

Modeling Multiple Views via Implicitly Preserving Global Consistency and Local Complementarity

Jiangmeng Li, Wenwen Qiang, Changwen Zheng, Bing Su, Farid Razzak, Ji-Rong Wen, *Senior Member, IEEE* and Hui Xiong, *Fellow, IEEE*

Abstract—While self-supervised learning techniques are often used to mine hidden knowledge from unlabeled data via modeling multiple views, it is unclear how to perform effective representation learning in a complex and inconsistent context. To this end, we propose a new multi-view self-supervised learning method, namely *consistency and complementarity network* (CoCoNet), to comprehensively learn global inter-view consistent and local cross-view complementarity-preserving representations from multiple views. To capture crucial common knowledge which is implicitly shared among views, CoCoNet employs a global consistency module that aligns the probabilistic distribution of views by utilizing an efficient discrepancy metric based on the generalized sliced Wasserstein distance. To incorporate cross-view complementary information, CoCoNet proposes a heuristic complementarity-aware contrastive learning approach, which extracts a complementarity-factor jointing cross-view discriminative knowledge and uses it as the contrast to guide the learning of view-specific encoders. Theoretically, the superiority of CoCoNet is verified by our information-theoretical-based analyses. Empirically, our thorough experimental results show that CoCoNet outperforms the state-of-the-art self-supervised methods by a significant margin, for instance, CoCoNet beats the best benchmark method by an average margin of 1.1% on ImageNet.

Index Terms—unsupervised learning, self-supervised learning, representation learning, multi-view, regularization, Wasserstein distance.

1 INTRODUCTION

SELF-SUPERVISED learning (SSL) aims to learn representations from unlabeled data that nonetheless can have wide-reaching benefits. The key to the problem lies in designing appropriate SSL objectives. Recent works explore how to maximize the mutual information (MI) between the inputs and outputs of the encoder. For example, the MI between high dimensional continuous random variables is effectively estimated by neural networks over gradient descent [2], and the MI between the high-level features and the local regions of the low-level features are jointly maximized [3]. The capacity of the encoder is crucial for estimating the MI between input-output pairs, and the ultimate goal of this approach is to encode the discriminative

information for downstream tasks (e.g., classification) into representations. However, maximizing the MI between the input and the output of an encoder over a single view may encode the view-specific task-irrelevant information into the learned representations. These methods differ from the normal human learning process in how observations are represented. Humans have a tendency to perceive items from a variety of perspectives, including aural, gustatory, and visual. Concretely, the single-view-based methods are unable to extract task-relevant information from other views.

Recently proposed self-supervised learning methods, for example, SimCLR [4], SwAV [5], MoCo [6], AMDIM [7], gLMSC [8], and CMC [9], have extended maximizing the MI between the encoder input and output on single-view data to maximizing the MI between the same samples under different views on multi-view data. As an example, given two views X and \hat{X} of image data, these methods concentrate on maximizing the MI $I(X; \hat{X})$. The assumption behind these methods is that the task-relevant information lies mostly in the shared information between the different views [10]. Since the background of the data in different views may be different, maximizing the MI between the same data in different views will cause the encoder to focus on the shared information of the foreground in different views. It should be noted that for each view, the task-relevant discriminative information that is unique to that view also exists, which is referred to the view-specific and task-relevant information. In Figure 1 (a), we show an example of such view-specific and task-relevant information in image data for classification. However, there are no additional terms in the objective of the benchmark methods to extract the view-specific and task-relevant information. In Figure 1 (c), we further show an application to demonstrate that only mining

- J. Li and W. Qiang are with the University of Chinese Academy of Sciences, Beijing, China. They are also with the Science & Technology on Integrated Information System Laboratory, Institute of Software Chinese Academy of Sciences, Beijing, China. E-mail: jiangmeng2019@iscas.ac.cn, a01114115@163.com. They contributed equally to this work.
- C. Zheng is with the Science & Technology on Integrated Information System Laboratory, Institute of Software Chinese Academy of Sciences, Beijing, China. E-mail: changwen@iscas.ac.cn.
- F.Razzak is with the New York University & Columbia University, New York, USA. E-mail: farid.razzak@nyu.edu.
- B. Su and J.-R. Wen are with the Beijing Key Laboratory of Big Data Management and Analysis Methods, Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, 100872, China. E-mail: subingats@gmail.com; jrwen@ruc.edu.cn. Corresponding author: Bing Su.
- H. Xiong is with Thrust of Artificial Intelligence, the Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China. He is also with Department of Computer Science & Engineering, the Hong Kong University of Science and Technology, Hong Kong SAR, China. E-mail: xionghui@ust.hk.
- ©2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

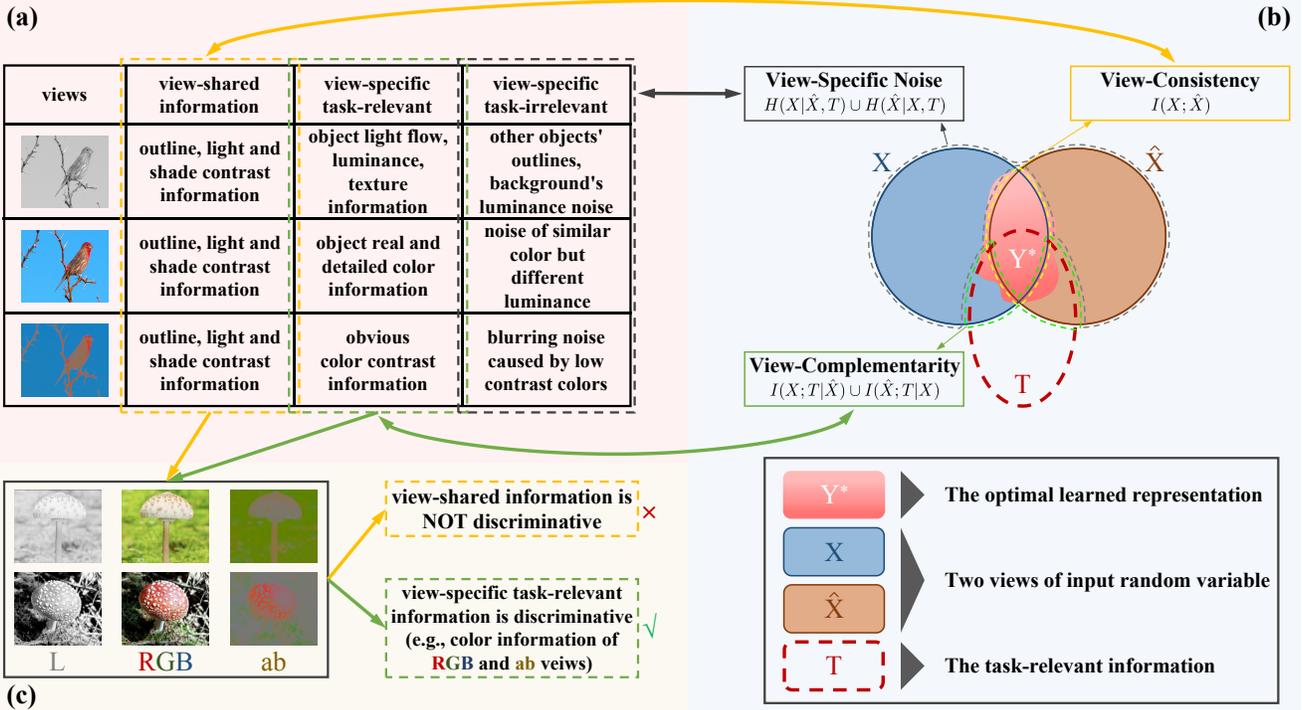


Figure 1: Illustration of the theoretical analysis based on information theory, where the *yellow* boxes show the shared view-consistency information, the *green* boxes represent the view-specific and task-relevant information which is also the desired view-complementarity information, and the *grey* boxes denote the view-specific but task-irrelevant noise information. (a) An example of the three mentioned information in practical application; (b) the definitions of the mentioned information in the information-theoretical perspective; (c) The application of using multiple views to learn discriminative information on a specific image classification task, which proves that only mining view-shared information is *not* enough to learn discriminative representations on benchmark datasets, e.g., ImageNet [1] has many similar fine-grained categories.

view-shared information is not enough so that mining the view-specific and task-relevant information can improve the general discriminability of the learned representations.

According to information theory, the information contained in the input is divided into three parts. Figure 1 (b) shows an example with a two-view dataset, where X and \hat{X} denote two views of a same sample, respectively, T denotes task-relevant information or label-relevant information, and Y^* denotes the optimal learned representation. The three parts of X and \hat{X} are as follows: the view-consistency information $I(X; \hat{X})$ to denote the view-shared information, which refers to the part surrounded by the yellow line; the view-complementarity information $I(X; T|\hat{X})$ and $I(\hat{X}; T|X)$ to denote the view-specific task-relevant information, which refer to the part surrounded by the green lines; and the view-specific noise $H(X|\hat{X}, T)$ and $H(\hat{X}|X, T)$ to denote the view-specific task-irrelevant information, which refers to the part enclosed by the grey lines. Meanwhile, we give their formal definitions in Section 5. Therefore, we suppose the discriminative learned representation should contain both view-consistency and view-complementarity information and discard view-specific noise, i.e., $H(Y^*) = I(X; \hat{X}) + I(X; T|\hat{X}) + I(\hat{X}; T|X)$.

However, benchmark methods are difficult to achieve such a objective. We rethink the learning paradigms of conventional self-supervised multi-view learning methods from the perspective of information theory, which is demonstrated in Figure 2. As shown in Figure 2 (a), methods that maximize

the MI between the inputs and the outputs of the encoder over a single view aim to extract the task-relevant information contained in a single view, e.g., $I(X; T)$, which refer to the red shaded part. However, such a built self-supervision problem is not enough to make the model to capture task-relevant information so that, after training, the optimal learned representation Y^* may contain the view-specific noise, i.e., $H(X, Y^* | T)$, which is denoted by the grey shaded part. Also, the task-relevant information contained in the other view, e.g., $I(\hat{X}; T|X)$, can not be extracted. As demonstrated in Figure 2 (b), the benchmark methods that maximize the MI between the different views of a same sample can only extract the view-consistency information contained in the $I(X; \hat{X})$ part and discard the view-specific noise $H(X|\hat{X}, T)$ and $H(\hat{X}|X, T)$. However, the view-complementarity information contained in $H(X, T|\hat{X})$ and $H(X, T|\hat{X})$ may also be discarded. Therefore, we motivate our method to sufficiently capture view-consistency and -complementarity information while discarding the view-specific noise, and the conceptual learning paradigm of our method is demonstrated in Figure 2 (c).

To this end, we propose an integrated SSL method for modeling multi-view data called *consistency and complementarity network* (CoCoNet). It projects all views into a latent space to obtain the feature representations and minimizes the generalized sliced Wasserstein distance discrepancy metric between the distribution of different views to enhance the consistency of multiple views in a global manner. For local

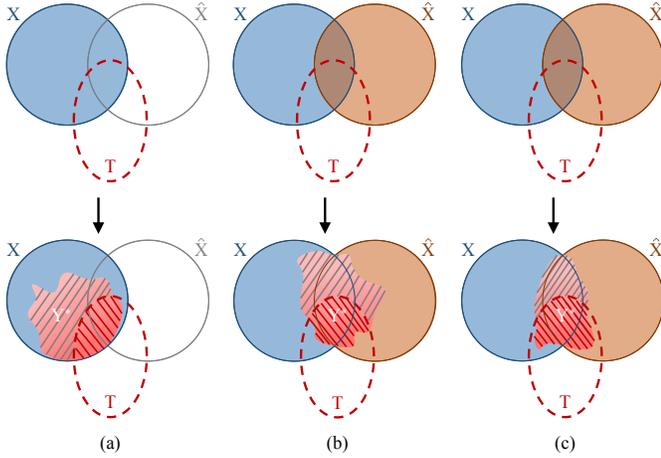


Figure 2: The conceptual learning paradigm plots of SSL methods. (a) the single-view SSL approach maximizing the MI between the inputs and the outputs of the encoder; (b) the conventional multi-view SSL approach maximizing the MI between the different views of a same sample; (c) our method. For the optimal learned representation Y^* , the red shaded part denotes the task-relevant information, and the grey shaded part denotes the task-irrelevant noise.

single views, CoCoNet proposes a novel complementarity-aware contrastive learning approach, which leverages the complementarity factor to guide the encoders to capture the view-complementary discriminative information and eliminate view-specific noise. In this way, CoCoNet aggregates the advances of multiple views and reduces the empirical risk of learning from each single view. Concretely, our proposed method aims to learn (albeit not fully) discriminative representations by using the strict consistency-preserving network to capture $I(X; \hat{X})$, and the proposed complementarity-aware contrastive learning approach prompts to capture $I(X; T|\hat{X})$ and $I(\hat{X}; T|X)$. It is worth noting that the proposed CoCoNet is a general unsupervised representation learning model and the learned representation can be applied in a wide range of downstream tasks. In the experiments, we verify the effectiveness of CoCoNet on image classification tasks using multi-view data. The major contributions are four-fold:

- We minimize a specific discrepancy metric to align the distributions of different views. As a result, the shared information between multiple views is extracted. This is to constrain our model to learn the view-consistency information, thereby reducing the impact of view-specific and task-irrelevant noise.
- We propose a heuristic complementarity-aware contrastive learning approach to enable the encoders to gain the view-specific and task-relevant information by using a novel complementarity-factor.
- We provide the information-theory-based analyses to demonstrate that preserving the global consistency and local complementarity can improve the discriminability of the learned multi-view representations.
- Following the protocol of [11], we perform empirical evaluations. Results show that CoCoNet outperforms previous works on benchmark and practical datasets. We have also demonstrated the generality of our

method to different forms and types of multiple-view data with different characteristics.

2 RELATED WORKS

Unsupervised learning. Unsupervised representation learning gets rid of the reliance on labeled data [12]. It starts with classical methods (without the use of deep neural networks), such as independent component analysis (ICA) [13], self-organizing maps [14], and principal components analysis (PCA) [15]. SSL is a specific kind of unsupervised learning. However, there is a principal difference between it and classical unsupervised learning; it requires a designed generator of supervised learning problems. SSL methods must capture helpful information about the data to solve the generated problems.

SSL performs well with use of deep neural networks, and it started with seminal techniques (e.g., Boltzmann machines [16], [17], autoencoders [18], variational autoencoders [19], β variational autoencoders [20], generative adversarial networks [21], adversarial autoencoders [22], autoregressive models [23], BiGAN (a.k.a. adversarially learned inference with a deterministic encoder [24]), Split-Brain Autoencoders (SplitBrain) [25], etc.). Recently, SSL is used in many fields, e.g., the NLP, vision, and robotics communities fields [26], [27], and with the development of contrastive learning, several approaches based on it have come to the forefront, for instance, Noise As Targets (NAT) [28], Contrastive Predictive Coding (CPC) [29], Deep InfoMax (DIM) [3], a simple framework for contrastive learning of visual representations (SimCLR) [4], Momentum Contrast (MoCo) [6], and Contrastive Multiview Coding (CMC) [9]. Existing generative models that maximize MI are also popular in this research area [3], [30]. However, learning the representations from a single view does not significantly improve performance of the task.

