INVERSE OPTIMAL TRANSPORT WITH APPLICATION TO CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Previous works in contrastive learning (CL) mainly focus on pairwise views to learn the representations by attracting the positive samples and repelling negative ones. In this work, we understand the CL with a collective point set matching view and solve this problem with the formulation of inverse optimal transport(IOT), which is a min-min optimization to learn the features. By varying the relaxation degree of constraints in inner minimization of IOT, one can naturally get three different contrastive losses and reveal that InfoNCE is a special case of them, which shows a new and more generalized understanding view of CL. Besides, with our soft matching view, a uniformity penalty is also proposed to improve the representation learning. Experimental results show the effectiveness of our methods.

1 INTRODUCTION

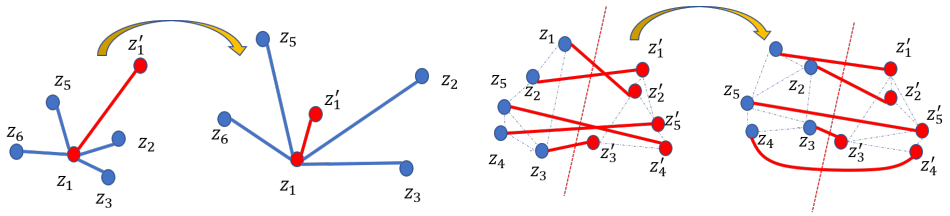
Unsupervised/self-supervised learning of representation has received increasing attention, whose frontier is advanced by contrastive learning (CL) (Hu et al., 2021; Grill et al., 2020). In mainstream CL methods (Chen et al., 2020; Gao et al., 2021), the representation is learned by first identifying one anchor and then looking for its positive/negative samples, whereby the contrastive loss based on feature similarity is adopted to discriminate positive and negative pairs. However, the comparisons of the positive and negative pairs behave often empirically and the popular contrastive loss (i.e. InfoNCE) has some disagreements by interpreting with the lower bound of mutual information (Tschannen et al., 2019), e.g., maximizing a tighter bound often leads to worse performance for downstream tasks. The underlying theoretical understanding of CL remains open.

We propose to understand CL with a collective point set matching view, which differs from the existing pairwise contrasting. As shown in Fig. 1, traditional methods (Chen et al., 2020; He et al., 2020) focus on improving the similarity for the positive pair and decreasing that for the negative pairs, which considers only one anchor at a time to compose the positive/negative pair in isolation. Different from the traditional view, we consider the set of mini-batch samples as a whole and learn the representations by matching between two point sets as shown in Fig. 1(b), where the point features in the two sets are learned with different encoders/augmentations from a set of samples, e.g. the same mini-batch training samples.

With this collective point set matching view, we propose to learn the representations with IOT (Li et al., 2019; Stuart & Wolfram, 2020), which supervises with empirical matching and aims to learn the cost matrix instead of the Coupling (i.e. the probability of matching matrix) in OT. In this paper, for solving the IOT problem, we view it with a min-min problem (as shown in Eq. 8) , where the outer minimization is to learn the representations by supervising the Coupling calculated in inner minimization.

Moreover, following previous works (Wang & Isola, 2020; Wang & Liu, 2021) which emphasize the uniformity for CL, a new penalty term is proposed in this paper based on the Coupling, which increase the uniformity of matching probabilities among negative pairs. In a nutshell, this paper contributes in the following aspects:

1) We propose a novel set matching view for contrastive learning, which jointly involves a collection matching of points, rather than anchor-based pairwise comparison (i.e. positive/negative pairs) as done in previous CL works.



(a) pairwise contrasting (b) collective matching

Figure 1: The difference between traditional contrastive learning and our matching method. **(a)** In traditional pairwise contrasting protocol, given one anchor $\mathbf{z}_1$, mainstream methods learn to attract the positive samples and repel its negative ones; **(b)** In our collective matching protocol, we consider the mini-batch features as a whole and learn the representations by improving the matching between two feature sets from different encoders/augmentations with the same mini-batch data.

2) Based on the above perspective, we propose IOT-CL, which contains min-min optimization as shown in Eq. 8. It can be proved that the object of minimization is a family of new contrastive loss functions when varying the degree of constraint relaxation in the Coupling set: **i)** We find the equivalence between our loss and InfoNCE with a specified Coupling set, which represents a new interpretation of InfoNCE in addition to the lower bound of mutual information. And with this specified Coupling set, InfoNCE and softmax cross-entropy loss can achieve theoretical unity under this matching view. **ii)** Other two kinds of contrastive losses are proposed by loosening and tightening the degree of constraint relaxation compared with constraints of infoNCE loss. The former loss has a closed-form result and the latter one should do the iteration of the Coupling to get the final loss.

3) We give a new understanding of uniformity for CL, that is, the matching probabilities of negative pairs remain low and even. With this idea, we propose the uniformity penalty on the Coupling. Experiments show the effectiveness of the penalty term. Source code will be made public available.

2 Background and Related Works

2.1 Optimal Transport and entropic Regularizaion

Originally introduced by Kantorovitch (Kantorovich, 1942), the discrete Optimal Transport is to solve a linear program, which is widely used for many classical problems such as matching (Wang et al., 2013). Specifically, given the cost matrix $\mathbf{C}$, Kantorovich's OT can read by solving the Coupling $\mathbf{P}$ (i.e. the joint probability matrix):

$$\min_{\mathbf{P}\in U(\mathbf{a},\mathbf{b})} < \mathbf{C}, \mathbf{P} >= \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{C}_{ij}\mathbf{P}_{ij} \tag{1}$$

where $\mathbf{a} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^m$ are histograms (i.e. probability vectors), and $U(\mathbf{a}, \mathbf{b})$ is the set of the Couplings:

$$U(\mathbf{a}, \mathbf{b}) = \{\mathbf{P} \in \mathbb{R}_+^{n\times m} | \mathbf{P}1_m = \mathbf{a}, \mathbf{P}^\top 1_n = \mathbf{b}\} \tag{2}$$

which is bounded and defined by $n + m$ equality constraints. When $n = m$ and $\mathbf{a} = \mathbf{b} = \mathbf{1}/n$ for every i, j, the OT is equivalent to solve a balanced matching problem, while unbalanced matching problem can also be formulated with OT by setting $n \neq m$ and $\mathbf{a} = \mathbf{1}/n, \mathbf{b} = \mathbf{1}/m$.

