathcal{L}_{ssl}(\theta_\xi, \phi_\xi)$, the objective is to identify a configuration that minimize the magnitude of $\nabla_\xi \mathcal{L}_{ssl}(\theta_\xi, \phi_\xi)$. So, optimizing

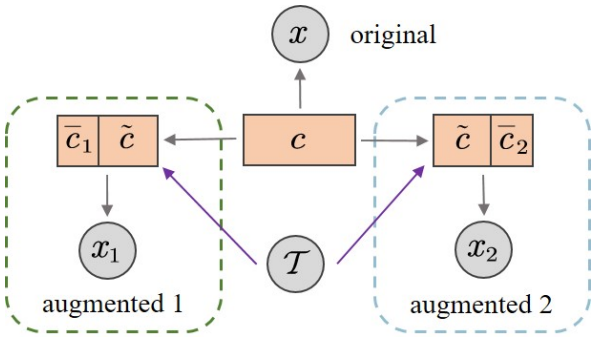

Figure 3: Structural causal model of latent variables. We assume that part of the semantic missing in the augmented view x_1 and x_2 compared to the original view x. These exclusive semantic $\bar{c}$ is affected by the semantic c and the applied augmentation strategy t.

$\nabla_\xi \mathcal{L}_{ssl}(\theta_\xi, \phi_\xi)$ can be simply regarded as the mechanism for modulating ξ to reduce its magnitude, thereby enhancing the similarity between matrices Z_1 and Z_2. As widely acknowledged, it is only when matrices Z_1 and Z_2 encapsulate a greater abundance of analogous semantic information that matrices Z_1 and Z_2 can exhibit increased similarity. Therefore, optimizing $\nabla_\xi \mathcal{L}_{ssl}(\theta_\xi, \phi_\xi)$ can compel g_ξ to extract a greater amount of semantic information. When coupled with the influence of $\mathcal{L}_{comp}$, this process enables Eq. 9 to coerce g_ξ into capturing comprehensive semantic information.

Causal Interpretation

We use a Structural Causal Model (SCM) (Pearl and Mackenzie 2018) to describe the causal relationship between variables in self-supervised learning, whereby data augmentation can be viewed as a counterfactual. We define c to be the semantic of original sample, $\tilde{c}$ is the invariant semantics across augmented views, $\bar{c}$ is the exclusive semantics in the augmented view. We assume that the samples are only generated by their semantic information (or label) and exogenous variables, as shown in Fig.3. The augmentation strategy t is randomly sampled from the strategy set $\mathcal{T}$. We can formulate the relation between these variables:

$$c := f_c(u_c), \tilde{c} := f_{\tilde{c}}(c, t), \bar{c} := f_{\bar{c}}(c, t), x := f(c, u_x);$$

where u_c, u_x is independent exogenous variable, and $f_{\tilde{c}}, f_{\bar{c}}$ are deterministic functions.

Given a factual observation $x' = f(c', u_{x'})$. If a new augmentation strategy t^* is applied, it is equivalent to an intervention mechanism affecting $f_{\tilde{c}}, f_{\bar{c}}$:

$$do(\tilde{c} := f_{\tilde{c}}(c', t^*), \bar{c} := f_{\bar{c}}(c', t^*))$$

Using the modified SCM by fixing all exogenous variables, the relation between the variables in the counterfactual sample $x'' = f(c', u_{x''})$ and its augmented views can be reformulated as:

$$c'' := c', \tilde{c}' := f_{\tilde{c}}(c', t^*), \bar{c}' := f_{\bar{c}}(c', t^*);$$

Thus, data augmentation in self-supervised learning can be viewed as counterfactual. Then, we can obtain:

Theorem 2. *Assume that the data generating process is consistent with the above description. Let $f : x \to z$ be any smooth function that can minimize the following objective:*

$$E_{(z_i^1, z_i^2) \sim (\mathcal{Z}_1, \mathcal{Z}_2)} \left[\left\| z_i^1 - z_i^2 \right\|_2^2 \right] - H_{min}(z_i^t) \qquad (10)$$

where $H_{min}(\cdot)$ denotes the minimum entropy of all augmented embeddings, $z_i^t = \{z_i^1, z_i^2\}$. In this way, the learned encoder f_θ can capture all semantic information related to the original sample.