Multi-view learning. In order to capture information from different views, existing Multi-View Learning methods jointly consider multiple views for different downstream tasks (e.g., clustering [31] and classification [32]). The multi-view representation learning methods based on Canonical Correlation Analysis (CCA) [33] are representative, which project different views into a common space, e.g., kernelized CCA [34], CCA-based deep neural network [35], and semi-pair and semi-supervised generalized correlation analysis (S2GCA) [36]. The unsupervised multi-view learning methods [3], [7], [9] have also shown remarkable success in multi-view representation learning, while there is a crucial issue that learning from unaligned multiple views can lead to the poor performance of the representations.

Distributions aligning. Refer to the alignment of domain distributions, and the existing distribution alignment methods can be divided into three categories [37]. Instance-based approaches [38] align the distributions by sub-sampling the training data of the two domains. Parameter-based approaches [39] add adaptation layers or adaptive normalization layers to align the distributions. Furthermore, representation learning (RL) based approaches [40], [41] primarily map the input to a common latent space and then align the two distributions in the latent space. Further, employing deep neural networks to align the distributions

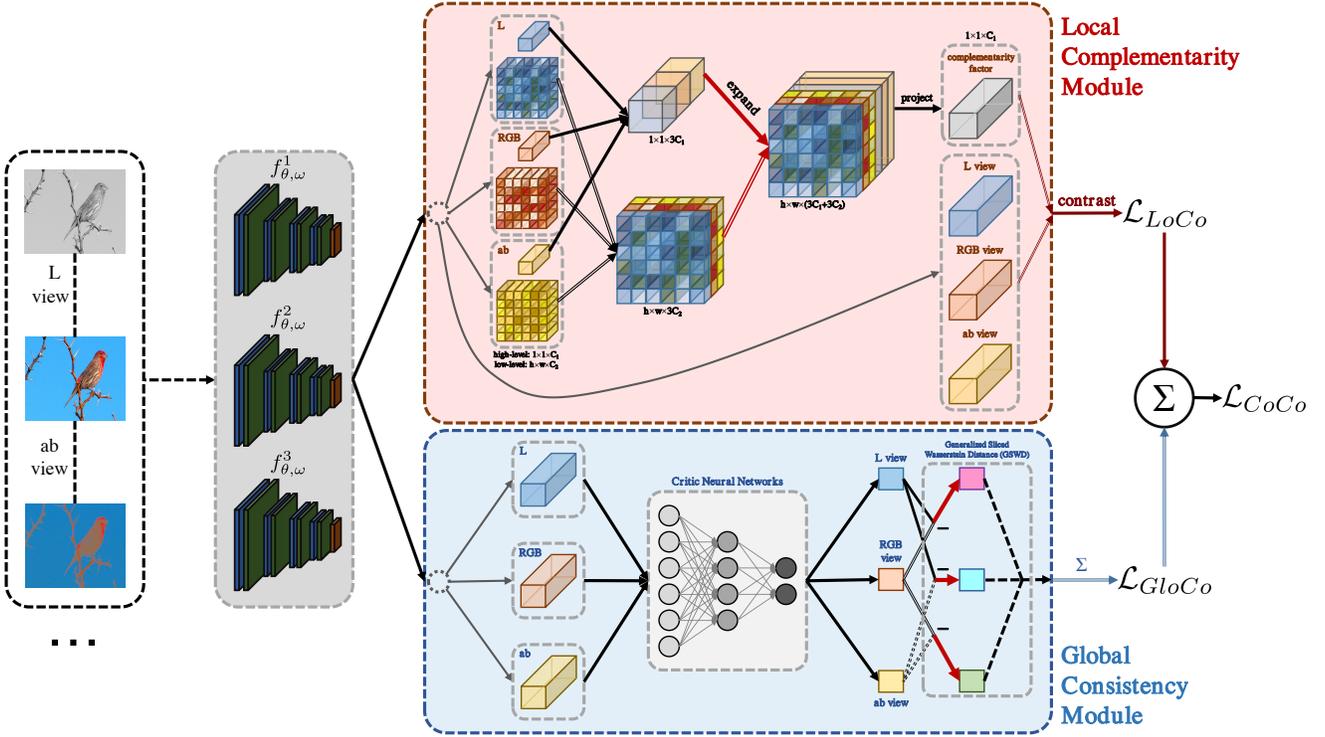


Figure 3: Overview of the proposed CoCoNet, and we demonstrate an example of adopting three views: RGB view, L view, and ab view. It consists of two modules: the local complementarity module for capturing the view-complementarity information and the global consistency module for acquiring the view-consistency information, which can jointly eliminate view-specific noise and capture both of view-shared and view-specific discriminative information.

by minimizing a certain metric [42], [43] is a standard method found in RL-based approaches. The metrics of the RL-base approaches include the KL-divergence, the H-divergence, and the Wasserstein distance [44], [45], [46], [47]. The strength of Wasserstein distance, compared with other metrics, is that it takes advantage of gradient superiority. The literature findings motivated our proposal of a metric based on Wasserstein distance to align the distribution of different views in a latent space globally.

3 PROBLEM DEFINITION

Formally, we consider the multi-view dataset $X^m = [x_1^m, x_2^m, \dots, x_N^m]$, where X^m represents the sample collection from the m -th view, and $x_i^m, i = 1, \dots, N, m = 1, \dots, M$ denotes the i -th sample of the m -th view. N is the number of samples in the m -th view and M is the number of views. We denote x^m as a random variable sampled *i.i.d* from the distribution $\mathcal{P}(x^m)$. Also, we denote $x_i = [x_i^1; x_i^2; \dots; x_i^M]$, which represents a complete sample that consists of the same samples from different views, $X = [x_1, x_2, \dots, x_N]$ represents a complete dataset, and x presents a variable sampled *i.i.d* from distribution $\mathcal{P}(x)$. The self-supervised multi-view learning aims to learn multiple encoders capable of extracting discriminative features for each view's data in an unsupervised manner, so that it can better serve downstream tasks such as classification. Specifically, we first project all multi-view data into the latent space through multiple encoders to obtain their feature representations, each view corresponds to a encoder, e.g., f^m for the m -th view. Then, a certain objective is minimized to the

parameters of the f^m . In this paper, the objective minimized in our proposed method consists of two parts including the loss function of the global consistency module and the complementarity-aware contrastive loss function. We will introduce these two loss function in the next section.

4 METHOD

In this section, we present the proposed *consistency and complementarity network* (CoCoNet) in detail, which aims at learning feature representations that can jointly model view-consistent factors and view-complementary factors from multi-view data. As shown in Figure 3, CoCoNet consists of two modules, i.e., the local complementarity module for capturing the view-complementarity information, and the global consistency module for acquiring the view-consistency information.

4.1 Global consistency module

The idea behind global consistency module is to learn a feature representation that globally captures information shared among multiple views. We first adopt the view-specific encoders to project the original inputs $\{X^1, X^2, \dots, X^M\}$ into the latent space. The resulting latent representations $\{H^1, H^2, \dots, H^M\}$ fit the distributions $\mathcal{P}(H^1), \mathcal{P}(H^2), \dots, \mathcal{P}(H^M)$. Then, we align $\mathcal{P}(H^1), \mathcal{P}(H^2), \dots, \mathcal{P}(H^M)$ in the latent space by minimizing the discrepancy among the distributions.

Wasserstein distance is widely used as the discrepancy measure. Let $\mathcal{P}(H)$ be the set of Borel probability measures. For $P_r, P_g \in \mathcal{P}(H)$ and the corresponding support set

Σ_r, Σ_g , respectively. Then, the p -th Wasserstein distance of the corresponding distributions is defined as:

$$W_p(P_r, P_g) = \left(\inf_{\mu(x_r, x_g) \in \Pi(x_r, x_g)} \int c(x_r, x_g)^p d\mu \right)^{\frac{1}{p}}, \quad (1)$$

where $x_r \in \Sigma_r, x_g \in \Sigma_g$, $c(x_r, x_g)$ denotes the distance of two patterns in Σ_r, Σ_g , and $\Pi(x_r, x_g)$ denotes the set of all joint distributions $\mu(x_r, x_g)$ that satisfies $P_r = \int_{x_g} \mu(x_r, x_g) dx_g, P_g = \int_{x_r} \mu(x_r, x_g) dx_r$.

Directly calculating $W_p(p_1, p_2)$ is computationally expensive. An alternative is using the popular dual version to calculate it, yet the Lipschitz constraint is difficult to meet. Therefore, we use the generalized sliced Wasserstein distances (GSWD), which is proposed by [47], to approximate $W_p(p_1, p_2)$. GSWD is defined as:

$$GSWD_p(P_r, P_g) = \int_{\Omega_\vartheta} W_p(GR_\vartheta P_r, GR_\vartheta P_g) d\vartheta \quad (2)$$

where Ω_ϑ denotes a compact set of feasible parameters for GR_ϑ , GR_ϑ represents one-dimensional nonlinear projection operation, also denoted as the critic neural network. Therefore, due to the non-linearity of GR_ϑ , the GSWD is expected to capture the complex structure of high-dimensional distributions (see the details of GR_ϑ in Appendix 9.2).

From the perspective of the gradient, we analyze the superiority of adopting GSWD as the discrepancy metric compared with other metrics, e.g., Kullback-Leibler (KL) divergence, which is described in Section 5.2. Empirically, we further explore the improvement of adopting various divergences as the discrepancy metric in Section 6.5.3.

Concretely, the loss function of the global consistency module is defined as:

$$\mathcal{L}_{GloCo} = \sum_{i=1}^{M-1} \sum_{j=i+1}^M GSWD_p(\mathcal{P}(H^i), \mathcal{P}(H^j)) \quad (3)$$

We denote the network by GloCo if only the global consistency preserving module is employed.

4.2 Local complementarity module

Each view may contain unique discriminative information complementary to other views, which cannot be captured by the conventional contrastive learning approach. The local complementarity module aims to encode such view-complementarity information from multiple views in an instance-based manner. To this end, we first extract a complementarity-factor and then maximize the MI between this factor and the latent features of each view. The pipeline of the local complementarity module is depicted in Figure3.

We incorporate the local discriminative knowledge of all views into a complementarity-factor. Specifically, given a sample x_i with M -views $\{x_i^1, x_i^2, \dots, x_i^M\}$, we first encode these views to obtain the low-level feature maps $\{z_i^1, z_i^2, \dots, z_i^M\}$ with $h \times w \times C_2$ dimensions by the view-specific feature extraction networks $\{f_\omega^1, f_\omega^2, \dots, f_\omega^M\}$, where C_2, h , and w are the number of channels, height, and width of the low-level feature maps, respectively. Then, we map $\{z_i^1, z_i^2, \dots, z_i^M\}$ into C_1 -dimensional high-level feature vectors $\{h_i^1, h_i^2, \dots, h_i^M\}$ by the view-specific mapping networks $\{f_\theta^1, f_\theta^2, \dots, f_\theta^M\}$. We

concatenate these high-level feature vectors to obtain a $M \cdot C_1$ -dimensional syncretic feature vector h_i , which is reckoned to combine the shared information among high-level feature vectors $\{h_i^1, h_i^2, \dots, h_i^M\}$. Then, the low-level feature maps are also concatenated to obtain a $h \times w \times M \cdot C_2$ -dimensional syncretic feature map z_i , which is considered to capture the shared low-level information from $\{z_i^1, z_i^2, \dots, z_i^M\}$.

For the sake of combining both high-level and low-level information, we expand h_i to a $h \times w \times M \cdot C_1$ feature map, and then concatenate it with the syncretic low-level feature map z_i to obtain a $h \times w \times M \cdot (C_1 + C_2)$ embedding. Then, we project this embedding to a C_1 -dimensional feature vector, called complementarity-factor CF_i . For a sample x_i , we finally obtain M high-level vectors $\{h_i^1, h_i^2, \dots, h_i^M\}$ and a complementarity-factor CF_i .

CF_i is expected to encode comprehensive and complementary information from different views, but may also contain redundant view-specific noises. To filter out such redundant information while maintain useful complementary information, we perform contrast learning to maximize the MI between the high-level features of different views and the complementarity-factor. In a minibatch with n samples $[x_1, x_2, \dots, x_n]$, for each sample x_i , we regard $\{h_i^1, h_i^2, \dots, h_i^M\}$ of all its views and the complementarity-factor CF_i as the positive terms, and $\{h_j^1, h_j^2, \dots, h_j^M\}$ and the corresponding complementarity-factors CF_j of the other samples as the negative terms where $j \in \{1, \dots, n\} \cap j \neq i$.

Conventional contrastive loss [29] can be formulated:

$$\mathcal{L}_{contrast} = - \mathbb{E}_{X^n} \left[\log \frac{S_\tau(p)}{S_\tau(p) + \sum_{j=1}^k S_\tau(n_j)} \right] \quad (4)$$

where S_τ is a score function to measure the positive pairs and the negative pairs, p denotes the positive pair, n_j denotes the negative pair sampled, and k denotes the number of the sampled negative pairs.

We insert the complementarity-factor into the contrastive loss to generate a novel complementarity-aware contrastive loss by reformulating the Equation 4 as follows:

$$\mathcal{L}_{cf-contrast} = -\alpha \cdot \mathbb{E}_{X^n} \left[\log \frac{S_\tau(\tilde{p})}{S_\tau(\tilde{p}) + \sum_{j=1}^k S_\tau(\tilde{n}_j)} \right] - \beta \cdot \mathbb{E}_{X^n} \left[\log \frac{S_\tau(p)}{S_\tau(p) + \sum_{j=1}^k S_\tau(n_j)} \right] \quad (5)$$

where, \tilde{p} denotes the positive pair, which consists of a positive term, i.e., a high-level feature of view h_i^m of the selected sample x_i , and the complementarity-factor CF_i . Also, \tilde{n}_j denotes the negative pair of a negative term and the according complementarity-factor. In order to further study the impacts of the two parts, we excessively set two hyper-parameters, i.e., α and β , to balance the two terms.

More specifically, the positive pair and negative pair are constructed as follows. For the complementarity-aware contrastive term, we group one of the positive terms, $\{h_i^1, h_i^2, \dots, h_i^M\}$ and CF_i , to form a positive pair. A negative term, $\{h_j^1, h_j^2, \dots, h_j^M\}$ and CF_j , are bonded as a negative pair. Then, expanding the Equation 5, we formalize the proposed complementarity-aware contrastive loss function as follows:

$$\begin{aligned} \mathcal{L}_{LoCo} = & -\alpha \cdot \sum_{i=1}^n \sum_{m=1}^3 \log \frac{S_\tau(h_i^m, CF_i)}{S_\tau(h_i^m, CF_i) + \sum_{j=1 \cap j \neq i}^n S_\tau(h_i^m, CF_j)} \\ & -\beta \cdot \sum_{i=1}^n \sum_{m=1}^3 \sum_{t=1 \cap t \neq m}^3 \log \frac{S_\tau(h_i^m, h_i^t)}{S_\tau(h_i^m, h_i^t) + \sum_{j=1 \cap j \neq i}^n \sum_{q=1}^3 S_\tau(h_i^m, h_j^q)} \end{aligned} \quad (6)$$

S_τ is the contrastive feature measurement function, which is implemented as:

$$S_\tau(a, b) = \exp\left(\frac{\langle (a), (b) \rangle}{\|(a)\| \cdot \|(b)\|} \cdot \frac{1}{\tau}\right) \quad (7)$$

where a and b denote the input high-level feature vectors, $\langle \cdot, \cdot \rangle$ is the inner product operator, $\|\cdot\|$ is the L_2 -norm, τ is the fixed temperature coefficient. We denote the local complementarity module as LoCo.