A lot of methods (Bertsimas & Tsitsiklis, 1997; Benamou & Brenier, 2000) are proposed to solve the Kantorovitch OT problem and relaxing with the entropic regularization (Wilson, 1969) is one of the simple but efficient methods, whose objective reads:

$$\min_{\mathbf{P}\in U(\mathbf{a},\mathbf{b})} < \mathbf{C}, \mathbf{P} > -\epsilon H(\mathbf{P}), \tag{3}$$

where $\epsilon > 0$ is the coefficient for entropic regularization $H(\mathbf{P})$. The regularization $H(\mathbf{P})$ can be specified as

$$H(\mathbf{P}) = -\sum_{i,j} \mathbf{P}_{ij}(\log(\mathbf{P}_{ij}) - 1). \tag{4}$$

Obviously, the objective in Eq. 3 is an ϵ-strongly convex function, and thus it can be solved quickly with iterative methods e.g. the Sinkorn method (Sinkhorn, 1967). If we use this entropic regularized OT to solve the matching problem, the hard matching problem may convert to soft matching, whose result is a non-sparse probability matching matrix.

2.2 Inverse Optimal Transport

(Discrete) Optimal Transport can always learn the matching with a known cost matrix, however, remains that the underlying cost criterion is unknown. Different from traditional OT, Inverse Optimal Transport (IOT) (Dupuy et al., 2016; Li et al., 2019; Stuart & Wolfram, 2020) is to infer the underlying cost matrix that gives rise to an observation on the Coupling. Recently, some works has focused on this problem. For example, with noisy observations of OT plans, (Stuart & Wolfram, 2020) propose a systematic approach to infer unknown costs and , (Chiu et al., 2022) develops the mathematical theory of IOT. (Li et al., 2019) emphasize that IOT can not only predict potential future matching, but is also able to explain what leads to empirical matching and quantifies the impact of changes in matching factors. Different from previous IOT which aims to learn the inferred cost, we find that IOT can further learn the representations, which gives us a new understanding of CL.

2.3 Contrastive Learning

Recently, self-supervised methods based on contrastive learning have drawn increasing attention due to the excellent performances (Logeswaran & Lee, 2018), which learns the representations without data labeling. (Wu et al., 2018) first proposes an instance discrimination method and adopts a contrastive loss (called NCE loss) to improve the discrimination for positive/negative pairs. CPC (Oord et al., 2018) learns context-invariant representations and proposes the InfoNCE loss to maximize the mutual information between different levels of features. Then in this subsection, we focus on revisiting the following typical contrastive losses to show the differences and generalization for our matching based methods.

InfoNCE loss. InfoNCE is one of the most widely use loss for contrastive learning introduced in (Oord et al., 2018). Given two unlabeled data sets $\{x_i\}_{i=1}^m$ and $\{y_j\}_{j=1}^n$ where (x_i, y_i) is semantically related, the InfoNCE loss is specified as

$$\mathcal{L}_{\text{InfoNCE}} = -\sum_{i=1}^{n} \log \left(\frac{\exp(s_{ii}/\tau)}{\sum_{k \neq i} \exp(s_{ik}/\tau) + \exp(s_{ii}/\tau)} \right) \tag{5}$$

Here s_{ij} is a similarity (e.g. cosine) between the feature z_i and z'_j, where $z_i = f(x_i)$ and $z'_j = g(y_j)$ with two feature extractor $f(\cdot)$ and $g(\cdot)$ mapping the (augmented) raw samples from raw space (e.g. image pixel) to the latent space. Previous works (Oord et al., 2018; Zbontar et al., 2021; Tian et al., 2020) mainly understand the InfoNCE from the perspective of maximizing the lower bound of mutual information between different levels of features. However, some works disagree with the lower bound interpretation, which has issues in practice, e.g., maximizing a tighter bound often leads to worse performance for downstream tasks (Tschannen et al., 2019). Different from the mutual information perspective, we will give a new interpretation with matching view in this paper.

Alignment and Uniformity. (Wang & Isola, 2020) views the contrastive learning with alignment and uniformity of feature distributions on the output unit hypersphere, which reads

$$\mathcal{L}_{\text{align}} = \sum_{i} ||z_i - z'_i||_2^2 \text{ and } \mathcal{L}_{\text{uniform}} = \log \sum_{i,j} e^{2||z_i - z'_j||_2^2} \tag{6}$$

where all features $\{z_i\}$ and $\{z'_i\}$ are L2 normalized (i.e. $||z_i||_2 = ||z'_i||_2 = 1$). Note $\mathcal{L}_{\text{uniform}}$ is designed with Gaussian potential kernel, which tries to learn the uniformity among negative pairs. (Wang & Isola, 2020) tries to learn with $\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{uniform}}$ for effectively restricting the output space to the unit hypersphere. In this paper, we give another view for alignment and uniformity, which improve the two properties with matching probabilities instead of L2 norm.