The above theorem states that when we constrain the embeddings to have the maximum entropy, the SSL model can constrain the learned representation to contain more semantics. Note that the Eq. 10 is similar to Eq. 9. Therefore, this theorem also provides a theoretical basis for the proposed bi-level learning method.

Experimental Results

In this section, we first evaluate it on classification task using linear evaluation and semi-supervised learning settings. Then, we validate our method on object detection and instance segmentation tasks in computer vision.

Experimental Setting

Datasets. For the classification task, we evaluate our proposed method on the following six image datasets, including CIFAR-10 and CIFAR-100 dataset (Krizhevsky 2009), STL-10 dataset (Coates, Ng, and Lee 2011), Tiny ImageNet dataset (Le and Yang 2015), ImageNet-100 dataset (Russakovsky et al. 2015), and ImageNet dataset (Russakovsky et al. 2015). For transfer learning, we validate our method by the performance on the object detection and semantic segmentation tasks on COCO (Lin et al. 2014) dataset.

Default Setting. Each input sample generates two corresponding positive samples in the experiment. The image augmentation strategies comprise the following image transformations: random cropping, resizing, horizontal flipping, color jittering, converting to grayscale and gaussian blurring. Detailed experimental settings for different downstream tasks can be found in Appendix B. In the experiment, we use Resnet18 or Resnet50 as our base encoder network, along with a 3-layer MLP projection head to project the representation to a embedding space.

CompMod Details. The CompMod consists of a multilayer linear network, which is set to $2d' - d' - d$, where d' is the output dimension of the backbone network f_θ.

Downstream Tasks

Self-supervised Learning We conduct a classification task to test our proposed method. For comparison, we take SimCLR (Chen et al. 2020), BarlowTwins (Zbontar et al. 2021), BYOL (Grill et al. 2020), SimSiam (Chen and He 2021), W-MES (Ermolov et al. 2021), SwAV (Caron et al. 2020), MoCo (He et al. 2020), CMC (Tian, Krishnan, and Isola 2020), SSL-HSIC (Li et al. 2021) and VICReg (Bardes, Ponce, and Lecun 2022) as baselines. We validate the proposed method with the results of a linear classifier and a 5-nearest neighbor classifier. Table 1 shows the

Methods	CIFAR-10		CIFAR-100		STL-10		Tiny ImageNet	
	linear	5-nn	linear	5-nn	linear	5-nn	linear	5-nn
SimCLR	91.80	88.42	66.83	56.56	90.51	85.68	48.82	32.86
BarlowTwins	90.88	88.78	66.67	56.39	90.71	85.31	49.74	33.61
BYOL	91.73	89.45	66.60	56.82	91.99	88.64	51.00	36.24
SimSiam	91.51	89.31	66.73	56.87	91.92	88.54	50.92	35.98
W-MSE	91.99	89.87	67.64	56.45	91.75	88.59	49.22	35.44
SwAV	90.17	86.45	65.23	54.77	89.12	84.12	47.13	31.07
SSL-HSIC	91.95	89.91	67.22	57.01	92.06	88.87	51.42	36.03
VICReg	91.08	88.93	67.15	56.47	91.11	86.24	50.17	34.24
SimCLR+	93.91	91.21	**68.91**	**58.22**	92.77	87.98	51.01	35.23
BYOL+	**93.96**	**91.53**	68.74	58.01	**94.95**	**89.88**	**53.51**	**37.95**
BarlowTwins+	92.54	90.75	68.29	57.84	93.12	89.61	51.72	35.21

Table 1: Classification accuracy on small and medium datasets. Top 1 accuracy(%) of linear classifier and a 5-nearest neighbors classifier for different datasets with a ResNet-18. Best results are in bold.