4.3 Consistency and complementarity network

As shown in Figure 3, our proposed CoCoNet incorporates the local consistency module and the global consistency module. Overall, the loss for the proposed CoCoNet is the weighted sum of the losses for the two modules:

$$\mathcal{L}_{CoCo} = \mathcal{L}_{LoCo} + \gamma \cdot \mathcal{L}_{GloCo} \quad (8)$$

where γ is the coefficient that controls the balance between \mathcal{L}_{LoCo} and \mathcal{L}_{GloCo} . By substituting Equation 3 and Equation 6 into Equation 8, the objective is formulated as follows:

$$\begin{aligned} \min_{f_\omega, f_\theta} \left\{ & -\alpha \sum_{i=1}^n \sum_{m=1}^3 \log \frac{S_\tau(h_i^m, CF_i)}{S_\tau(h_i^m, CF_i) + \sum_{j=1 \cap j \neq i}^n S_\tau(h_i^m, CF_j)} \right. \\ & \left. -\beta \sum_{i=1}^n \sum_{m=1}^3 \sum_{t=1 \cap t \neq m}^3 \log \frac{S_\tau(h_i^m, h_i^t)}{S_\tau(h_i^m, h_i^t) + \sum_{j=1 \cap j \neq i}^n \sum_{q=1}^3 S_\tau(h_i^m, h_j^q)} \right. \\ & \left. + \gamma \sum_{i=1}^2 \sum_{j=i+1}^3 \underbrace{GSWD_p(\mathcal{P}(H^i), \mathcal{P}(H^j))}_{GloCo} \right\} \end{aligned} \quad (9)$$

Minimizing the first two terms of equation 9, i.e., the loss of LoCo, can guide the multi-view representations to model more discriminative local view-complementarity information, and minimizing the last term of equation 9, i.e., the loss of GloCo, can globally make the representations consistent. We conduct experiments to study the influence of α , β , and γ , respectively, which is manifested in Section 6.5.

Algorithm 1 CoCoNet

Input: Multi-view dataset $X^m = [x_1^m, x_2^m, \dots, x_N^m]$, mini-batch size n , critic network training steps s , the learning rates ℓ_ω and ℓ_θ for the view-specific feature extractors f_ω and mapping networks f_θ , the learning rate ℓ_{critic} for the critic network GR_ϑ , and hyperparameters α, β, γ .

Initialize $\ell_\omega, \ell_\theta, \ell_{critic}, f_\omega, f_\theta$, and GR_ϑ .

repeat

 Sample minibatch $\{x_i\}_{i=1}^n \in X^m$.

for $t = 1$ **to** s **do**

 # Fix f_ω and f_θ

 # Train the critic network GR_ϑ to get \max GSWD

$\vartheta \leftarrow \vartheta - \ell_{critic} \cdot \Delta_\vartheta(-\mathcal{L}_{GloCo})$

end for

 # Fix GR_ϑ and train f_ω and f_θ

$\omega \leftarrow \omega - \ell_\omega \cdot \Delta_\omega(\mathcal{L}_{LoCo} + \gamma \cdot \mathcal{L}_{GloCo})$

$\theta \leftarrow \theta - \ell_\theta \cdot \Delta_\theta(\mathcal{L}_{LoCo} + \gamma \cdot \mathcal{L}_{GloCo})$

until f_ω, f_θ converge.

Following [9], we maintain a memory bank to store latent features for each training sample. In addition, we build an extra memory bank for efficiently retrieving complementarity-factor features on the fly. We elaborate the training pipeline in Algorithm 1, and the code is available at <https://github.com/jiangmengli/CoCoNet>.

5 THEORETICAL ANALYSES

In this section, we analyze the proposed CoCoNet from the information-theoretical perspective, and we also provide the theoretical analysis about the advantages of using the generalized sliced Wasserstein distance as the selected metric.

5.1 The information-theoretical analysis of CoCoNet

Notation. Figure 1 demonstrates a visual illustration of CoCoNet by using information theoretical description. We regard the input random variable as X and another view of X as \hat{X} in the figure, e.g., $X = X^1$ and $\hat{X} = X^2$. T presents the downstream task-relevant information. Y^* is the optimal representation learned from the deterministic encoder $f_{\vartheta, \omega}(\cdot)$, i.e., $f_\vartheta(f_\omega(\cdot))$ that includes the feature extraction network f_ω and the mapping network f_ϑ . For random variables A, B , and C , $H(A)$ denotes the entropy of A , and $H(A|B)$ denotes the conditional entropy of $H(A) - H(B)$. Accordingly, $I(A; B)$ presents the MI of A and B , and $I(A; B|C)$ represents the conditional MI of $I(A; B) - H(C)$.

To clarify the information diagrams of CoCoNet, we detail the definitions as follows:

Definition 5.1. *View-Consistency information is the discriminative information that is shared among views.*

Definition 5.2. *View-Complementarity information is the task-relevant information that is view-specific.*

Definition 5.3. *View-Specific Noise is the task-irrelevant information that only exists in one specific view.*

Considering the Definitions 5.1, 5.2, and 5.3, we rewrite the common multi-view assumption [10], [48] to describe multi-view learning between multiple views:

Assumption 5.1. (Multi-view, rewriting Assumption 1 in work [10]). The different views are approximately redundant to each other for the task-relevant information, based on 5.2, which is the View-Complementarity information, denoted as $\epsilon^{\text{complementarity}}$. For the View-Complementarity information of each view, we have $I(X^i; T|X^j) \leq \epsilon^{\text{complementarity}}$ with $i, j \in \{1, \dots, m\} \cap i \neq j$.

Assumption 5.1 states that, for $\epsilon^{\text{complementarity}}$, when it is small, the task-relevant information mainly lies in the MI between the input and the self-supervised signal. Therefore, when the number of views, m , is not large, as m increases, $I(X^i; T|\{X^j\}_{j=1 \cap j \neq i}^m)$ gets more compressed, and the ratio of discriminative task-relevant information rise, since the constraints of the MI, $\{\max_{i, j=1 \cap i \neq j}^m I(X^i; X^j)\}$ become stronger. Accordingly, the latent representation is more discriminative, which is supported by the view-vanishing experiment (See Section 6.5).

Definition 5.4. (Consistent and Complementary Multi-view Representations for Self-supervision). Let Y denotes the initially learned multi-view representation and Y^* denotes the consistent and complementary multi-view representation with restricted view-shared and view-specific discriminative knowledge: $Y^* = \underset{Y}{\operatorname{argmax}} I(Y; \{X^i\}_{i=1}^m; T)$ s.t. $I(Y; \{X^i\}_{i=1}^m)$ is maximized.

To learn the view-consistent and view-complementary multi-view representations Y^* from the multiple views $\{X^i\}_{i=1}^m$, different from previous works that roughly maximize the MI $I(Y; X^1; X^2)$ in a multi-view manner by utilizing the conventional contrastive learning framework, we globally align the distribution of view to guide the encoders to model view-shared information by availing of the efficient generalized sliced Wasserstein distance, which, based on the Definitions 5.1 and 5.3, is defined as:

Theorem 5.1. (View-Consistency information with a potential loss of View-Specific Noise ϵ^{noise}). Y^* is the sufficiently compressed latent representation, while Y is the self-supervised representation with part of the view-specific but task-irrelevant information ϵ^{noise} . Formally, only considering two views, i.e., X^1 and X^2 , $I(X^1; Y) \geq I(X^1; Y|\epsilon^{\text{noise}}) = I(X^1; Y^*) = I(X^1; X^2)$.

Proof. See Proof 9.1.1 in Appendix 9.1 for details. \square

Based on Theorem 5.1, we propose an implicit View-Consistency preserving regularization, which is approximated by the global consistency network in Section 4.1.

Theorem 5.2. (View-Complementarity information, which is view-specific and task-relevant). Y is the sufficiently compressed latent representation, and Y^* is the latent representation of adding view-specific and task-relevant information into Y . Formally, considering X^1 and X^2 , $I(X^1; Y; T) = I(X^2; Y; T) = I(Y; T) \leq I(Y; T) + I(X^1; Y^*; T|X^2) + I(X^2; Y^*; T|X^1) = I(Y^*; T)$.

Proof. See Proof 9.1.2 in Appendix 9.1 for details. \square

By the same token, with the intuition of Theorem 5.2, we introduce an implicit View-Complementarity preserving regularization and implement it by the local complementarity network in Section 4.2.

5.2 The gradient advantages of the generalized sliced Wasserstein distance

In order to align the distributions of views, we minimize the divergence between distributions $\{\mathcal{P}(X^1), \mathcal{P}(X^2), \dots, \mathcal{P}(X^m)\}$. To this end, we map data into a common latent space and then measure the distance based on a specific discrepancy metric, which reduces the dimensionality of representations in latent space. The representation's wide distribution may exist throughout the latent space. Then, for the conventional discrepancy metric (e.g., KL-divergence), the data points located in a region where the probability of a certain distribution is extremely greater than other distributions have little contribution to the gradient with cross-entropy loss. At the same time, the generalized sliced Wasserstein distance can provide stable gradients for every data point. Learning from [45], [49], we found that there will be a gradient vanishing problem if making data indistinguishable based on the conventional discrepancy metric in the case of the distributions has supports lying on low dimensional manifolds in the latent space. Also, we can get stable gradients by adopting the generalized sliced Wasserstein distance. Theoretically, consistent performance is achievable by using the generalized sliced Wasserstein distance.

6 EXPERIMENTS

In this section, we compared the proposed method against a fully-supervised classifier similar to the Alexnet architecture and various benchmark unsupervised methods to evaluate the performance of CoCoNet. To comprehensively evaluate the performance of the propose method, we imposed CoCoNet on *four* major downstream tasks: 1) benchmark image classification; 2) benchmark graph prediction; 3) benchmark action recognition; 4) practical object detection.

6.1 Preparation

We conducted experiments on neural network methods (i.e., the convolutional (conv) neural network-based method and the fully-connected (fc) network-based method) on benchmark datasets. Furthermore, we studied the performance of the ablation models of CoCoNet in conducted experiments, and we took CIFAR10 as the target dataset for the deepgoing exploration.

For setting the ablation study of CoCoNet, we compared with two main ablation models: GloCo, and LoCo. In details, CoCoNet refers to the complete model that considering the global consistency preserving, and the local complementarity preserving (i.e., $\alpha = 1, \beta = 0.5, \gamma = 10^{-4}$). GloCo refers to an ablation model of CoCoNet by removing the local complementarity preserving module (i.e., $\alpha = 0, \beta = 0, \gamma = 10^{-4}$). Since GloCo only employs the global consistency module, it can be treated as a view-alignment method. Therefore, GloCo can also be applied to conventional SSL methods. We combined GloCo and conventional self-supervised methods, e.g., GloCo+SwAV [5] and GloCo+CMC [9], to evaluate the performance of GloCo, where GloCo serves as a trimmer to align the feature distribution of different views. The features for each view are generated by these SSL methods. Similarly, LoCo refers to the model with only the local complementarity module (i.e., $\alpha = 1, \beta = 0.5, \gamma = 0$).

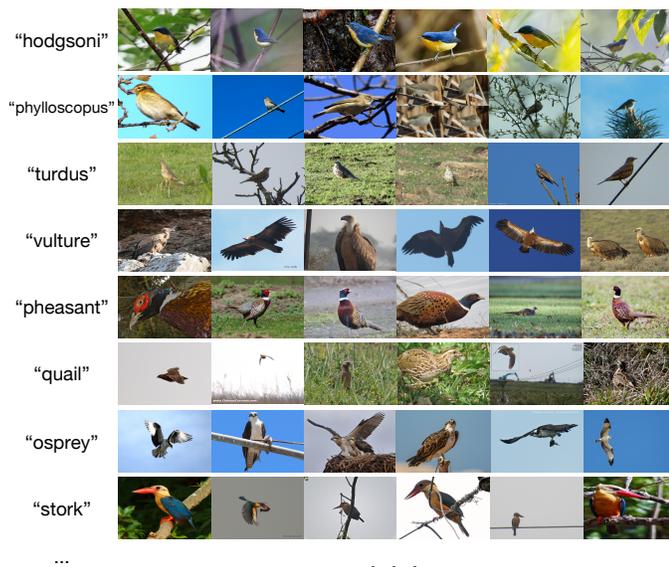


Figure 4: Example images in the WHEBD-759-SIM dataset.

6.2 Datasets

6.2.1 Benchmark image classification datasets

We conducted experiments on 5 established image representation learning datasets:

CIFAR10 dataset [50] is a small-scale labeled dataset composed of 32×32 images with 10 classes. CIFAR10 has 45,000 examples for training the classifier, and 5,000 examples for testing.

CIFAR100 dataset [50] is a small-scale labeled dataset consisting of 32×32 images of 100 categories, and each category contains 500 examples for training the classifier and 100 examples for testing.

STL-10 dataset [51] is a dataset derived from ImageNet composed of 96×96 images, and it contains a mixture of 100,000 unlabeled training examples and 500 labeled examples per class.

Tiny ImageNet dataset [50] is a reduced version of ImageNet ILSVRC [52], and images are scaled down to 64×64 with 200 classes. Each class has 500 images. The test set contains 10,000 images.

ImageNet dataset [1] consists of 1,000 image classes and is frequently considered as a testbed for unsupervised representation learning algorithms.

6.2.2 Benchmark graph prediction datasets

We conducted experiments on 5 established benchmark molecule prediction downstream tasks: molesol (mole), molipo (moll), molbbbp (molb), moltox21 (molt) and molsider (mols) from **Open Graph Benchmark (OGB)** [53].

6.2.3 Benchmark action recognition datasets

We conducted experiments on 2 established video datasets:

UCF-101 dataset [54] is a dataset for realistic action videos, providing 13,320 videos from 101 action categories.

HMDB-51 dataset [55] contains 6,849 samples, divided into 51 categories, each category contains at least 101 samples.

6.2.4 Practical object detection datasets

We introduced CoCoNet on a practical bird detecting and classifying dataset to validate the effectiveness of the proposed method under real-world circumstances, as follows:

Waterfowl and Habitat of Earth Big Data dataset (WHEBD) is a real-world and extensible dataset, which is automatically updated in cycles. We truncated WHEBD-759-2020 (WHEBD-759) from the complete WHEBD. An example of WHEBD-759-SIM is demonstrated in Figure 4, and the derived WHEBD-759 consists of 759 image classes and 699,815 samples in total.