Hard Negative Sampling (HNSampling). Based on InfoNCE, (Wang & Liu, 2021) gives a more straightforward hard negative sampling strategy which truncates the gradients with respect to the

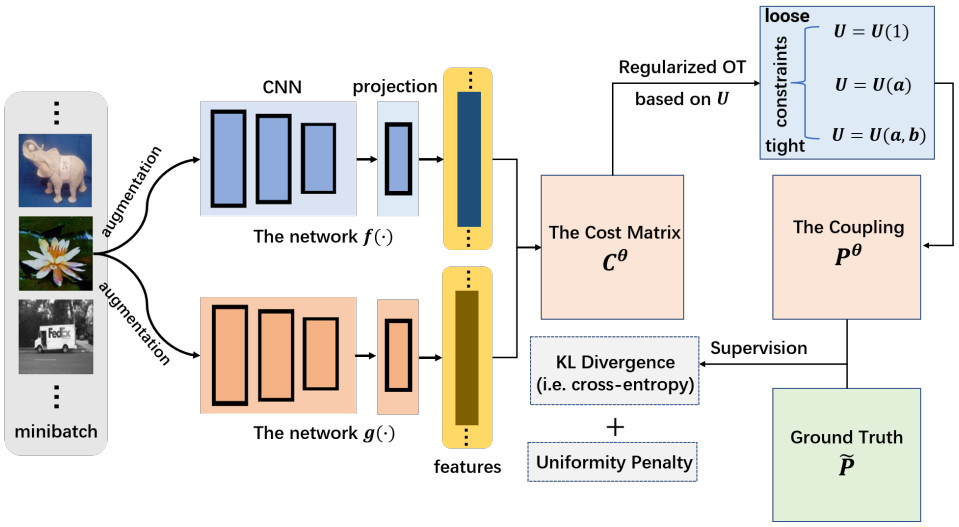


Figure 2: The overview of our approach for CL. The regularized OT is used to analyze and estimate the (matching) coupling, which is supervised with ground truth for learning the representations.

uninformative negative samples. The contrastive loss with hard negative sampling is specified as

$$\mathcal{L}_{\text{hard}} = -\sum_{i=1}^{n} \log \left(\frac{\exp(s_{ii}/\tau)}{\sum_{k:s_{ik}>s_\alpha^i} \exp(s_{ik}/\tau) + \exp(s_{ii}/\tau)} \right) \tag{7}$$

where s_α^i is the upper α quantile of the similarities $s_{i,:}$, which samples the negative pairs with high similarity ones. It is believed that with the selection of hard negative samples, the learned features will behave more uniformity. We will compare it with our method for uniformity learning.

3 Set Matching Framework for Contrastive Learning

3.1 Formulating Contrastive Learning as Set Matching

In this section, we propose a collective set matching approach for CL (IOT-CL), which studies contrastive learning by matching two feature point sets $\{z_i\}_{i=1}^n$ and $\{z'_j\}_{j=1}^m$ where $z_i = f(x_i)$ and $z'_j = g(y_j)$. With the cost matrix $\mathbf{C}^\theta \in \mathbb{R}_+^{n\times m}$ designed with features $\{z_i\}_{i=1}^n$ and $\{z'_j\}_{j=1}^m$, the optimization of IOT-CL is defined with two minimization, which reads (the same with the formula in Sec. 1):

$$\min_\theta KL(\tilde{\mathbf{P}}|\mathbf{P}^\theta) \quad \text{where} \quad \mathbf{P}^\theta = \arg\min_{\mathbf{P}\in U} < \mathbf{C}^\theta, \mathbf{P} > -\epsilon H(\mathbf{P}) \tag{8}$$

where $H(P)$ is the entropic regularization as defined in Eq. 4 and U is the set of the couplings based on the constraints. In the outer minimization, the coupling $\mathbf{P}^\theta$ is the matching probability matrix and $\tilde{P}$ is the ground truth. The aim of outer minimization is to supervise the soft matching with the ground truth to learn the representation. In inner minimization, the soft matching problem is formulated with the entropic regularized Optimal Transport. Our goal is to solve the coupling $\mathbf{P}^\theta$ with the cost matrix $\mathbf{C}^\theta$. In addition to setting $U = U(\mathbf{a}, \mathbf{b})$ where $\mathbf{a} = \mathbf{b} = \mathbf{1}/n$, we can loosen the constraint relaxation in U as

$$U(\mathbf{a}) = \{\mathbf{P} \in \mathbb{R}_+^{n\times m} | \mathbf{P}1_m = \mathbf{a}\}, \tag{9}$$

which only contains half of contraints in $U(\mathbf{a}, \mathbf{b})$ and we can also further loosen the relaxation as

$$U(1) = \{\mathbf{P} \in \mathbb{R}_+^{n\times m} | \sum_{i,j} \mathbf{P}_{ij} = 1\} \tag{10}$$

which only asks the basic probability requirements for the coupling. Thus, by varying the degree of constraint relaxation, we can get different contrastive losses of IOT-CL. In the following subsections, we will analyze the contrastive loss by setting $U = U(1), U(\mathbf{a})$ and $U = U(\mathbf{a}, \mathbf{b})$ in detail and show the generality of our contrastive loss.

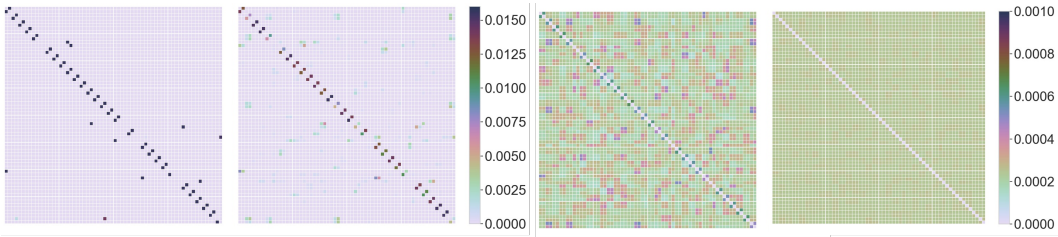


Figure 3: Results of couplings $\mathbf{P}^\theta$ by varying ϵ with given 64 trained features on CIFAR-10 based on the SimCLR framework(Chen et al., 2020). When $\epsilon \to 0$, $\mathbf{P}^\theta$ becomes more 'sharp' for the probability prediction. With the increment of ϵ, $\mathbf{P}^\theta$ becomes more uniform and when $\epsilon \to +\infty$, $\mathbf{P}^\theta$ is approximating to a uniform distribution, which has nothing to do with the quality of the learned features.