Methods	ImageNet-100		ImageNet	
	top-1	top-5	top-1	top-5
SimCLR	70.15	89.75	69.32	89.15
MoCo	72.81	91.64	71.13	-
CMC	73.58	92.06	66.21	87.03
BYOL	74.89	92.83	74.31	91.62
SwAV	75.77	92.86	75.30	-
DCL	74.60	92.08	-	-
RELIC	-	-	74.81	92.23
SSL-HSIC	-	-	72.13	90.33
ICL-MSR	72.08	91.60	70.73	90.43
BarlowTwins	72.88	90.99	73.22	91.01
SimCLR+	72.21	91.23	71.89	91.52
BYOL+	**76.95**	93.94	75.11	**93.55**
BarlowTwins+	76.88	**94.11**	**75.62**	92.13

Table 2: Evaluation on ImageNet-100 and ImageNet datasets. The representations are obtained with a ResNet-18 with our method on top 1 accuracy(%) of linear classifier and a 5-nn classifier. Best results are in bold.

Methods	Epochs	1%		10%	
		top-1	top-5	top-1	top-5
SimCLR	1000	48.3	75.5	65.6	87.8
BYOL	1000	53.2	78.4	68.8	89.0
SwAV	1000	53.9	78.5	**70.2**	**89.9**
BarlowTwins	1000	55.0	79.2	69.7	89.3
SimCLR+	1000	49.1	75.8	65.8	88.0
BYOL+	1000	54.6	78.9	69.2	89.3
Barlow Twins+	1000	**56.1**	**79.8**	**70.2**	**89.9**

Table 3: Semi-supervised classification. We finetune the pre-trained model using 1% and 10% training samples of ImageNet following (Zbontar et al. 2021), and the top-1 and top-5 under linear evaluation are reported.

performance of different SSL methods, where "method+" denotes our proposed method. The results show that our proposed method improves the classification performance, in which SimCLR+ and BYOL+ improve by more than 2% on CIFAR10 and CIFAR100 dataset, while BYOL+ improves by about 2.5% on Tiny ImageNet dataset.

Furthermore, we test our method for classification on two larger datasets, ImageNet-100 and ImageNet. For comparison, we also add several other methods including MoCo (He et al. 2020), CMC (Tian, Krishnan, and Isola 2020), ICL-MSR (Qiang et al. 2022) and RELIC (Mitrovic et al. 2020). The results in Table 2 demonstrate that our method still improves over the baseline, e.g., BarlowTwins+ achieves 4% performance improvement on ImageNet-100, SimCLR+ and BarlowTwins+ achieve more than 2% on ImageNet.

Semi-supervised Learning The detailed experimental setup follows the most common evaluation protocol for semi-supervised learning, as in Appendix B. Table 3 reports the classification results on ImageNet compared with existing methods using two pre-trained models. From the results, Barlow Twins+ is 1.1% better than Barlow Twins, and BYOL+ increases by about 1.4% at the 1% subset setting.

Transfer Learning We evaluate our method for the localization based tasks of object detection and instance segmentation on COCO (Lin et al. 2014) datasets. ImageNet supervised pre-training is often used as initialization for finetuning downstream tasks. Several different self-supervised methods are used for performance comparison. We report the results of our proposed method compared with baselines in Table 5, showing that the proposed method brings performance improvements on different downstream tasks.

Ablation Experiments

Parametric Sensitivity In this section, we conduct an experimental investigation of the model trade-off parameters. Specifically, we vary λ_1 and λ_1 in the range of [0.001, 0.01, 0.1, 1], and record the classification accuracy of our method using a ResNet-18 on CIFAR-10 dataset with the SimCLR+ method. The results in Table 6 indicates that our method has minimal variation in accuracy, indicating that hyperparameter tuning is easy in practice.

ID	Data augmentations					Methods		
	horizontal flip	rotate	random crop	random grey	color jitter	SimCLR+	BYOL+	Barlow Twins+
1	✓	✓				93.65	92.64	91.75
2			✓			92.31	92.78	92.09
3				✓		92.78	93.16	92.15
4					✓	93.36	92.99	91.95
5	✓		✓			93.47	92.74	92.23
6		✓			✓	93.72	92.89	91.89
7	✓		✓	✓		93.75	93.78	92.48
8	✓	✓	✓	✓	✓	**93.91**	**93.96**	**92.54**

Table 4: Comparison of different data augmentations by using a ResNet-18 on CIFAR-10 dataset.