6.3 Implementations

6.3.1 Setup of benchmark image classification

In the training, we used fixed hyper-parameters and a batch size of 128. In the test, we built conventional classifiers on the high-level vector representations extracted from the previous conv or fc network and then evaluated the performance of models by averaging the results of the last 40 epochs of optimizations. Here, we selected three views: the Red-Green-Blue (RGB) view of the original image, the luminance channel (L) view, and the ab-color channel (ab) view. Hence, in order to improve the discriminability of the learned features, we adopted a series of data augmentation methods: color jittering, random grayscale, and random cropping. We uniformly adopted the MI estimator based on Noise-Contrastive Estimation [3], [9].

In the comparisons of Table 1, 2, and 4, the encoder function $f_{\vartheta, \omega}$, i.e., $f_{\vartheta}(f_{\omega}(\cdot))$ that includes the feature extraction network f_{ω} and the mapping network f_{ϑ} , is approximated by a designed Alexnet [52] for the classification tasks. Inspired by the backbone splitting setting of SplitBrain [25], we evenly split the Alexnet into sub-networks across the channel dimension and then used the sub-networks as encoders. According to the principle of building the encoders, the Alexnet is split across the channel dimension with a conjecture that split-Alexnet can also perform well in learning representations between views, and the split-Alexnet only has the halved learnable parameters [25]. We, therefore, built the Alexnet with five convolutional layers (attached with additional batchnorm layers, ReLU activation functions, and corresponding maxpool functions), two linear layers (with corresponding batchnorm layers and ReLU activation functions), and a fully connected layer followed by an l2 normalization function, which is to tackle the problem of distribution drift. Then the split-Alexnets (i.e., the sub-networks) are served as the encoders. In the experiments, we used the convolutional (conv) neural network and the fully-connected (fc) network as the encoders, which use respectively the layers of Alexnet as the encoders, i.e., conv has 5 convolutional layers and one fully connected layer, and fc is the complete Alexnet. For the sake of further exploring the influence of the network architecture on the performance of CoCoNet and the ablation models, we chose conv_n as the backbone network, where n denotes the number of the convolutional layers. The comparisons on the large-scale ImageNet [1] are shown in Table 2.

In addition, we conducted extended experiments on the CIFAR10 and STL-10 datasets, and Table 3 depicts the results. We followed the experimental settings of the CPC and DIM

experiment [3], and then a stroke crop architecture [29] (i.e., eight \times eight crops with four \times four strides on the CIFAR10 dataset, and 16 \times 16 crops with eight \times eight strides on the STL-10 dataset) is adopted. We chose ResNet-50 [56] architecture as the encoder $f_{\vartheta, \omega}$, and the same classifier as in Table 1 is used, which is conducted to study the performance of CoCoNet based on a different backbone network.

For a fair comparison, all benchmark datasets use backbone encoders without pretraining. In training, we held the perspective that the representations learned the crucial features of views through different encoders. Then, we directly concatenated representations layer-wise from the encoders into one to achieve the ultimate representation of an input sample. The classifier’s development leverages a basic Multi-Layer Perception network (MLP) followed by the softmax output function. All downstream classification tasks are subject to the classifiers (i.e., the linear networks) on the high-level vector representations extracted from the designed encoders. For building the discrepancy metric calculation critic network based on generalized sliced Wasserstein distance (i.e., the critic network), the discrepancy metric of CoCoNet measures the differences between views in the learned latent space. Hence, the critic network is designed based on MLP [57].

6.3.2 Setup of benchmark graph prediction

To evaluate our method on self-supervised graph classification and regression tasks. We combined CoCoNet with GraphCL [58], which is a benchmark graph contrastive learning method. In detail, given a graph $G = \{G_i | i \in N\}$, two augmented graphs $\tilde{G}_i^1, \tilde{G}_i^2 : \tilde{G}_i \sim \text{aug}(\tilde{G}_i | G_i)$ are generated by random graph augmentations [58], [59]. G_i, \tilde{G}_i^1 , and \tilde{G}_i^2 are treated as three views of a graph. We compared CoCoNet with 4 baselines, and the reason we selected such baselines is that the experimental results of InfoGraph [60] and GraphCL [58] show that they achieve the state-of-the-art and outperform graph kernel and network embedding approaches [61], [62], [63], [64], [65], [66]. We followed the experimental protocol of GraphCL and AD-GCL. The average classification accuracy and standard deviation of the test results over the 20 runs are reported in Table 5. For a fair comparison, baselines and our methods adopt GIN as the encoder and use a downstream linear classifier or regressor with the same hyperparameters. To perform the ablation study, we constructed LoCo and two GloCo variants by combining GloCo with GraphCL and AD-GCL.

6.3.3 Setup of benchmark action recognition

We combined CoCoNet with CMC [9], and then evaluate the performance of the combination variant on benchmark action recognition tasks by following the experimental setting of [9], [67]. We trained our methods on UCF-101 [54] by using CaffeNets [68] to learn features from images and optical flows. Two streams are applied in the method: 1) the ventral stream, which performs object recognition and connects the target frame (image) of a video stream with a neighbouring frame; 2) the dorsal stream, which processes motion and associates the target frame to optical flow (centered at the target frame) in video data. In the training, we adopted both ventral and dorsal streams, which can be treated as two

views, and the target frame (image) in a video stream is the third view. In the test, the compared methods are tested on UCF-101 to evaluate the *task* transferability and on HMDB-51 [55] to evaluate the *task* and *dataset* transferability. It is worthy to note that, we performed CoCoNet based on the reimplemented CMC, i.e., CMC*.

6.3.4 Setup of practical object detection

In a real application, many newly added samples of CWD are unlabeled, and only a few samples are labeled per category on the procedure. Therefore, we introduced the proposed self-supervised method to enhance the performance of the main detection and classification models, e.g., B-CNN [69] and Faster RCNN [70], in a semi-supervised manner.

For the experimental settings, we trained the main models by fully utilizing the labeled samples of WHEBD-759 as the control group (the results are manifested in the first 2 rows of Table 7), which is denoted as $X = \{(x_1, y_1), (x_2, y_2), \dots, (x_{|X|}, y_{|X|})\}$. In order to simulate the real scene of this application, we generated a general simulated dataset, i.e., WHEBD-759-SIM, from the original labeled dataset, where only a quarter of the labeled data is retained for each category, and the labels of the remaining data are discarded. Hence, WHEBD-759-SIM consists of a labeled set $L = \{(x_1^L, y_1^L), (x_2^L, y_2^L), \dots, (x_{|L|}^L, y_{|L|}^L)\}$, and an unlabeled set $U = \{x_1^U, x_2^U, \dots, x_{|U|}^U\}$. We trained plain main models on WHEBD-759-SIM as another compared group. For the sake of taking advantage of the unlabeled data of U , we introduced the SSL methods, as the auxiliary methods, into the main models to form the integral semi-supervised learning methods. In details, the alternative SSL methods includes our proposed CoCoNet and the state-of-the-art self-supervised methods, for instance, SimCLR [4], SwAV [5], and CMC [9]. The supervised methods contains B-CNN [69] and Faster-RCNN [70], and both of VGG-16 [71] and VGG-19 [71] are selected as the alternative encoders. We used the self-supervised methods to pretrain the backbone networks and then leveraged the supervised methods to train the model.

6.4 Results and discussion

6.4.1 Comparisons on benchmark image classification

We extensively evaluated our proposed CoCoNet method on several benchmark datasets and tasks against the state-of-the-art methods. Table 1 shows the comparison results on the CIFAR10, CIFAR100, Tiny ImageNet, and STL-10 benchmark datasets respectively. The last 4 rows of tables represent the results of our proposed methods. Specifically, GloCo+SwAV, GloCo+CMC, and LoCo are the ablation models designed to eliminate different parts’ influence. In general, CoCoNet outperforms all models presented here by a significant margin when using the benchmark datasets. CoCoNet even outperforms the fully-supervised classifier without fine-tuning for the specific architectures presented, which shows that the representations learned by CoCoNet are better than the original images. However, in different experimental settings, we found that a designed fully-supervised classifier can outperform the state-of-the-art methods by a wider margin. Meanwhile, when more powerful backbone networks are used as encoders and specific data augmentations are adopted, the approaches perform better

Table 1: Performance of top-1 classification accuracy (%) on the CIFAR10, CIFAR100, Tiny ImageNet, and STL-10 datasets. In the experiments, we evaluated CoCoNet and the ablation models. Fully-supervised classification results are provided for comparison. † indicates that the results are reproduced by our reimplement. For a fair comparison, we adopted the same backbone networks with benchmarks. Note that the results of SplitBrain and CMC on STL-10 are reported in [9].

Model	CIFAR10			CIFAR100			Tiny ImageNet			STL-10		
	conv	fc	Average	conv	fc	Average	conv	fc	Average	conv	fc	Average
Fully supervised	75.39			42.27			36.60			68.70		
VAE [19]	60.71	60.54	60.63	37.21	34.05	35.63	18.63	16.88	17.76	58.27	56.72	57.50
AE [18]	62.19	55.78	58.99	31.50	23.89	27.70	19.07	16.39	17.73	58.19	55.57	56.88
β -VAE [20]	62.40	57.89	60.15	32.28	26.89	29.59	19.29	16.77	18.03	57.15	55.14	56.15
AAE [22]	59.44	57.19	58.32	36.22	33.38	34.80	18.04	17.27	17.66	59.54	54.47	57.01
BiGAN [24]	62.57	62.74	62.66	37.59	33.34	35.47	24.38	20.21	22.30	71.53	67.18	69.36
NAT [28]	56.19	51.29	53.74	29.18	24.57	26.88	13.70	11.62	12.66	64.32	61.43	62.88
SplitBrain† [25]	77.56	76.80	77.18	51.74	47.02	49.38	32.95	33.24	33.10	72.35	63.15	67.75
DIM [3]	73.25	73.62	73.44	48.13	45.92	47.03	33.54	36.88	35.21	72.86	70.85	71.86
SimCLR† [4]	80.58	80.07	80.33	50.03	49.82	49.93	36.24	39.83	38.04	75.57	77.15	76.36
SwAV† [5]	66.18	69.23	67.71	50.87	51.23	51.05	39.56	38.87	39.22	70.32	71.40	70.86
CMC† [9]	81.31	83.28	82.30	58.13	56.72	57.43	41.58	40.11	40.85	83.03	85.06	84.05
GloCo+SwAV	74.63	73.58	74.11	57.09	55.21	56.15	40.20	41.02	40.61	72.38	71.06	71.72
GloCo+CMC	82.27	82.95	82.61	59.02	57.38	58.20	42.21	39.62	40.92	84.12	85.03	84.58
LoCo	82.74	82.31	82.53	57.86	58.29	58.08	42.74	40.94	41.84	82.63	83.75	83.19
CoCoNet	83.10	83.24	83.17	58.64	58.21	58.43	42.28	43.63	42.96	85.34	83.82	84.58

Table 2: Performance of top-1 classification accuracy (%) on the ImageNet dataset. We followed the experiments of CMC [9], and we further reimplemented SimCLR and SwAV based on the same backbone networks.

Model	ImageNet					
	conv1	conv2	conv3	conv4	conv5	Average
Fully supervised	19.3	36.3	44.2	48.3	50.5	39.7
Context [72]	16.2	23.3	30.2	31.7	29.6	26.2
Colorization [73]	13.1	24.8	31.0	32.6	31.8	26.7
Jigsaw [74]	19.2	30.1	34.7	33.9	28.3	29.2
BiGAN [24]	17.7	24.5	31.0	29.9	28.0	26.2
SplitBrain [25]	17.7	29.3	35.4	35.2	32.8	28.7
Counting [75]	18.0	30.6	34.3	32.5	25.7	28.2
Inst-Dis [76]	16.8	26.5	31.8	34.1	35.6	29.0
RotNet [77]	18.8	31.7	38.7	38.2	36.5	32.8
DeepCluster [78]	12.9	29.2	38.2	39.8	36.1	32.2
DIM [3]	14.5	24.9	29.1	32.4	35.9	27.4
SimCLR† [4]	15.9	22.4	34.5	34.0	37.7	28.9
SwAV† [5]	13.6	23.8	32.2	27.3	38.0	27.0
CMC [9]	18.4	33.5	38.1	40.4	42.6	34.6
GloCo+SwAV	16.8	28.5	33.7	26.2	34.9	28.0
GloCo+CMC	17.7	36.2	39.6	41.1	43.0	35.5
LoCo	17.9	34.4	38.4	38.6	43.7	34.6
CoCoNet	18.2	36.3	39.8	40.5	43.8	35.7

on the benchmark datasets (albeit in different settings, e.g., AMDIM [7]). However, these approaches leverage different and deeper networks as their backbone encoders, so we excluded these methods from the benchmarks. Hence, the ablation models, i.e., GloCo+SwAV, GloCo+CMC, and LoCo, outperform most state-of-the-art approaches on all datasets. Yet, our proposed methods only outperform CMC with a small advantage. After comparison, we found out that CMC adopts a specialized architecture with carefully-chosen data augmentations, and in general, our proposed GloCo can additionally enhance CMC, e.g., GloCo+CMC outperforms CMC. To our knowledge, in the field of unsupervised learning, the results of CoCoNet are state-of-the-art following

Table 3: Performance of top-1 classification accuracy (%) on the CIFAR10 and STL-10 datasets. We compared the proposed method with the state-of-the-art unsupervised methods. We adopted ResNet-50 [56] as the encoders.

Model	CIFAR10	STL-10	Average
CPC [29]	77.45	77.81	77.63
DIM [3]	77.51	78.21	77.86
SwAV† [5]	83.15	82.93	83.04
SimCLR† [4]	84.63	83.75	84.19
CMC† [9]	86.10	86.83	86.47
GloCo+SwAV	84.62	85.81	85.22
GloCo+CMC	87.78	89.11	88.45
LoCo	89.06	88.21	88.64
CoCoNet	89.58	89.37	89.48

the proposed experimental settings. Specifically, the results support the proposed CoCoNet effectiveness to preserve the consistency of unlabeled data across views. As shown in Tables 1, the best results are in the last row, indicating that the feature representations learned by CoCoNet are discriminative.

Benchmarking CoCoNet on a large-scale dataset. As demonstrated in Table 2, the proposed CoCoNet has consistent performance even on a large benchmark dataset (e.g., ImageNet) within different network architectures. CoCoNet beats the state-of-the-art unsupervised method (e.g., CMC) by 1.1% on average. The performance of LoCo is on par with that of CMC, which demonstrates that our proposed local complementarity preserving module can improve the discriminability of the learned features by utilizing the complementarity-factor to capture the complementary information from different views. Furthermore, the ablation model GloCo+SwAV outperforms SwAV by 1.8%, and GloCo+CMC outperforms CMC by 0.9% respectively, which indicates that the global consistency preserving network enhances the baseline methods by aligning the distribution of views in the hidden space. We also observed that the performance of the compared methods is unstable within

weaker backbone networks, such as conv1 or conv2, and our consideration behind this phenomenon is that oversimplified networks do not have enough mapping capabilities to learn discriminative high-dimensional feature representations by utilizing complex self-supervised tasks.