3.2 INFONCE IS A SPECIAL CASE UNDER $U(\mathbf{a})$

We first perform the analysis of contrastive loss when $U = U(\mathbf{a})$, which can be proven equivalent to the infoNCE loss. We begin to rewrite the inner minimization for solving the coupling $\mathbf{P}^\theta$:

$$\mathbf{P}^\theta = \arg\min_{\mathbf{P} \in U(\mathbf{a})} < \mathbf{C}^\theta, \mathbf{P} > -\epsilon H(\mathbf{P}), \tag{11}$$

which can be easily solved as an analytical form with the Lagrangian method:

$$\mathbf{P}^\theta_{ij} = \frac{\exp(-\mathbf{C}^\theta_{ij}/\epsilon)}{n \sum_{k=1}^m \exp\left(-\mathbf{C}^\theta_{ij}/\epsilon\right)} \tag{12}$$

The proof detail is given in Appendix A.1. So the solution of coupling is in the Softmax form under $U(\mathbf{a})$ for inner minimization. Then in the outer minimization, if we set $\tilde{P}_{ii} = \frac{1}{n}$ for each i and $\tilde{P}_{ij} = 0$ when $i \neq j$, we get the contrastive loss under $U(\mathbf{a})$:

$$\mathcal{L}_{\text{IOT-CL}} = -\frac{1}{n} \sum_{i=1}^n \log \left(\frac{\exp(-\mathbf{C}^\theta_{ii}/\epsilon)}{\sum_{j=1}^m \exp(-\mathbf{C}^\theta_{ij}/\epsilon))} \right) + \text{Constant} \tag{13}$$

Then we can easily find the equivalence between Eq. 13 and InfoNCE loss in Eq. 5 if we set $\mathbf{C}^\theta_{ij} = 1 - s_{ij}$, i.e. $\mathbf{C}^\theta_{ij}$ is the cosine distance between the features z_i and z'_j. It shows we can understand the InfoNCE with our soft matching view and theoretical results in entropic regularized OT may can help us analyze the properties for CL.

Regularization Coefficient ϵ. With the equivalence between InfoNCE and our loss in Eq. 13, we can also find that the temperature τ in InfoNCE exactly equals to the regularization coefficient ϵ (i.e. $\tau = \epsilon$). This finding is new and interesting to our best knowledge. Specially, when $\tau \to 0$, (Wang & Liu, 2021) proves that the InfoNCE will be converted to triplet loss:

$$\mathcal{L}_{\text{triplet}} = \lim_{\tau \to 0} \mathcal{L}_{\text{InfoNCE}} = \lim_{\tau \to 0} \frac{1}{\tau} \sum_i \max[s^i_{max} - s_{ii}, 0] \tag{14}$$

where s^i_{max} is the maximum of $\{s_{i,:}\}$ with the anchor feature z_i. In our matching understandings, $\epsilon \to 0$ means the hard matching without entropic regularization. In this view, it satisfies the matching requirement by making s_{ii} be the largest in the set $\{s_{i,:}\}$. On the other hand, when $\epsilon \to \infty$, the coupling will become more uniform as shown in Fig. 3. However, the uniformity for negative pairs is what we need to learn instead of conversion results with a very large ϵ. Thus too large ϵ is not conducive to the uniformity learning.

Balanced Matching in the Same Space. In addition to memory bank based methods, it is also popular to select the negative samples within the same mini-batch samples and augmentations e.g. InvaSpread (Ye et al., 2019) and SimCLR (Chen et al., 2020). Specifically, they randomly sample a mini-batch of N examples and learn the representations on pairs of augmented examples derived from the same mini-batch, resulting in $2N$ data points. They do not sample negative examples explicitly from memory bank. Instead, given a positive pair, they treat the other $2(N-1)$augmented examples within a mini-batch as negative examples.

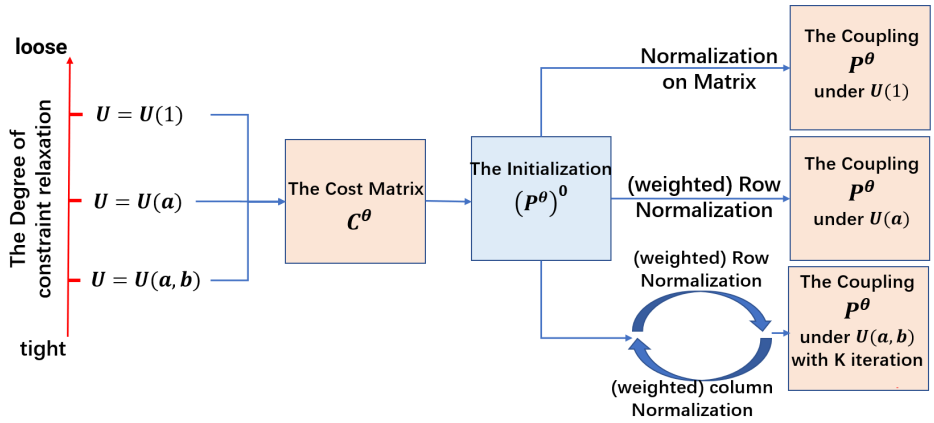


Figure 4: Comparison of the coupling by varying the degree of constraint relaxation.