Methods	Object Det.			Instance Seg.		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Supervised	38.2	58.2	41.2	33.3	54.7	35.2
SimCLR	37.9	57.7	40.9	33.2	54.6	35.3
SwAV	37.6	57.6	40.2	33.0	54.2	35.1
BYOL	37.9	57.8	40.9	33.1	54.3	35.0
SimSiam	37.9	57.5	40.9	33.3	54.2	35.2
BarlowTwins	39.2	59.0	42.5	34.2	56.0	36.5
SimCLR+	38.1	58.1	41.0	33.7	54.6	35.1
BYOL+	**39.7**	59.1	42.9	**35.4**	56.1	36.2
Barlow Twins+	39.1	**59.3**	**43.1**	35.2	**56.2**	**36.9**

Table 5: Transfer learning. We pre-train the network on ImageNet dataset. Then, we learn representation on the object detection and instance segmentation tasks on COCO dataset using Mask P-CNN. Evaluation is on AP, AP_{50} and AP_{75}.

Analysis of Data Augmentation we compare the linear classification accuracy under different augmentation strategies on CIFAR-10 dataset as shown in Table 4. As can be seen, there is no significant difference in classification accuracy, indicating that our method can be applied to different augmentation strategies.

Fusion Strategy and Optimization In this section, we first investigate the impact of different fusion strategies.

Mixup (Verma et al. 2021) can be used to fuse features in representation space. We can obtain the more comprehensive representation using the following formula:

$$\hat{h}_i = \alpha * h_i^1 + (1 - \alpha) * h_i^2 \tag{11}$$

where α is a coefficient sampled from a uniform distribution, $\alpha \sim U(0, 1)$. By adjusting α, we can control the semantic information to be biased towards h_i^1 or h_i^2. Another strategy is to achieve semantic fusion in the embedding space:

$$\tilde{z}_i := z_i^1 \oplus z_i^2 = \left[z_i^1 \ z_i^2\right] \tag{12}$$

Then $\tilde{z}_i$ is mapped onto the embedding space to obtain $\hat{z}_i$ via a projection head g_ζ composed of a multi-layer linear network $(2d - d - d)$: $\hat{z}_i = g_\zeta(\tilde{z}_i)$.

Additionally, our proposed method utilizes a bi-level optimization mechanism for model optimization during training.

λ_2 \ λ_1	0.001	0.01	0.1	1
0.001	91.79	92.13	92.77	91.75
0.01	92.56	91.89	93.91	91.11
0.1	93.35	92.23	92.35	90.78
1	93.03	91.26	92.77	90.08

Table 6: Parametric analysis of λ_1 and λ_2.

	Mixup					M(h)	M(z)
α	0.1	0.3	0.5	0.7	0.9	-	-
Acc.	91.03	91.33	91.87	91.34	91.56	**93.96**	92.58
No bi-level						92.05	

Table 7: Analysis of Fusion Strategy and Optimization. Experimental results are based on the classification with BYOL+. M(h) means fusion in representation space, while, M(z) means fusion in embedding space

In this section, we also analyze the impact of not using this strategy in experiments.

Table 7 shows the results of BYOL+ utilizing different feature fusion strategies and optimization on the classification task on CIFAR-10 dataset. Experimental results demonstrate that the fusion strategy and optimization we adopt achieve the best results.

Conclusion

In this paper, we find that data augmentation in SSL may lead to the lack of task-related information from information theory,resulting in a reduction of the model's performance in downstream tasks. To this end, we design a novel module CompMod with Meta Comprehensive Regularization as a complement to existing SSL frameworks. CompMod exploits a bi-level optimization mechanism and constraint based on maximum entropy coding to enable more information to be discovered, thereby enhancing the generalization of the learned model. Moreover, a causal interpretation provide theoretical support for the proposed method. Finally, the performance of various downstream tasks validates the effectiveness of our proposed method.

Acknowledgements

This work was supported by China Postdoctoral Science Foundation under Grant 2023M743639, National Key R&D Program of China (2021YFB3500700), NSFC Grant 62172026, National Social Science Fund of China 22&ZD153, the Fundamental Research Funds for the Central Universities and SKLSDE.

References

Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.-H.; Chang, S.-F.; Cui, Y.; and Gong, B. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34: 24206–24221.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Balestriero, R.; and LeCun, Y. 2022. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. *Advances in Neural Information Processing Systems*, 35: 26671–26685.