Performing CoCoNet with ResNet. We performed extended classification comparisons on the CIFAR10 and STL-10 datasets with results shown in Table 3. We set the experiments by following the same principle of the CPC and DIM comparison [3]. The results show that CoCoNet and the experimental ablation models outperform state-of-the-art methods on the CIFAR10 and STL-10 datasets, respectively. Since ResNet-based encoders are adopted on the comparisons, and our proposed methods outperform the benchmarks, we reckoned that CoCoNet has strong adaptability to different encoders.

Evaluation with F1-Measure. In Table 4, we evaluated the compared methods with F1-Measure. We observed that CoCoNet is still state-of-the-art. It is widely acknowledged that there would be a big difference between the results of Accuracy and F1-Measure in the case of imbalanced sample categories (e.g., long-tail datasets). The benchmark datasets we conducted are all balanced datasets, including CIFAR10, CIFAR100, Tiny ImageNet, STL-10, and ImageNet. So, we considered that the results of our comparisons would not show a big difference, whether it is based on Accuracy or F1-Measure. However, in the case of imbalanced datasets, we have also included the F1-measure scores for the comparisons on the test sets of benchmark datasets to provide support for the superiority of our proposed CoCoNet, which is demonstrated in Table 4. Note that the comparisons are based on Macro F1-Measure, and the results indicate that CoCoNet can still achieve the best performance.

6.4.2 Comparisons on benchmark graph prediction

To evaluate the generalization of our proposed CoCoNet, we conducted comparisons in the field of graph prediction. As shown in Table 5, our methods beat the compared baselines on most benchmark downstream tasks. However, comparing the results of the combination variants of GloCo, e.g., GloCo + GraphCL, with the original baselines, e.g., GraphCL, we observed that the improvement is limited (or even arbitrary). We attributed such a phenomenon to the learning paradigm of our proposed GloCo, which prompts the model to capture *view-consistency* information by globally aligning the distributions of views. Such a process requires the data of different views to form *different* and *constant* distributions. Yet, we constructed views of graphs by using the *same* and *random* graph augmentations, which is contrary to our original intention of developing GloCo. So in the setting of conventional graph contrastive learning, GloCo’s poor improvement is understandable. The experimental results further support the effectiveness of CoCoNet and the ablation variant LoCo, which proves that our proposed *view-complementarity* information also applies to the graph self-supervised representation learning task, and mining such information can improve the benchmark methods. However, the effect of mining *view-complementarity*’s boost on the model is reduced, because randomly and inconsistently generated multiple views degenerate the performance of our proposed CoCoNet.

6.4.3 Comparisons on benchmark action recognition

To evaluate the effectiveness of CoCoNet on data of another modality, we conducted comparisons in the field of action recognition, which is based on video data. The results, reported in Table 6, support that our proposed methods can improve the performance of benchmark methods on the action recognition task of video data. Comparing the test results on UCF-101, we observed that CoCoNet and variants have remarkable *task* transferability, and comparing the test results on HMDB-51, we found that our methods have the good *task* and *dataset* transferability. Therefore, on the action recognition task of video data, our proposed *view-consistency* and *-complementary* information is valuable to mine in the paradigm of self-supervised multiview video representation learning, and CoCoNet can effectively model such information.

6.4.4 Comparisons on practical object detection

In order to deal with real-world issues, we further performed the proposed CoCoNet on a practical dataset against state-of-the-art methods. As demonstrated in Table 7, the comparison results on the practical dataset show that the methods in bold generally have better performance than other compared methods in the same experimental settings. The first 2 rows in the table represent the results of the pure supervised methods trained on the completely labeled WHEBD-759 dataset, and we observed that the models under such training strategy have the best performance, e.g., B-CNN trained on WHEBD-759 beats B-CNN w/ CoCoNet trained on WHEBD-759-SIM by 17.9%, and Faster-RCNN trained on WHEBD-759 beats Faster-RCNN w/ CoCoNet trained on WHEBD-759-SIM by 20.2% on average. This phenomenon indicates that the label information is important for models to learn discriminative representations, which is hard to be replaced by the self-supervised method. However, considering the fundamental idea that SSL can enhance the models to learn discriminative features, we used self-supervised methods to pretrain the encoders and then trained the supervised models on WHEBD-759-SIM to get better classification performance, and it is proved by the experiments. We first introduced the advanced self-supervised methods (e.g., SimCLR, SwAV, and CMC) to pretrain the encoders, and the results support that this pretraining procedure can improve the supervised models, where CMC has the best results, and, in detail, B-CNN w/ CMC beats the single B-CNN by 3.4%, and Faster-RCNN w/ CMC beats Faster-RCNN by 2.8% on average. Furthermore, the best results are always acquired by our proposed CoCoNet, for example, B-CNN w/ CoCoNet improves B-CNN w/ CMC by 1.5%, and Faster-RCNN w/ CoCoNet improves Faster-RCNN w/ CMC by 1.3% on average. The ablation models also have better performance than other self-supervised methods, which proves the effectiveness of the proposed method ulteriorly.

6.5 Deepgoing exploration

We conducted further experiments to explore the deep properties of CoCoNet.

6.5.1 CoCoNet with different settings of views

To validate whether CoCoNet has consistent performance under different settings of views, we selected several views

Table 4: Performance of classification (F1-Measure) on the CIFAR10, CIFAR100, STL-10, Tiny ImageNet, and ImageNet.

Model	CIFAR10		CIFAR100		Tiny ImageNet		STL-10		ImageNet
	conv	fc	conv	fc	conv	fc	conv	fc	conv
DIM [3]	0.7280	0.7276	0.4729	0.4435	0.3308	0.3461	0.7264	0.7042	0.3210
SwAV [5]	0.6578	0.6842	0.4990	0.4932	0.3808	0.3642	0.7000	0.7076	0.3407
SimCLR [4]	0.7980	0.7937	0.4856	0.4810	0.3508	0.3919	0.7542	0.7681	0.3445
CMC [9]	0.8101	0.8210	0.5753	0.5611	0.3952	0.3858	0.8280	0.8461	0.3918
GloCo + SwAV	0.7409	0.7324	0.5567	0.5338	0.3858	0.4056	0.7223	0.7039	0.3372
LoCo	0.8232	0.8203	0.5609	0.5746	0.3996	0.3983	0.8234	0.8308	0.4085
CoCoNet	0.8258	0.8241	0.5786	0.5745	0.4122	0.4239	0.8512	0.8347	0.4062

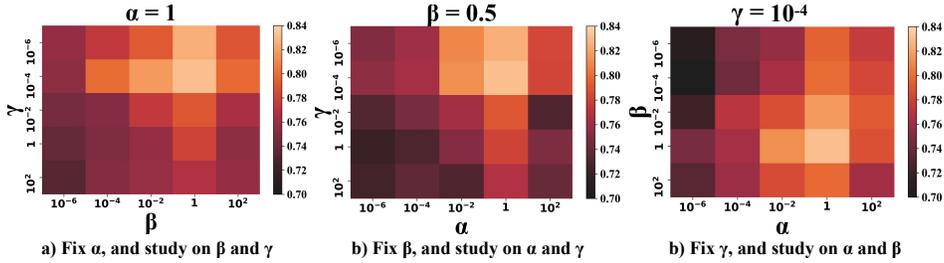
Figure 5: Influences of hyper-parameters α , β and γ of CoCoNet. We conducted comparisons on the CIFAR10 dataset.

Table 5: Performance of chemical molecules property prediction in OGB datasets, including two downstream tasks: graph regression and graph classification.

Model	mole	moll	molb	molt	mols
	Regression		Classification		
	(RMSE ↓)		(ROC-AUC% ↑)		
GIN RIU [79]	1.706	1.075	64.48	71.53	62.29
InfoGraph [60]	1.344	1.005	66.33	69.74	60.54
GraphCL [58]	1.272	0.910	68.22	72.40	61.76
AD-GCL [59]	1.270	0.926	68.26	71.08	61.83
GloCo + GraphCL	1.272	0.913	68.21	72.42	61.76
GloCo + AD-GCL	1.271	0.925	68.24	71.09	61.78
LoCo	1.268	0.910	68.49	71.32	61.81
CoCoNet	1.269	0.907	68.53	72.37	61.80

from the RGB optical (RGB) view, the luminance (L) view, and the ab-color (ab) view and conducted experiments on CIFAR10 using conv encoder. As manifested in the Table 8, CoCoNet has consistent performance and outperforms the compared methods on most tasks, and in details, CoCoNet beats the best benchmark method, i.e., CMC, by 0.96% with RGB and L views, by 0.34% with RGB and ab views, by 0.57% with L and ab views, and by 1.79% with all alternative views. We further added an experimental study on the comparison of the proposed CoCoNet and a typical contrastive learning method with more views. On the CIFAR10 dataset, we increased the number of views from 1 to 5 by sequentially adding the L, ab, RGB, Grayscale, and CbCr (belongs to YCbCr color space, where CB and Cr are the concentration offset components of blue and red) views. Results are shown in Figure 6. CoCoNet maintains its advantage over SimCLR when different view-settings are used for training. Compared with the addition of L, ab, and RGB, the additions of Grayscale and CbCr improve the performance of methods by a limited margin, and we considered the reason is that most information of Grayscale and CbCr is already contained by L, ab, and RGB. Concretely, our proposed consistency and complementarity regularization can indeed enhance the ability of the encoders to model multiple views, and such

Table 6: Action recognition accuracy (%) to evaluate *task* and *dataset* transferability on benchmark video datasets. We followed the setting of [9]. † denotes different network architecture. * denotes our reimplementation.

Method	Number of Views	UCF-101	HMDB-51
Random	-	48.2	19.5
ImageNet	-	67.7	28.0
VGAN [†] [80]	2	52.1	-
LT-Motion [†] [81]	2	53.0	-
TempCoh [82]	1	45.4	15.9
Shuffle and Learn [83]	1	50.2	18.1
Geometry [84]	2	55.1	23.3
OPN [85]	1	56.3	22.1
ST Order [86]	1	58.6	25.0
Cross and Learn [87]	2	58.7	27.2
CMC [9]	3	59.1	26.7
CMC*	3	58.8	26.3
GloCo + CMC	3	59.5	27.0
LoCo	3	59.2	26.8
CoCoNet	3	59.4	27.4

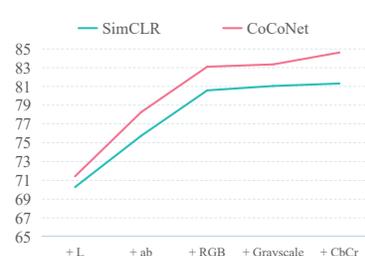


Figure 6: Comparisons with sequentially adding views on CIFAR10 using conv, which further indicates the superiority of CoCoNet over the compared baseline under different view-settings.

superiority is consistent under different view-settings.

6.5.2 Hyper-parameter heatmap

Specifically, we performed several experiments to study the influence of the tunable hyper-parameters. The hyper-parameter α balances the impact of the local complementarity preserving module. β balances the impact of conventional contrastive learning loss. γ balances the impact of the global consistency preserving module. To explore the in-

Table 7: Comparison of classification top-1 accuracies (%) on the real-world WHEBD-759-SIM dataset. We incorporated various SSL methods, e.g., SimCLR [4], SwAV [5], CMC [9], the proposed CoCoNet, and the ablation models of CoCoNet, into the main supervised detection and classification methods, e.g., B-CNN [69], and Faster-RCNN [70], to form the improved semi-supervised learning methods for the tasks, respectively. VGG-16 [71] and VGG-19 [71] are the backbone networks.

Training set	Supervised model	Self-supervised model	Backbone network		Average
			VGG-16	VGG-19	
WHEBD-759	B-CNN	N/A	87.7	86.2	87.0
	Faster-RCNN	N/A	86.3	90.2	88.3
WHEBD-759-SIM	B-CNN	N/A	63.7	64.6	64.2
		SimCLR	64.5	67.0	65.8
		SwAV	64.1	65.3	64.7
		CMC	66.5	68.7	67.6
		GloCo+SwAV	64.7	66.2	65.5
		GloCo+CMC	67.4	68.5	68.0
		LoCo	65.5	69.1	67.3
		CoCoNet	69.2	69.0	69.1
	Faster-RCNN	N/A	61.3	66.7	64.0
		SimCLR	61.8	67.4	64.6
		SwAV	63.3	66.8	65.1
		CMC	65.0	68.5	66.8
		GloCo+SwAV	63.6	67.0	65.3
		GloCo+CMC	65.9	69.6	67.8
	LoCo	64.3	68.4	66.4	
	CoCoNet	65.2	70.9	68.1	

Table 8: Comparison of applying different settings of views.

Views			Methods	Results
RGB	L	ab		
✓	✓		SimCLR [4]	76.37
			SwAV [5]	66.14
			CMC [9]	79.92
			CoCoNet	80.88
✓		✓	SimCLR [4]	74.30
			SwAV [5]	64.72
			CMC [9]	76.01
			CoCoNet	76.35
	✓	✓	SimCLR [4]	75.74
			SwAV [5]	65.90
			CMC [9]	77.69
			CoCoNet	78.26
✓	✓	✓	SimCLR [4]	80.58
			SwAV [5]	66.18
			CMC [9]	81.31
			CoCoNet	83.10

fluence of α and β , we fixed γ and selected α from the range of $\{10^{-6}, 10^{-4}, 10^{-2}, 1, 10^2\}$ and β from the range of $\{10^{-6}, 10^{-4}, 10^{-2}, 1, 10^2\}$. Following the same principle, we selected γ from the range of $\{10^{-6}, 10^{-4}, 10^{-2}, 1, 10^2\}$. As *a*), *b*), and *c*) shown in Figure 5, we observed that good classification performance is highly dependent on the local complementarity preserving module, i.e., α . An appropriate tuning of the impact of the contrastive loss, i.e., β , is needed for CoCoNet to enhance the cross-view feature discriminability. As such, the global consistency preserving module helps in classification performance with a small amount of γ , because it aligns the distribution of views, which helps to model the view-shared information.

6.5.3 CoCoNet with different discrepancy metrics

We conducted an ablation comparison by employing different discrepancy metrics for the proposed method. As shown in

Table 9, we directly replaced the discrepancy metric in GloCo module with KL, WD, etc. We observed that no matter which discrepancy metric is based on, GloCo + CMC can improve CMC, which proves the effectiveness of aligning the distributions of multiple views. Yet the improvements in taking different discrepancy metrics are inconsistent. Generally, the Wasserstein distance-based methods beat the KL divergence-based method, and we discussed the reasons in Section 5.2. Since directly calculating the high-dimensional Wasserstein distance is extremely computationally expensive, the difference between the Wasserstein distance-based methods is that the approaches to approximately calculate Wasserstein distances. In detail, WD uses the dual form of Wasserstein distance, yet the Lipschitz constraint is difficult to meet. SWD first obtains the one-dimensional representation of the high-dimensional probability distribution through linear mapping and then calculates the Wasserstein distance of the one-dimensional representation of the two probability distributions. Likewise, GSWD uses a similar approach except that generalized nonlinear mapping is used instead of linear mapping. The results demonstrate that, in the setting of multi-view learning, GSWD can retain more discriminative information than SWD in dimensionality reduction.