In the view of matching, the two collective point sets are the same within $2N$ data points and in this case, the matching is balanced one defined in the same space as shown in Fig. 6(b). Specifically, with features $\{z_i\}_{i=1}^N$ and $\{z'_j\}_{j=1}^N$, we can reset the features as $\tilde{z}_{2k-1} = \mathbf{z}_k$ and $\tilde{z}_{2k} = z'_k$ when $k = 1, 2, \ldots, N$. The new cosine similarity is specified as $\tilde{s}_{ij} = \tilde{z}_i \cdot \tilde{z}_j / (||\tilde{z}_i|| \cdot ||\tilde{z}_j||)$. In this SimCLR case, the cost matrix $\mathbf{C}^\theta$ and ground truth $\tilde{\mathbf{P}}$ read

$$\mathbf{C}^\theta_{ij} = \begin{cases} +\infty, & i = j \\ 1 - \tilde{s}_{ij}, & else. \end{cases} \quad \text{and} \quad \tilde{\mathbf{P}}_{ij} = \begin{cases} \frac{1}{n}, & (i,j) \in S, \\ 0, & else. \end{cases} \tag{15}$$

where S is self-supervised set for positive pairs with $S = S_1 \cup S_2$. Here S_1 and S_2 are specified as

$$\begin{aligned} S_1 &= \{(i,j) | i = 2k, j = 2k-1, k = 1, \ldots, N\} \\ S_2 &= \{(i,j) | i = 2k-1, j = 2k, k = 1, \ldots, N\} \end{aligned} \tag{16}$$

When $i = j$, we set $\mathbf{C}^\theta_{ij} \to +\infty$, which means matching itself is not available for every sample. Then we can get that $\exp(-\mathbf{C}^\theta_{ii}/\epsilon) \to 0$. For the contrastive loss:

$$\mathcal{L}_{\text{IOT-CL}} = -\frac{1}{2N} \sum_{(i,j) \in S} \log \left(\frac{\exp(-\mathbf{C}^\theta_{ii}/\epsilon)}{\sum_{s=1}^{2N} 1_{i \neq s} \exp(-\mathbf{C}^\theta_{is}/\epsilon))} \right) \tag{17}$$

which is exactly the contrastive loss in SimCLR (Chen et al., 2020). Thus SimCLR can be interpreted with balanced matching view as shown in Fig. 6(b). For unbalanced matching case, we discuss it in Appendix C with the MoCo framework.

3.3 New Contrastive Losses under $U(1)$ and $U(\mathbf{a}, \mathbf{b})$

As shown above, our IOT-CL loss can be viewed as InfoNCE under $U = U(\mathbf{a})$. In this subsection, we give the results of contrastive loss when $U = U(1)$ and $U(\mathbf{a}, \mathbf{b})$. Fig. 4 shows the difference of calculating operations for the coupling.

Contrastive Loss under $U(1)$. To propose the new contrastive loss, we first loosen the constraint relaxation by setting $U = U(1)$. Then we can get the coupling matrix $\mathbf{P}^\theta$ as

$$\mathbf{P}^\theta_{ij} = \frac{\exp(-\mathbf{C}^\theta_{ij}/\epsilon)}{\sum_{t=1}^n \sum_{s=1}^m \exp(-\mathbf{C}^\theta_{ts}/\epsilon))} \tag{18}$$

Different from the coupling in Eq. 12 under $U(\mathbf{a})$, this new coupling matrix is symmetric if $\mathbf{C}^\theta$ is a symmetric matrix. Then we can get the loss under $U(1)$ as

$$\mathcal{L}_{\text{IOT-CL}} = -\frac{1}{n} \sum_{i=1}^n \log \left(\frac{\exp(-\mathbf{C}^\theta_{ii}/\epsilon)}{\sum_{t=1}^n \sum_{s=1}^m \exp(-\mathbf{C}^\theta_{ts}/\epsilon))} \right) \tag{19}$$

The proof of Eq. 18 and Eq. 19 are in Appendix A.2. The main difference between above loss and InfoNCE is that $\mathbf{P}^{\theta}_{ij}$ is only determined by the $i-$th row of cost matrix $\mathbf{C}^{\theta}$, while the above coupling cares all the values in cost matrix.

Contrastive Loss under $U(\mathbf{a}, \mathbf{b})$. Next we propose another new loss for IOT-CL by tightening the constraint relaxation (i.e. fulfilling full constraints of matching in U). Similarly in Sec. C, we first solve the inner minimization in Eq. 8. As discussed in (Cuturi, 2013), the closed-form coupling may not exist, which differs from the Couplings under $U(1)$ and $U(\mathbf{a})$. By setting

$$\left(\mathbf{P}^{\theta}\right)^{0} = \exp\left(-\mathbf{C}^{\theta}/\epsilon\right) \tag{20}$$

we adopt the popular Sinkhorn algorithm (Adams & Zemel, 2011; Cuturi, 2013; Wang et al., 2019) to approximate the optima:

$$\left(\mathbf{P}^{\theta}\right)^{k}_{\text{temp}} = \frac{1}{n}\left(\mathbf{P}^{\theta}\right)^{k-1} \oslash \left(\left(\mathbf{P}^{\theta}\right)^{k-1}\mathbf{1}_{m\times m}\right) \tag{21}$$

$$\left(\mathbf{P}^{\theta}\right)^{k} = \frac{1}{m}\left(\mathbf{P}^{\theta}\right)^{k}_{\text{temp}} \oslash \left(\left(\mathbf{1}_{n\times n}\mathbf{P}^{\theta}\right)^{k}_{\text{temp}}\right) \tag{22}$$

where $\oslash$ means element-wise division, and $\mathbf{1}_{m\times m}$ and $\mathbf{1}_{n\times n}$ are the matrices whose elements are all ones. Exactly the Sinkhorn algorithm works iteratively by taking $1/n$ weighted row normalization of Eq. 21 and $1/m$ weighted column normalization of Eq. 22 alternatively.