Bardes, A.; Ponce, J.; and Lecun, Y. 2022. VICReg: Variance-Invariance-Covariance Regularization For Self-Supervised Learning. In *ICLR 2022-International Conference on Learning Representations*.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33: 9912–9924.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15750–15758.

Chen, X.; Xie, S.; and He, K. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9640–9649.

Coates, A.; Ng, A.; and Lee, H. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 215–223. JMLR Workshop and Conference Proceedings.

Cover, T. M. 1999. *Elements of information theory*. John Wiley & Sons.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Ermolov, A.; Siarohin, A.; Sangineto, E.; and Sebe, N. 2021. Whitening for self-supervised representation learning. In *International Conference on Machine Learning*, 3015–3024. PMLR.

Garrido, Q.; Chen, Y.; Bardes, A.; Najman, L.; and Lecun, Y. 2022. On the duality between contrastive and non-contrastive self-supervised learning. *arXiv preprint arXiv:2206.02574*.

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

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

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

Klir, G.; and Wierman, M. 1999. *Uncertainty-based information: elements of generalized information theory*, volume 15. Springer Science & Business Media.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. *Master's thesis, University of Tront*.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Li, J.; Qiang, W.; Zheng, C.; Su, B.; and Xiong, H. 2022a. Metaug: Contrastive learning via meta feature augmentation. In *International Conference on Machine Learning*, 12964–12978. PMLR.

Li, Y.; Pogodin, R.; Sutherland, D. J.; and Gretton, A. 2021. Self-supervised learning with kernel dependence maximization. *Advances in Neural Information Processing Systems*, 34: 15543–15556.

Li, Z.; Chen, Y.; LeCun, Y.; and Sommer, F. T. 2022b. Neural manifold clustering and embedding. *arXiv preprint arXiv:2201.10000*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, X.; Wang, Z.; Li, Y.-L.; and Wang, S. 2022. Self-supervised learning via maximum entropy coding. *Advances in Neural Information Processing Systems*, 35: 34091–34105.

Mitrovic, J.; McWilliams, B.; Walker, J.; Buesing, L.; and Blundell, C. 2020. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*.

Pearl, J.; and Mackenzie, D. 2018. *The book of why: the new science of cause and effect*. Basic books.

Qiang, W.; Li, J.; Zheng, C.; Su, B.; and Xiong, H. 2022. Interventional Contrastive Learning with Meta Semantic Regularizer. In *International Conference on Machine Learning*, 18018–18030. PMLR.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Shorten, C.; and Khoshgoftaar, T. M. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1): 1–48.

Takahashi, R.; Matsubara, T.; and Uehara, K. 2019. Data augmentation using random image cropping and patching for deep CNNs. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9): 2917–2931.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 776–794. Springer.

Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; and Isola, P. 2020. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33: 6827–6839.

Verma, V.; Luong, T.; Kawaguchi, K.; Pham, H.; and Le, Q. 2021. Towards domain-agnostic contrastive learning. In *International Conference on Machine Learning*, 10530–10541. PMLR.

Wang, H.; Guo, X.; Deng, Z.-H.; and Lu, Y. 2022. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16041–16050.

Xiao, T.; Wang, X.; Efros, A. A.; and Darrell, T. 2020. What Should Not Be Contrastive in Contrastive Learning. In *International Conference on Learning Representations*.

Yang, K.; Zhou, T.; Tian, X.; and Tao, D. 2022. Identity-disentangled adversarial augmentation for self-supervised learning. In *International Conference on Machine Learning*, 25364–25381. PMLR.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, 12310–12320. PMLR.

Zhang, J.; and Ma, K. 2022. Rethinking the augmentation module in contrastive learning: Learning hierarchical augmentation invariance with expanded views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16650–16659.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, 649–666. Springer.

Zheng, M.; You, S.; Wang, F.; Qian, C.; Zhang, C.; Wang, X.; and Xu, C. 2021. ReSSL: Relational Self-Supervised Learning with Weak Augmentation. *Advances in Neural Information Processing Systems*, 34.

Zhou, W.; Lee, D.-H.; Selvam, R. K.; Lee, S.; and Ren, X. 2020. Pre-training Text-to-Text Transformers for Concept-centric Common Sense. In *International Conference on Learning Representations*.