6.5.4 Validating the effectiveness of modeling low-level information of the LoCo module

To decouple the design of the architecture and the design of the learning objective, we conducted a further exploration with CoCoNet, LoCo, and an ablation model HLoCo by removing the low-level feature maps from LoCo, i.e., HLoCo only uses the high-level feature vectors, not the information of low-level feature maps. As shown in Figure 8 and Table 1, we observed that HLoCo beats the baselines with the same high-level representations, which shows the effectiveness of the learning objective \mathcal{L}_{LoCo} . Moreover, both LoCo and CoCoNet can outperform HLoCo on benchmark datasets,

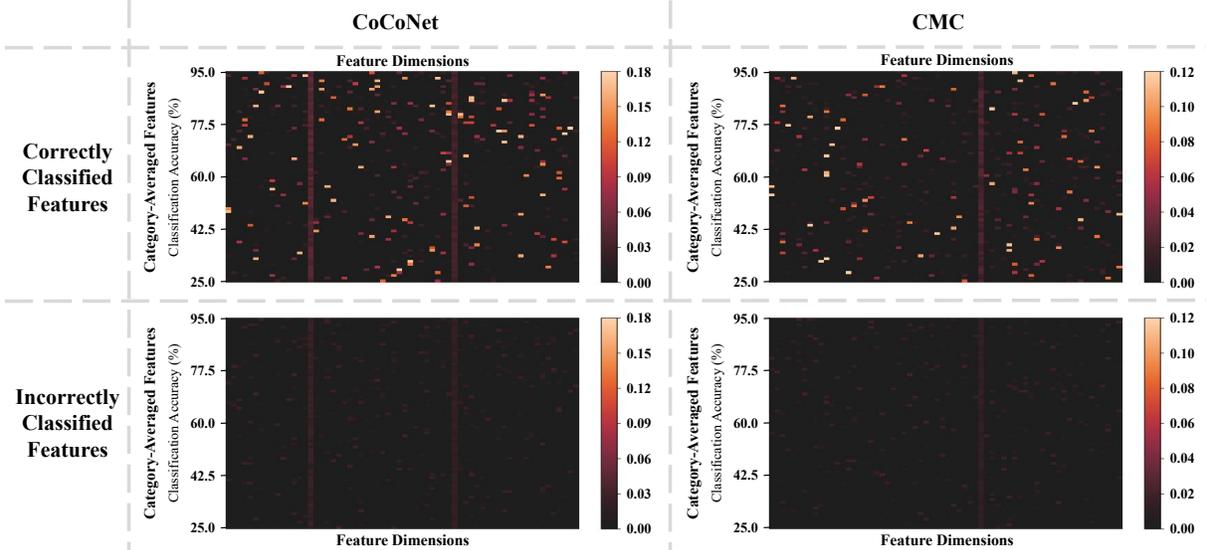


Figure 7: Visual comparisons of the top category-averages features of correct and incorrect classifications in the representation space of CoCoNet and CMC. We derived averaged features for each category, and according to the classification results, we retrieved the top category-averages features to evaluate the *activated* feature elements of CoCoNet and CMC, which is conducted on the Tiny ImageNet dataset by following the experimental principle of [88]. We observe that the correct classification contains specific activated feature elements that are more salient (colorful) than other feature elements, whereas the incorrect classifications do not.

Table 9: Classification top-1 accuracy (%) on the CIFAR10 and Tiny ImageNet datasets. We conducted several experiments based on the *conv* encoder and classifier as in Table 1. We introduced the optional discrepancy metrics, e.g., KL-divergence (KL) [89], Wasserstein distance (WD) [46], sliced Wasserstein distance (SWD) [90], and generalized sliced Wasserstein distance (GSWD) [47], to GloCo + CMC. Notably, the GSWD-based GloCo + CMC outperforms benchmark methods but falls short compared to CoCoNet.

Model	CIFAR10	Tiny ImageNet	Average
CMC	81.31	41.58	61.45
GloCo + CMC w/ KL	81.62	42.08	61.85
GloCo + CMC w/ WD	82.02	42.14	62.08
GloCo + CMC w/ SWD	82.07	42.05	62.06
GloCo + CMC w/ GSWD	82.27	42.21	62.24
CoCoNet	83.10	42.28	62.69

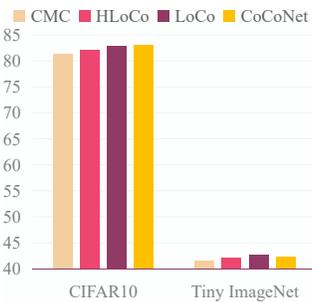


Figure 8: Research on the effectiveness of a specific sub-structure of the LoCo module. Specifically, we decoupled modeling the low-level feature information from learning the complementarity-factor CF and evaluate the ablation model.

indicating that modeling the discriminative information from low-level feature maps can generally improve the performance of our method. The reason behind such a phenomenon is that from the perspective of the information theory, compared with the low-level feature map, the high-level feature vector may lose some complementarity information. Therefore, according to the amount of information entropy, HLoCo, like typical contrastive learning methods [4], [5], [9],

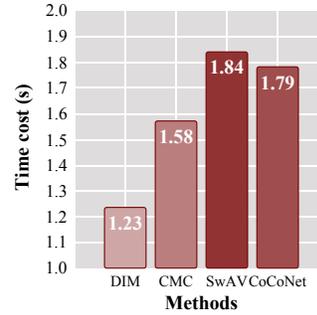


Figure 9: The average computational time costs of the training of a batch during the first 20 epochs. The process includes the feed-forward calculation and the back-propagation training of the encoders.

only contains the information of high-dimensional feature vectors, while the useful complementarity information may be lost within the encoding process so that compared with LoCo and CoCoNet, complementarity information is not sufficiently explored by HLoCo. We further observed that comparing HLoCo and LoCo, LoCo improves HLoCo by a larger margin on Tiny ImageNet than on CIFAR10. We reckoned that the classification on Tiny ImageNet requires more complementarity information, since Tiny ImageNet contains 200 categories while CIFAR10 only contains 10 categories.

6.5.5 Case study

As shown in Figure 7, each class has a discriminatory set of feature elements that contribute to correct classifications, i.e., each category requires different sets of discriminative feature elements for classifications. The observations on the incorrect classifications demonstrate that the over-consistency (trivial) of feature elements is a crucial reason that the feature cannot be correctly classified, indicating that the misclassified features model much task-irrelevant information. We observe that compared with CMC, CoCoNet learns features with more salient elements, indicating that CoCoNet can learn more activated feature elements for each category in classifications. The reason behind such an observation is that CMC

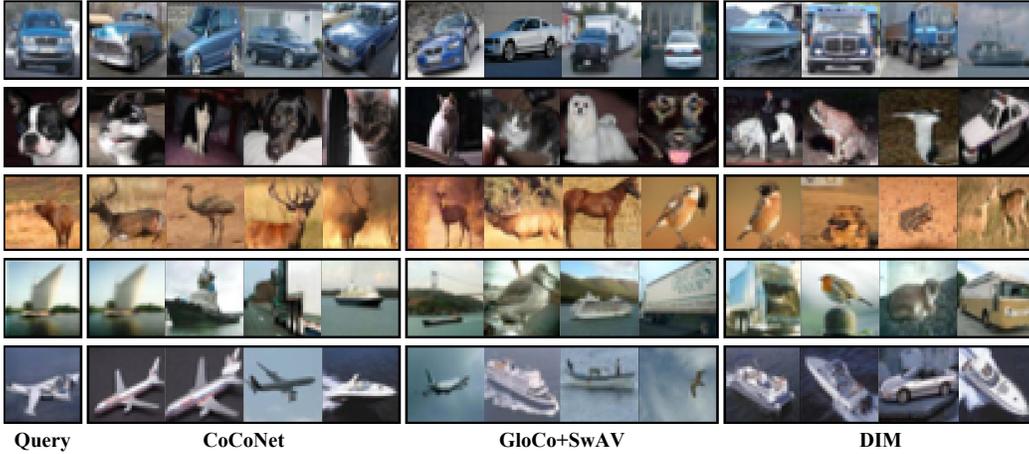


Figure 10: Visual comparisons for studying the merits of CoCoNet on the CIFAR10 dataset. We retrieved the 4 nearest neighbors to evaluate the discriminability by using L_1 distance. The leftmost images are randomly selected images as queries, and the other images are their nearest neighbors measured in the representations of compared methods.

only leverages the typical contrastive approach to model consistency information, while CoCoNet further imposes the proposed LoCo module to extract complementarity information. Hence, in addition to the feature elements modeling consistency information, the feature elements modeling complementary information can also contribute to the classification of each category. The extra elements in the features learned by our method can be regarded as the elements modeling complementary information. Therefore, a larger amount of discriminative information can empower CoCoNet to be robust to task-irrelevant noisy information, resulting in better performance on downstream tasks.

6.5.6 Limitations and discussion

Discussion on the time complexity. In head-to-head comparisons, CoCoNet achieves the state-of-the-art, which supports that mining view-consistency and -complementarity knowledge can improve to model multiple views in multi-view SSL. Yet compared with benchmark self-supervised methods, CoCoNet has relatively higher time complexity in training. As shown in Figure 9, the computational time cost of CoCoNet is lower than SwAV but higher than the baseline CMC. We reckoned the reasons are 1) the training of the critic network of GloCo; 2) the matrix operations of LoCo. However in the test, the compared methods adopt the same paradigm, and the test time complexities are the same.

Threats to validity [91]. For the *conclusion validity*, we followed the benchmark experimental settings [3], [9], e.g., choice of statistical tests, choice of sample size, etc. In order to avoid the threat to validity caused by imbalanced datasets, in addition to accuracy, we further adopted F1-Measure as a metric to measure the experiments, which is shown in Table 4. For the *internal validity*, we introduced sufficient ablation studies, demonstrated in Section 6.4, to prove the effectiveness of the proposed parts of CoCoNet, i.e., GloCo and LoCo. To further explore whether replacing specific components of CoCoNet with variants may affect the conclusion that “the improvement in results is due to the proposed method”, we conducted comparisons in Table 9 and Figure 8, and the results support the effectiveness of CoCoNet’s components. For the *construct validity*, a foundational assumption of multi-view SSL is stated in Assumption 5.1, which is theoretically

proved by [10], [48], [92]. Moreover, such an assumption is empirically proved by [4], [5], [9] to be applicable to image-related tasks, by [58] to be applicable to graph-related tasks, and by [9] to be applicable to video-related tasks. For the *external validity*, to avoid the influence of random factors (such as random seeds) in the experiment on the results, we collected the results of 5 trials for comparisons. The average result of the last 10 epochs is used as the final result of each trial. The average results from all trials are presented in tables. We conducted comparisons on multiple downstream tasks, including image classification tasks, graph prediction tasks, and action recognition tasks, to avoid artificial experimental settings that may affect the generalization of the model. In order to further verify whether the experiments on benchmark datasets can be generalized to actual real-world scenarios, we conducted comparisons on a practical dataset, i.e., WHEBD-759, and the results demonstrate that CoCoNet can still improve the performance of benchmark supervised methods in a self-supervised manner.

6.5.7 Visual comparisons

As shown in Figure 10, the representations learned by CoCoNet lead to more interpretable metric structures since neighboring representations correspond to visually similar images of the same category. There are three reasons for this circumstance: 1) CoCoNet learns representations from multiple views instead of a single view; 2) LoCo helps to refine the representations by improving the feature’s view-specific discriminability; 3) GloCo further enhances the learned representations’ view-shared discriminability by measuring the discrepancy metric between views.

7 CONCLUSIONS

This paper proposes a novel CoCoNet to mine discriminative knowledge from multi-view data in an unsupervised manner. To this end, CoCoNet globally aligns the distributions of views in the latent space by adopting an efficient alignment method based on GSWD, which helps to capture view-consistency information. CoCoNet leverages the proposed complementarity-factor to maintain the cross-view complementarity of the latent representations on the local

stage. Compared with the conventional methods, CoCoNet explores more, albeit still not full, discriminative information from multiple views. The provided theoretical and experimental analyses support the effectiveness of CoCoNet.

8 ACKNOWLEDGEMENTS

The authors would like to thank the associate editor and anonymous reviewers for their valuable comments. This work is supported in part by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDA19020500, National Natural Science Foundation of China No. 61976206 and No. 61832017, Key Special Project for Introduced Talents Team of Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), No. GML2019ZD0603, Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, Beijing Academy of Artificial Intelligence (BAAI), China Unicom Innovation Ecological Cooperation Plan, the Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University of China 21XNLG05, and Public Computing Cloud, Renmin University of China. This work is also supported in part by Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China, and Public Policy and Decision-making Research Lab of Renmin University of China.