By iterating Eq. 21 and Eq. 22 for $k = 1, 2\ldots, K$, we can get the coupling results. Exactly when $K = 1$, the intermediate matrix $\left(\mathbf{P}^{\theta}\right)^{1}_{\text{temp}}$ equals the coupling under $U(\mathbf{a})$, which has the loosen relaxation for constraints. And when $K \to \infty$, $\left(\mathbf{P}^{\theta}\right)^{K}$ will converge to optimal solution under $U(\mathbf{a}, \mathbf{b})$. Thus increasing the value of K is exactly tightening the constraint relaxation in U. Besides, this Sinkhorn operation is fully differentiable because only element-wise division and matrix multiplication are used in iterations. Thus it can be efficiently implemented by PyTorch's automatic differentiation functions. Finally, we can get the contrastive loss under $U(\mathbf{a}, \mathbf{b})$

$$\mathcal{L}_{\text{IOT-CL}} = \sum_{i=1}^{m}\sum_{j=1}^{n}\tilde{\mathbf{P}}_{ij}\log\left(\mathbf{P}^{\theta}_{ij}\right)^{K} \tag{23}$$

with a proper iterative number K and the ground truth $\tilde{\mathbf{P}}$.

3.4 Enhancing Uniformity for IOT-CL

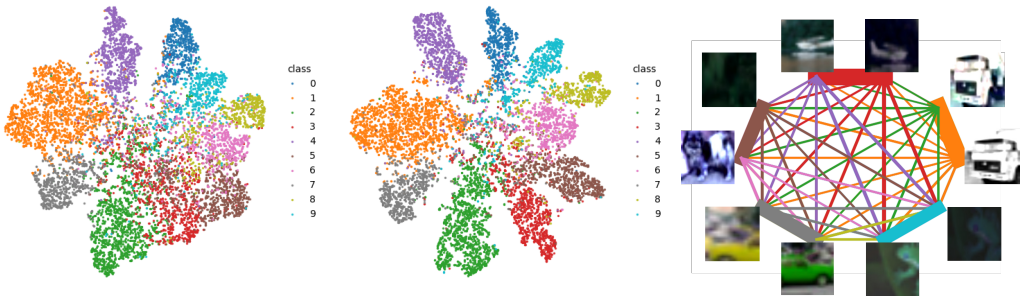
Following the works (Wang & Isola, 2020; Wang & Liu, 2021) emphasizing alignment and uniformity on the hypersphere for CL, we can rethink these two key properties from the matching perspective. The alignment requires similar samples to have similar features (Wang & Isola, 2020), which must have a high probability for matching. While uniformity prefers a uniform distribution for features on the unit hypersphere (Wang & Isola, 2020), which can be understood as uniformly matching among negative pairs. Thus with the coupling $\mathbf{P}^{\theta}$, the alignment and uniformity loss with matching view can be specified as

$$\min_{\theta} \mathcal{L}_{\text{IOT-CL}} + \lambda_p KL(\bar{\mathbf{Q}}^{\theta}|\mathbf{P}^{\theta}) \tag{24}$$

where λ_p is the uniformity penalty coefficient and $\bar{\mathbf{Q}}$ reads:

$$\bar{\mathbf{Q}}^{\theta}_{ij} = \begin{cases} \mathbf{P}^{\theta}_{ij}, & i = j, \\ \underset{i\neq j}{\text{mean}}\,\mathbf{P}^{\theta}_{ij}, & i \neq j, . \end{cases} \tag{25}$$

The first term in Eq. 24 represents the matching alignment, which increases the probability of positive pair matching, while the first term is for increasing the uniformity, where $\bar{\mathbf{Q}}_{\mathbf{ij}}$ is the mean of matching probability for negative pairs when $i \neq j$. We let $\mathbf{P}^{\theta}$ approximate to $\mathbf{Q}^{\theta}$ by using KL divergence, which decreases the volatility as well as uniformity.



(a) Without pnalty (b) With penalty (c) Matching results with penalty

Figure 5: (a) and (b) are T-SNE visualizations of the embedding distribution without and with uniformity penalty on the coupling for studying the uniformity on CIFAR-10. (c) is the matching result based on the uniformity penalty on coupling $\mathbf{P}^{\theta}$. In detail, Given 5 original images of CIFAR-10 and their augmented images, we can get matching visualization under SimCLR framework. The thickness of the lines is proportional to the element value in $\mathbf{P}^{\theta}$. Color has no particular meaning but for visual effects

Table 1: ACC results (%) of IOT-CL (without uniformity penalty) evaluated by linear networks when varying the relaxation of constraints in U with 100/200 epoch training. Here U is set to $U(1)$, $U(\mathbf{a})$, and $U(\mathbf{a}, \mathbf{b})$ (using Sinkhorn Algorithm with $K = 1, 2, 4, 8$ instead) to get different degrees of constraint relaxation for contrastive loss.