REFERENCES

- [1] D. Jia, D. Wei, R. Socher, L. J. Li, L. Kai, and F. F. Li, "Imagenet: A large-scale hierarchical image database," *Proc of IEEE Computer Vision and Pattern Recognition*, pp. 248–255, 2009.
- [2] I. Belghazi, S. Rajeswar, A. Baratin, R. D. Hjelm, and A. C. Courville, "MINE: mutual information neural estimation," *CoRR*, vol. abs/1801.04062, 2018.
- [3] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. 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*, 2019.
- [4] T. Chen, S. Kornblith, M. Norouzi, and G. 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*, vol. 119. PMLR, 2020, pp. 1597–1607.
- [5] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," 2020.
- [6] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [7] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *NeurIPS 2019*, 2019, pp. 15 509–15 519.
- [8] C. Zhang, H. Fu, Q. Hu, X. Cao, Y. Xie, D. Tao, and D. Xu, "Generalized latent multi-view subspace clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 86–99, 2020.
- [9] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12356. Springer, 2020, pp. 776–794.
- [10] K. Sridharan and S. M. Kakade, "An information theoretic framework for multi-view learning," *Conference on Learning Theory*, 2008.
- [11] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [12] Bengio, Yoshua, Courville, Aaron, Vincent, and Pascal, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
- [13] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.
- [14] N. Kakuda, T. Miwa, M. Nagaoka, and T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1, 1998.
- [15] I. T. Jolliffe, "Principal component analysis," *Journal of Marketing Research*, vol. 87, no. 4, p. 513, 2002.
- [16] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," *Parallel Distributed Process*, vol. 1, 01 1986.
- [17] R. Salakhutdinov and G. E. Hinton, "Deep boltzmann machines," *Journal of Machine Learning Research*, vol. 5, no. 2, pp. 1967 – 2006, 2009.
- [18] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science (New York, N.Y.)*, vol. 313, pp. 504–7, 08 2006.
- [19] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.
- [20] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *CoRR*, vol. abs/1612.00410, 2016.
- [21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *ArXiv*, 06 2014.
- [22] A. Makhzani, J. Shlens, N. Jaitly, and I. J. Goodfellow, "Adversarial autoencoders," *CoRR*, vol. abs/1511.05644, 2015.
- [23] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19–24, 2016*, M. Balcan and K. Q. Weinberger, Eds., vol. 48. JMLR.org, 2016, pp. 1747–1756.
- [24] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24–26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.
- [25] R. Zhang, P. Isola, and A. A. Efros, "Split-brain autoencoders: Unsupervised learning by cross-channel prediction," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017*, 2017.
- [26] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.
- [27] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, and G. Brain, "Time-contrastive networks: Self-supervised learning from video," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018.
- [28] P. Bojanowski and A. Joulin, "Unsupervised learning by predicting noise," in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017*, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 2017, pp. 517–526.
- [29] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.
- [30] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. 12, pp. 3371–3408, 2010.
- [31] Z. Yang, Q. Xu, W. Zhang, X. Cao, and Q. Huang, "Split multiplicative multi-view subspace clustering," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 5147–5160, 2019.
- [32] H. Zhang, V. M. Patel, and R. Chellappa, "Hierarchical multimodal metric learning for multimodal classification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017*. IEEE Computer Society, 2017, pp. 2925–2933.
- [33] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1935.
- [34] S. Akaho, "A kernel method for canonical correlation analysis," *CoRR*, vol. abs/cs/0609071, 2006.
- [35] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "On deep multi-view representation learning," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6–11 July 2015*, F. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 1083–1092.
- [36] X. Chen, S. Chen, H. Xue, and X. Zhou, "A unified dimensionality reduction framework for semi-paired and semi-supervised multi-view data," *Pattern Recognit.*, vol. 45, no. 5, pp. 2005–2018, 2012.

- [37] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, 2021. [Online]. Available: <https://doi.org/10.1109/JPROC.2020.3004555>
- [38] M. Chen, K. Q. Weinberger, and J. Blitzer, "Co-training for domain adaptation," in *Advances in neural information processing systems*, 2011, pp. 2456–2464.
- [39] K. You, X. Wang, M. Long, and M. Jordan, "Towards accurate model selection in deep unsupervised domain adaptation," in *International Conference on Machine Learning*, 2019, pp. 7124–7133.
- [40] Y. Wu, E. Winston, D. Kaushik, and Z. Lipton, "Domain adaptation with asymmetrically-relaxed distribution alignment," *arXiv preprint arXiv:1903.01689*, 2019.
- [41] H. Zhao, R. T. d. Combes, K. Zhang, and G. J. Gordon, "On learning invariant representation for domain adaptation," *arXiv preprint arXiv:1901.09453*, 2019.
- [42] F. M. Cariucci, L. Porzi, B. Caputo, E. Ricci, and S. R. Bulò, "Autodial: Automatic domain alignment layers," in *2017 IEEE International Conference on Computer Vision*, 2017.
- [43] K. Sohn, W. Shang, X. Yu, and M. Chandraker, "Unsupervised domain adaptation for distance metric learning," 2018.
- [44] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Advances in neural information processing systems*, 2007, pp. 137–144.
- [45] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *CoRR*, 2017.
- [46] S. Kuroki, N. Charoenphakdee, H. Bao, J. Honda, I. Sato, and M. Sugiyama, "Unsupervised domain adaptation based on source-guided discrepancy," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4122–4129.
- [47] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. K. Rohde, "Generalized sliced wasserstein distances," 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.
- [48] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *Computer Science*, 2013.
- [49] H. Narayanan and S. K. Mitter, "Sample complexity of testing the manifold hypothesis," in *International Conference on Neural Information Processing Systems*, 2010.
- [50] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Handbook of Systemic Autoimmune Diseases*, vol. 1, no. 4, 2009.
- [51] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," *Journal of Machine Learning Research*, vol. 15, pp. 215–223, 2011.
- [52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [53] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec, "Open graph benchmark: Datasets for machine learning on graphs," *Neural Information Processing Systems (NeurIPS)*, 2020.
- [54] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *Computer Science*, 2012.
- [55] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in *2011 International Conference on Computer Vision*, 2011, pp. 2556–2563.
- [56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778.
- [57] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *biol math biophys*, 1943.
- [58] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5812–5823, 2020.
- [59] S. Suresh, P. Li, C. Hao, and J. Neville, "Adversarial graph augmentation to improve graph contrastive learning," *arXiv preprint arXiv:2106.05819*, 2021.
- [60] F.-Y. Sun, J. Hoffman, V. Verma, and J. Tang, "Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization," in *International Conference on Learning Representations*, 2020.
- [61] N. M. Kriege, F. D. Johansson, and C. Morris, "A survey on graph kernels," *Applied Network Science*, vol. 5, no. 1, pp. 1–42, 2020.
- [62] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855–864.
- [63] B. Adhikari, Y. Zhang, N. Ramakrishnan, and B. A. Prakash, "Sub2vec: Feature learning for subgraphs," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2018, pp. 170–182.
- [64] P. Yanardag and S. Vishwanathan, "Deep graph kernels," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1365–1374.
- [65] A. Narayanan, M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu, and S. Jaiswal, "graph2vec: Learning distributed representations of graphs," *arXiv preprint arXiv:1707.05005*, 2017.
- [66] N. Shervashidze, P. Schweitzer, E. J. Van Leeuwen, K. Mehlhorn, and K. M. Borgwardt, "Weisfeiler-lehman graph kernels." *Journal of Machine Learning Research*, vol. 12, no. 9, 2011.
- [67] Z. Christopher, P. Thomas, and B. Horst, "A duality based approach for realtime tv-l 1 optical flow," *Joint pattern recognition symposium*, 2007.
- [68] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [69] T. Y. Lin, A. Roychowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," 2015.
- [70] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
- [71] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv*, 2014.
- [72] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction." IEEE Computer Society, 2015.
- [73] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," *European Conference on Computer Vision*, 2016.
- [74] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Springer, Cham*, 2016.
- [75] M. Noroozi, H. Pirsivash, and P. Favaro, "Representation learning by learning to count," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [76] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, *Unsupervised feature learning via non-parametric instance discrimination*, 2018.
- [77] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," 2018.
- [78] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," *European Conference on Computer Vision*, 2018.
- [79] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *International Conference on Learning Representations*, 2018.
- [80] C. Vondrick, H. Pirsivash, and A. Torralba, "Generating videos with scene dynamics," *arXiv*, 2016.
- [81] Z. Luo, B. Peng, D. A. Huang, A. Alahi, and F. F. Li, "Unsupervised learning of long-term motion dynamics for videos," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [82] H. Mobahi, R. Collobert, and J. Weston, "Deep learning from temporal coherence in video," in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, ser. ACM International Conference Proceeding Series, 2009.
- [83] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: Unsupervised learning using temporal order verification," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, 2016.
- [84] C. Gan, B. Gong, K. Liu, S. Hao, and L. J. Guibas, "Geometry guided convolutional neural networks for self-supervised video representation learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [85] H. Y. Lee, J. B. Huang, M. Singh, and M. H. Yang, "Unsupervised representation learning by sorting sequences," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [86] U. Büchler, B. Brattoli, and B. Ommer, "Improving spatiotemporal self-supervision by deep reinforcement learning," 2018.
- [87] N. Sayed, B. Brattoli, and B. Ommer, "Cross and learn: Cross-modal self-supervision," *Springer, Cham*, 2018.
- [88] N. M. Kalibhat, K. Narang, L. Tan, H. Firooz, M. Sanjabi, and S. Feizi, "Understanding failure modes of self-supervised

learning," *CoRR*, vol. abs/2203.01881, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2203.01881>

- [89] 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.
- [90] C.-Y. Lee, T. Batra, M. H. Baig, and D. Ulbricht, "Sliced wasserstein discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 285–10 295.
- [91] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, and B. Regnell, *Experimentation in Software Engineering*. Springer, 2012. [Online]. Available: <https://doi.org/10.1007/978-3-642-29044-2>
- [92] Y. Wang, Q. Zhang, Y. Wang, J. Yang, and Z. Lin, "Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap," *arXiv preprint arXiv:2203.13457*, 2022.
- [93] S. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.



Jiangmeng Li received the BS degree in the department of software engineering, Xiamen University, Xiamen, China, in 2016, and the MS degree from New York University, New York, USA, in 2018. He is currently a doctoral student at the University of Chinese Academy of Sciences. His research interests include self-supervised learning, deep learning, and machine learning. He has published more than five papers in journals and conferences such as IEEE Transactions on Knowledge and Data Engineering (TKDE),

International Conference on Machine Learning (ICML), International Joint Conference on Artificial Intelligence (IJCAI), etc.



Wenwen Qiang received the MS degree in the department of mathematics, college of science, China Agricultural University, Beijing, in 2018. He is currently a doctoral student at the University of Chinese Academy of Sciences. His research interests include transfer learning, deep learning, and machine learning. He has published more than five papers in journals and conferences such as IEEE Transactions on Knowledge and Data Engineering (TKDE), International Conference on Machine Learning (ICML), International Joint

Conference on Artificial Intelligence (IJCAI), etc.



Changwen Zhen received the Ph.D. degree in Huazhong University of Science and Technology. He is currently a professor in Institute of Software, Chinese Academy of Science. His research interests include computer graph and artificial intelligence.



Bing Su received the BS degree in information engineering from the Beijing Institute of Technology, Beijing, China, in 2010, and the PhD degree in electronic engineering from Tsinghua University, Beijing, China, in 2016. From 2016 to 2020, he worked with the Institute of Software, Chinese Academy of Sciences, Beijing. Currently, he is an associate professor with the Gaoling School of Artificial Intelligence, Renmin University of China. His research interests include pattern recognition, computer vision, and machine learning. He has

published more than ten papers in journals and conferences such as IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), IEEE Transactions on Image Processing (TIP), International Conference on Machine Learning (ICML), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE International Conference on Computer Vision (ICCV), etc.



Farid Razzak received the Ph.D. degree from the Management Science & Information Systems Department at Rutgers, the State University of New Jersey in 2020, the M.S. degree from the School of Professional Studies at New York University in 2009, and received a B.B.A degree from Bernard M. Baruch College at City University of New York in 2007. He is currently an Financial Quantitative Data Scientist for the Securities and Exchange Commission as well as adjunct faculty at New York University & Columbia University. His

research interests include applied data mining for financial regulations, business applications and services.



Ji-Rong Wen is a full professor at Gaoling School of Artificial Intelligence, Renmin University of China. He worked at Microsoft Research Asia for fourteen years and many of his research results have been integrated into important Microsoft products (e.g. Bing). He serves as an associate editor of ACM Transactions on Information Systems (TOIS). He is a Program Chair of SIGIR 2020. His main research interests include web data management, information retrieval, data mining and machine learning.



Hui Xiong received his Ph.D. in Computer Science from the University of Minnesota - Twin Cities, USA, in 2005, the B.E. degree in Automation from the University of Science and Technology of China (USTC), Hefei, China, and the M.S. degree in Computer Science from the National University of Singapore (NUS), Singapore. He is a chair professor at the Hong Kong University of Science and Technology (Guangzhou). He is also a Distinguished Professor at Rutgers, the State University of New Jersey, where he received the

2018 Ram Charan Management Practice Award as the Grand Prix winner from the Harvard Business Review, RBS Dean's Research Professorship (2016), two-year early promotion/tenure (2009), the Rutgers University Board of Trustees Research Fellowship for Scholarly Excellence (2009), the ICDM-2011 Best Research Paper Award (2011), the Junior Faculty Teaching Excellence Award (2007), Dean's Award for Meritorious Research (2010, 2011, 2013, 2015) at Rutgers Business School, the 2017 IEEE ICDM Outstanding Service Award (2017), and the AAAI-2021 Best Paper Award (2021). Dr. Xiong is also a Distinguished Guest Professor (Grand Master Chair Professor) at the University of Science and Technology of China (USTC). For his outstanding contributions to data mining and mobile computing, he was elected an ACM Distinguished Scientist in 2014, an IEEE Fellow and an AAAS Fellow in 2020. His general area of research is data and knowledge engineering, with a focus on developing effective and efficient data analysis techniques for emerging data intensive applications. He currently serves as a co-Editor-in-Chief of Encyclopedia of GIS (Springer) and an Associate Editor of IEEE Transactions on Data and Knowledge Engineering (TKDE), IEEE Transactions on Big Data (TBD), ACM Transactions on Knowledge Discovery from Data (TKDD) and ACM Transactions on Management Information Systems (TMIS). He has served regularly on the organization and program committees of numerous conferences, including as a Program Co-Chair of the Industrial and Government Track for the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), a Program Co-Chair for the IEEE 2013 International Conference on Data Mining (ICDM), a General Co-Chair for the IEEE 2015 International Conference on Data Mining (ICDM), and a Program Co-Chair of the Research Track for the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2018).

9 APPENDIX

9.1 Theoretical proofs

In Section 5, we propose two theorems: Theorem 5.1, i.e., the View-Consistency information with a potential loss of View-Specific Noise ϵ^{noise} theorem; Theorem 5.2, i.e., the View-Complementarity information, which is view-specific and task-relevant theorem. Here, we provide a formalized view to describe the proofs of them.