Method	CIFAR-10		CIFAR-100		SVHN	
	100 epochs	200 epochs	100 epochs	200 epochs	100 epochs	200 epochs
$U(1)$	81.75	84.43	51.60	55.63	85.11	87.12
$U(\mathbf{a})$	81.31	84.62	**52.85**	56.04	84.53	87.08
$K = 1$	**81.79**	84.55	51.88	**56.14**	**85.22**	**87.14**
$K = 2$	81.77	**84.80**	51.57	55.98	84.76	86.95
$K = 4$	81.68	84.50	51.32	55.72	85.03	86.66
$K = 8$	81.54	84.48	51.50	55.59	84.95	86.87

4 EXPERIMENTS

4.1 PROTOCOL

Pretraining. We conduct experiments on CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2011) and ImageNet-100 (Deng et al., 2009). The label selection of ImageNet-100 is in line with (Wang & Liu, 2021). In the pretraining stage, resnet18 (He et al., 2016) is adopted as the backbone on CIFAR-10, CIFAR-100, and SVHN. The resnet50 (He et al., 2016) is used as the backbone on ImageNet-100. Besides, the augmentations follow (Chen et al., 2020) with random color distortions, random Gaussian blur, and random cropping followed by resizing back to the original size. For model architecture, we mainly follow the framework of SimCLR (Chen et al., 2020), which uses an encoder network and a projector head to maximize agreement for different contrastive losses. We train all models with Adam (Kingma & Ba, 2014) for 200 epochs by 3e-4 learning rate with a mini-batch size of 256. Besides, the temperature τ is set to 0.5 for softmax based methods.

Evaluation. With all convolutional layers frozen, we first validate the performance of the pretrained models on linear classification. Specifically, we train the linear layer for 100 epochs with 256 mini-batch sizes. Besides, in our experiments, we also adopt the accuracy of a k-nearest neighbors classifier (k-NN, k = 5 here) as the evaluation. The advantage of this classifier is that it does not require additional parameters, which is applicable without training.

Baselines. In addition to the contrastive losses discussed in Sec. 2.3 (i.e. InfoNCE (Oord et al., 2018), HNSamping (Wang & Liu, 2021), and $\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{uniform}}$ (Wang & Isola, 2020)), we compare our losses of IOT-CL with triplet loss (Schroff et al., 2015) and the loss proposed in InvaSpread (Ye et al., 2019). The triplet loss exactly can be understand a special case of InfoNCE when the temperature $\tau \to 0$, while the loss in InvaSpread is exactly similair with the perturbation loss of graph matching (Wang et al., 2019), which learns the probability of positive/negative pairs with Bernoulli distribution.

Table 2: Accuracy (%) of uniformity penalty evaluated by Linear network (Lin.) and k-NN method for CIFAR-10, CIFAR-100 and SVHN based on 100 epochs training.

Method	CIFAR-10		CIFAR-100		SVHN	
	Lin.	k-NN	Lin.	k-NN	Lin.	k-NN
Triplet (Schroff et al., 2015)	70.97	64.83	40.58	30.75	71.62	53.57
InvaSpread (Ye et al., 2019)	81.51	77.21	51.15	39.74	85.34	73.43
InfoNCE (Oord et al., 2018)	81.31	77.01	52.85	40.26	84.53	73.09
$\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{uniform}}$ (Wang & Isola, 2020)	82.84	79.03	55.02	45.13	89.71	83.76
HNSampling (Wang & Liu, 2021)	82.50	79.28	54.20	42.53	87.49	77.94
IOT-CL(K=1 with penalty)	**84.18**	**80.19**	**56.88**	**46.60**	**90.98**	**84.16**

Table 3: Ablation study (evaluated by linear network) on CIFAR-10 with different uniformity coefficients. We test the contrastive loss under $U(\mathbf{a}, \mathbf{b})$ by setting $K = 1, 2$.

λ_p value	0.5	1	1.5	2	2.5	3	3.5
$K = 1$	82.06	82.41	**83.11**	82.97	83.01	82.95	82.98
$K = 2$	82.14	82.61	82.83	82.86	82.78	**83.33**	83.20

Table 4: ACC results (%) evaluated by Linear network (Lin.) and k-NN method for ImangeNet-100 based on 100 epochs training.

Method	Network	ACC of Lin.	ACC of k-NN
InfoNCE (Oord et al., 2018)	ResNet50	64.49	49.47
$\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{uniform}}$ (Wang & Isola, 2020)	ResNet50	68.02	56.12
HNSampling (Wang & Liu, 2021)	ResNet50	64.06	47.94
IOT-CL	ResNet50	68.18	54.81

4.2 Main Results

Impact of Constraint Relaxation. Table 1 shows the ACC results evaluated by liner network on CIFAR-10. By varying the degree of constant relaxation, we can find that the loosest (i.e. $U = U(1)$) or the tightest (i.e. $K = 8$ in this experiment) may not be the best of both. With the limited training epoch, finding proper relaxation can be an important thing. From Table 1, we find that loosening relaxation may be more suitable for smaller epoch numbers, while tight relaxation may be better for large epoch numbers.

Results with Uniformity Penalty on Coupling. Fig. 5 shows the embedding distribution on CIFAR-10 without penalty (i.e. the loss of IOT-CL with $K = 1$), with Gaussian Potential Kernal penalty ($\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{uniform}}$ (Wang & Isola, 2020)) and with our penalty on the Coupling (i.e. Eq. 24). The embedding is based on the logits of linear classification network. We can find a similar performance between ours and the work in (Wang & Isola, 2020), and outperform the results without penalty. Table 2 shows the results for CIFAR-10, CIFAR-100, and SVHN based on the SimCLR framework. At first, we can find the uniformity penalty can improve the performance both in Linear classification (Lin.) and K-NN. Specifically, compared with InfoNCE and its variants, the accuracy performances increase with a large value in all of the datasets for our method (i.e. IOT-CL($K = 1$ under $U(\mathbf{a}, \mathbf{b})$) with uniformity penalty on coupling). Thus, compared with state of art methods, we can find our method outperforms in most cases.

Besides, as shown in Fig. 5, we can find that by adding uniformity penalty based on the loss of IOT-CL, the representation of features will be clearer and features between different classes will be more separated. Table 3 shows the sensitivity for uniformity penalty by varying λ_p with running 100 epochs. We can find that the performance of the downstream task is insensitive to λ with different K. Besides, to the litmit of the paper length, other experiments are presented in Appendix. Besides, we test the results based on MoCo to show the generalization of our methods in Appendix C.