9.1.1 Proof of Theorem 5.1

We validate that $I(X^1; Y) \geq I(X^1; Y | \epsilon^{noise})$ by introducing the KL-divergence [93] measurement into the calculation of mutual information:

Proof. To proof $I(X^1; Y) \geq I(X^1; Y | \epsilon^{noise})$

$$\begin{aligned} \therefore I(X; Y) &= \sum_{x \in X} \sum_{y \in Y} \mathcal{P}(x, y) \log \frac{\mathcal{P}(x, y)}{\mathcal{P}(x) \cdot \mathcal{P}(y)} \\ \therefore I(X^1; Y) &= \sum_{x \in X^1} \sum_{y \in Y} \mathcal{P}(x, y) \log \frac{\mathcal{P}(x, y)}{\mathcal{P}(x) \cdot \mathcal{P}(y)} \\ \therefore I(X^1; Y | \epsilon^{noise}) &= \sum_{x \in X^1} \sum_{y \in \{Y - \epsilon^{noise}\}} \mathcal{P}(x, y) \log \frac{\mathcal{P}(x, y)}{\mathcal{P}(x) \cdot \mathcal{P}(y)} \end{aligned}$$

And KL-divergence is defined as:

$$D_{KL}(P||Q) = \int \mathcal{P}(x) \log \frac{\mathcal{P}(x)}{\mathcal{Q}(x)} dx$$

The discrete form of KL-divergence is:

$$D_{KL}(P||Q) = \sum \mathcal{P}(x) \log \frac{\mathcal{P}(x)}{\mathcal{Q}(x)}$$

We try to use KL divergence to fit the calculation of mutual information, and the \mathcal{P} and \mathcal{Q} are approximated by:

$$\hat{\mathcal{P}}(x) = \mathcal{P}(x, y)$$

$$\hat{\mathcal{Q}}(x) = \mathcal{P}(x) \cdot \mathcal{P}(y)$$

Put $\hat{\mathcal{P}}(x)$ and $\hat{\mathcal{Q}}(x)$ into the above formula of the discrete KL-divergence:

$$D_{KL}(\mathcal{P}_{XY} || \mathcal{P}_X \mathcal{P}_Y) = \sum_{x \in X} \sum_{y \in Y} \mathcal{P}(x, y) \log \frac{\mathcal{P}(x, y)}{\mathcal{P}(x) \cdot \mathcal{P}(y)}$$

Then, we get:

$$D_{KL}(\mathcal{P}_{XY} || \mathcal{P}_X \mathcal{P}_Y) = I(X; Y)$$

$$\therefore I(X^1; Y) = D_{KL}(\mathcal{P}_{X^1 Y} || \mathcal{P}_{X^1} \mathcal{P}_Y)$$

$$\therefore I(X^1; Y | \epsilon^{noise}) = D_{KL}(\mathcal{P}_{X^1 \{Y - \epsilon^{noise}\}} || \mathcal{P}_{X^1} \mathcal{P}_{\{Y - \epsilon^{noise}\}})$$

Because Y is not fully compressed, which means $I(X^1; Y | T) \geq 0$, it is acknowledged that $\epsilon^{noise} \geq 0$. For the KL-divergence, \mathcal{P}_{X^1} is constant, and $Y \geq \{Y - \epsilon^{noise}\}$. Therefore, compared with the joint $\mathcal{P}_{X^1 Y}$ and $\mathcal{P}_{X^1} \cdot \mathcal{P}_Y$, the distributions of the joint $\mathcal{P}_{X^1 \{Y - \epsilon^{noise}\}}$ and $\mathcal{P}_{X^1} \cdot \mathcal{P}_{\{Y - \epsilon^{noise}\}}$ are more consistent, and then we get:

$$D_{KL}(\mathcal{P}_{X^1 \{Y - \epsilon^{noise}\}} || \mathcal{P}_{X^1} \mathcal{P}_{\{Y - \epsilon^{noise}\}}) \leq D_{KL}(\mathcal{P}_{X^1 Y} || \mathcal{P}_{X^1} \mathcal{P}_Y)$$

$$\therefore I(X^1; Y) \geq I(X^1; Y | \epsilon^{noise})$$

□

9.1.2 Proof of Theorem 5.2

We validate that $I(Y; T) \leq I(Y; T) + I(X^1; Y^*; T | X^2) + I(X^2; Y^*; T | X^1)$ by introducing the KL-divergence [93] measurement into the calculation of mutual information:

Proof. To proof $I(Y; T) \leq I(Y; T) + I(X^1; Y^*; T | X^2) + I(X^2; Y^*; T | X^1)$

Transpose the mentioned equation:

$$I(Y; T) - I(Y; T) \leq I(X^1; Y^*; T | X^2) + I(X^2; Y^*; T | X^1)$$

$$I(X^1; Y^*; T | X^2) + I(X^2; Y^*; T | X^1) \geq 0$$

Since, we assume that Y^* is a extended representation of Y , and it can contain part of the View-Complementarity information, i.e., $I(X^1; T | X^2) + I(X^2; T | X^1)$. Therefore, we only need to proof that $I(X^1; T | X^2)$ or $I(X^2; T | X^1)$ is not null, because the mutual information cannot be negative. The proof is reformed to:

$$I(X^1; T | X^2) \geq 0$$

$$I(X^2; T | X^1) \geq 0$$

$$\therefore I(X; Y) = \sum_{x \in X} \sum_{y \in Y} \mathcal{P}(x, y) \log \frac{\mathcal{P}(x, y)}{\mathcal{P}(x) \cdot \mathcal{P}(y)}$$

$$\therefore I(X^1; T | X^2) = \sum_{x \in X^1} \sum_{y \in \{T - X^2\}} \mathcal{P}(x, y) \log \frac{\mathcal{P}(x, y)}{\mathcal{P}(x) \cdot \mathcal{P}(y)}$$

$$\therefore I(X^2; T | X^1) = \sum_{x \in X^2} \sum_{y \in \{T - X^1\}} \mathcal{P}(x, y) \log \frac{\mathcal{P}(x, y)}{\mathcal{P}(x) \cdot \mathcal{P}(y)}$$

As the equation deducing in Proof 9.1.1, we use the discrete form of KL divergence to fit the calculation of mutual information, and then we get:

$$D_{KL}(\mathcal{P}_{XY} || \mathcal{P}_X \mathcal{P}_Y) = I(X; Y)$$

$$\therefore I(X^1; T | X^2) = D_{KL}(\mathcal{P}_{X^1 \{T - X^2\}} || \mathcal{P}_{X^1} \mathcal{P}_{\{T - X^2\}})$$

$$\therefore I(X^2; T | X^1) = D_{KL}(\mathcal{P}_{X^2 \{T - X^1\}} || \mathcal{P}_{X^2} \mathcal{P}_{\{T - X^1\}})$$

The downstream task-relevant information T is not fully contained in any view of data, e.g., X^1 or X^2 , with a strong possibility meant for $H(T | X^1) \geq 0$ and $H(T | X^2) \geq 0$, and so, based on the view of KL-divergence, we reckon that $\mathcal{P}_{\{T - X^1\}}$ and $\mathcal{P}_{\{T - X^2\}}$ exist. For the KL-divergence, \mathcal{P}_{X^1} or \mathcal{P}_{X^2} is constant, and therefore it is very likely that $D_{KL}(\mathcal{P}_{X^1 \{T - X^2\}} || \mathcal{P}_{X^1} \mathcal{P}_{\{T - X^2\}})$ or $D_{KL}(\mathcal{P}_{X^2 \{T - X^1\}} || \mathcal{P}_{X^2} \mathcal{P}_{\{T - X^1\}})$ exists in like manner:

$$D_{KL}(\mathcal{P}_{X^1 \{T - X^2\}} || \mathcal{P}_{X^1} \mathcal{P}_{\{T - X^2\}}) \geq 0$$

$$D_{KL}(\mathcal{P}_{X^2 \{T - X^1\}} || \mathcal{P}_{X^2} \mathcal{P}_{\{T - X^1\}}) \geq 0$$

$$\therefore I(X^1; T | X^2) \geq 0 \text{ and } I(X^2; T | X^1) \geq 0$$

$$\therefore I(X^1; Y^*; T | X^2) \geq 0 \text{ and } I(X^2; Y^*; T | X^1) \geq 0$$

$$\therefore I(Y; T) \leq I(Y; T) + I(X^1; Y^*; T | X^2) + I(X^2; Y^*; T | X^1)$$

□

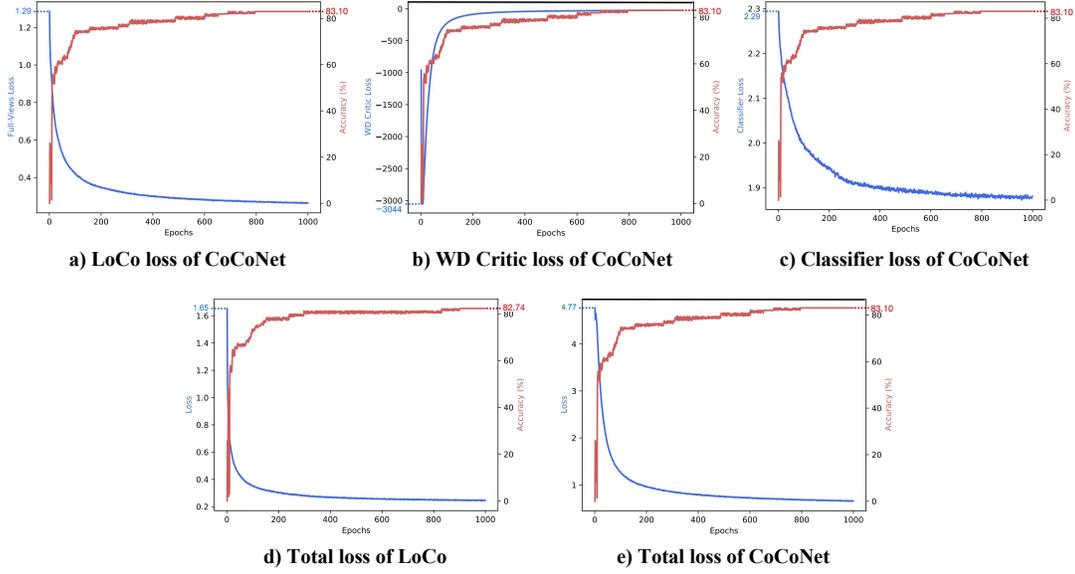


Figure 11: Extended verification for studying the loss convergence properties of CoCoNet and LoCo on the CIFAR10 dataset.

9.2 How does GSWD implement nonlinear mapping?

As mentioned in Section 4.1, GR_{ϑ} represents one-dimensional nonlinear projection operation on the probability measure P_r and P_g , which is defined as:

$$GR_{\vartheta}P_i(x_i) = \int_{\Sigma_i} P_i(x_i)\delta(t - ge(x_i, \vartheta)) dx_i \quad (10)$$

where $i \in \{r, g\}$, $\delta(\cdot)$ is the one-dimensional Dirac delta function, $t \in \mathbb{R}$, and $ge(\cdot, \vartheta)$ is a pre-defined nonlinear function that must satisfy the following four conditions:

- $ge(\cdot, \vartheta)$ is a real-valued C^∞ function.
- $ge(\cdot, \vartheta)$ is homogeneous of degree one in ϑ , i.e.,

$$\forall v \in R, ge(\cdot, v\vartheta) = vge(\cdot, \vartheta) \quad (11)$$

- $ge(\cdot, \vartheta)$ is non-degenerate in the sense that

$$\forall \vartheta \in \Omega_{\vartheta} \setminus \{0\}, x \in X, \frac{\partial ge}{\partial x}(x, \vartheta) \neq 0 \quad (12)$$

- The mixed Hessian of $ge(\cdot, \vartheta)$ is strictly positive, i.e.

$$\det\left(\left(\frac{\partial^2 ge}{\partial x_i \partial \vartheta_j}\right)_{i,j}\right) > 0 \quad (13)$$

$ge(\cdot, \vartheta)$ is a nonlinear function so that the GSWD achieves to map high-dimensional representations to one-dimensional representations in a nonlinear manner.

9.3 Extended comparisons

In this section, we conduct further experiments to study the intrinsic property of our proposed method.

9.3.1 Study on the Wasserstein distance changing trends

As *a)* and *b)* in Figure 12, they show the changing trends of the sum of Wasserstein distances between views in optimization. In *a)*, it is based on GloCo+CMC, and the result of CoCoNet is shown in *b)*. We found that, although the Wasserstein distance in both *a)* and *b)* can reach convergence, the Wasserstein distance in *b)* converges slower than in *a)*. Additionally, the Wasserstein distance's peak value in

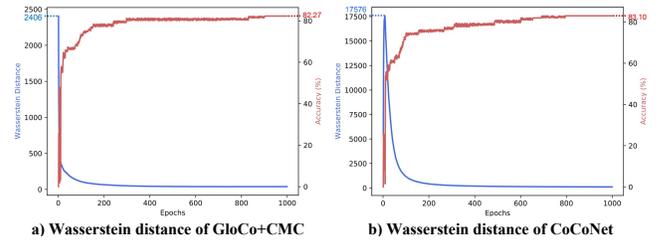


Figure 12: Extended verification of studying the Wasserstein distance changing trend properties of GloCo+CMC and CoCoNet in optimization on the CIFAR10 dataset.

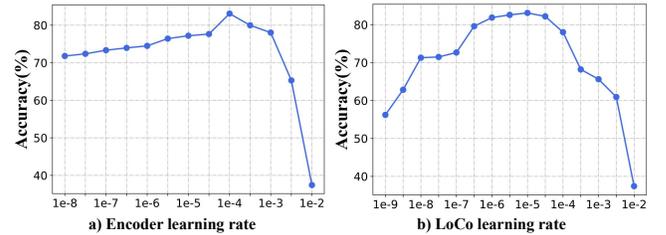


Figure 13: Extended verification of studying learning rates of the encoders and LoCo network in optimization on the benchmark CIFAR10 dataset.

b) is much higher than that in *a)*. The additional local complementarity preserving module affects the convergence of Wasserstein distance, and eventually, both the loss of the local complementarity preserving module and Wasserstein distance can converge. As such, the findings indicate that the game of simultaneously training the local complementarity and global consistency preserving modules is similar to the adversarial learning process, which helps CoCoNet to learn features with local complementarity and global consistency.

9.3.2 Study on optimizations

As manifested in Figure 13, we studied the learning rates of the encoders and the local complementarity preserving module (i.e., LoCo) respectively, while we excluded the learning rate parameter study of the global consistency

preserving network (i.e., GloCo) from the experiment, because we found that the learning rate of the GloCo has little effect on the performance of the proposed method. Furthermore, the objective of GloCo is to calculate the Wasserstein distances between views, so the learning rate of GloCo does not enhance the accuracy of CoCoNet. To explore the influence of different parts of CoCoNet, we selected wide ranges for the target learning rates, e.g., a range of $\{10^{-9}, 10^{-8}, \dots, 10^{-3}, 10^{-2}\}$ is for the uniform learning rate of the encoders, a range of $\{10^{-8}, 10^{-7}, \dots, 10^{-3}, 10^{-2}\}$ is for the learning rate of LoCo and fixed the other learning rates. We observed that the appropriate learning rates of the encoders and LoCo can promote the performance of our proposed method by a wide margin. Consider that the learning rate determines the step length of the weight iteration, so it is a very sensitive parameter. It has a significant effect on the model performance (e.g., the initial learning rate must have an optimal value. If it is too large, the model will not converge, and if it is too small, the model will converge slowly or fail to learn). Therefore, we concluded that the encoders and LoCo both have great impacts on the performance of CoCoNet.

9.3.3 Study on the loss convergence

From *a)*, *b)*, and *c)* in Figure 11, it can be found that all losses can reach convergence smoothly, which proves that the gradient descent of the loss of each part will not conflict with others in optimization. Moreover, it also verifies the integrity, robustness, and consistency of the proposed method.

We also conducted additional experiments to clarify the loss convergence of CoCoNet and the ablation model, i.e., LoCo. As in Figure 11, plots *d)* and *e)* separately show the relationships between the total loss and the accuracy based on LoCo and CoCoNet. We can find that in both of the optimization processes of LoCo and CoCoNet models, the total losses will eventually converge, while CoCoNet will be slightly slower to reach convergence during the training process. Meanwhile, as demonstrated in the classification comparisons, CoCoNet outperforms LoCo, which indicates that the global consistency preserving module can indeed enhance the performance of the proposed method. Hence, with the addition of GloCo, the initial loss tends to increase, and this shows that the additional module allows CoCoNet to have greater optimization potential.