Experiments on Imagenet-100 We further test our model on ImangeNet-100 and using Resnet50 as backbone. The pretraining protocol is given in Sec. 4.1. As shown in Table 4, we can find our model with with uniformity penalty can greatly improve the performance in this larger experiments.

5 Conclusion

In this paper, we have presented a set matching-based framework to interpret the contrastive loss widely used in (self-supervised) representation learning. Under this framework, we develop a family of new loss functions and apply optimal transport techniques to contrastive learning. In particular, the existing popular loss e.g. InfoNCE can be viewed as a special case. New space for improvement has been shown in our designed new algorithms, and experimental results on public datasets verify the effectiveness of our models.

REFERENCES

Ryan Prescott Adams and Richard S Zemel. Ranking via sinkhorn propagation. *arXiv preprint arXiv:1106.1925*, 2011.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

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

Wei-Ting Chiu, Pei Wang, and Patrick Shafto. Discrete probabilistic inverse optimal transport. In *International Conference on Machine Learning*, pp. 3925–3946. PMLR, 2022.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances. *arXiv preprint arXiv:1306.0895*, 2013.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Arnaud Dupuy, Alfred Galichon, and Yifei Sun. Estimating matching affinity matrix under low-rank constraints. *arXiv preprint arXiv:1612.09585*, 2016.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1074–1083, 2021.

L Kantorovich. On the transfer of masses (in russian). In *Doklady Akademii Nauk*, volume 37, pp. 227–229, 1942.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Ruilin Li, Xiaojing Ye, Haomin Zhou, and Hongyuan Zha. Learning to match via inverse optimal transport. *Journal of machine learning research*, 20, 2019.

Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*, 2018.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

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

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

Andrew M Stuart and Marie-Therese Wolfram. Inverse optimal transport. *SIAM Journal on Applied Mathematics*, 80(1):599–619, 2020.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020.

Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2495–2504, 2021.

Runzhong Wang, Junchi Yan, and Xiaokang Yang. Learning combinatorial embedding networks for deep graph matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3056–3065, 2019.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Wei Wang, Dejan Slepčev, Saurav Basu, John A Ozolek, and Gustavo K Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101(2):254–269, 2013.

Alan Geoffrey Wilson. The use of entropy maximising models, in the theory of trip distribution, mode split and route split. *Journal of transport economics and policy*, pp. 108–126, 1969.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6210–6219, 2019.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

A LAGRANGIAN FOR REGULARIZED OT

A.1 LAGRANGIAN UNDER $U(a)$

Now we show the collective matching framework for contrastive Learning with the simplified constraints:

$$U(\mathbf{a}) = \{\mathbf{P} \in \mathbb{R}_+^{n\times m} | \mathbf{P}1_m = \mathbf{a}\} \tag{26}$$

where $\mathbf{a} = \mathbf{1}/m$ and 1_m is the m-dimensional column vector whose elements are all ones. With the objective of the regularized OT:

$$\mathbf{P}^\theta = \arg\min_{P\in U(\mathbf{a})} < \mathbf{C}^\theta, \mathbf{P} > -\epsilon H(\mathbf{P}), \tag{27}$$

We introduce the dual variable $\mathbf{f} \in R^n$. The Lagrangian of the above equation is:

$$L(\mathbf{P}, \mathbf{f}) =< \mathbf{C}^\theta, \mathbf{P} > -\epsilon H(\mathbf{P}) - \sum_{i=1}^{n} \mathbf{f}_i \cdot \left(\sum_{j=1}^{m} \mathbf{P}_{ij} - \frac{1}{n} \right) \tag{28}$$

The first order conditions then yield by:

$$\frac{\partial L(\mathbf{P}, \mathbf{f})}{\partial \mathbf{P}_{ij}} = \mathbf{C}_{ij}^\theta + \epsilon \log \mathbf{P}_{ij} - \mathbf{f}_i = 0 \tag{29}$$

Thus we have $\mathbf{P_{ij}} = e^{(\mathbf{f}_i - C_{ij}^\theta)/\epsilon}$ for every i and j, for optimal $\mathbf{P}$ coupling to the regularized problem. Due to $\sum_j \mathbf{P}_{ij} = 1/n$ for every i, we can calculate the Lagrangian parameter $\mathbf{f}_i$ and the solution of the coupling is given by:

$$\mathbf{P}_{ij} = \frac{\exp\left(-\mathbf{C}_{ij}^\theta/\epsilon\right)}{n \sum_{t=1}^{m} \exp\left(-\mathbf{C}_{it}^\theta/\epsilon\right)} \tag{30}$$

Then in outer minimization, if we set $\tilde{P}_{ii} = \frac{1}{n}$ for each i and $\tilde{P}_{ij} = 0$ when $i \neq j$, we get the contrastive loss under $U(\mathbf{a})$

$$\mathcal{L}_{\text{IOT-CL}} = -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{\exp(-\mathbf{C}_{ii}^\theta/\epsilon)}{\sum_{j=1}^{m} \exp(-\mathbf{C}_{ij}^\theta/\epsilon))} \right) + \text{Constant} \tag{31}$$

We have therefore got the loss of IOT-CL under $U(\mathbf{a})$.

A.2 Lagrangian under $U(1)$

If the relaxation in U is further loosen by setting $U = U(1)$:

$$\tilde{U}(1) = \{\mathbf{P} \in \mathbf{R}_+^{n \times m} | \sum_{i,j} \mathbf{P} = 1\}$$

The objective in inner optimization can be specified as

$$\mathbf{P}^\theta = \arg \min_{\mathbf{P} \in U(1)} < \mathbf{C}^\theta, \mathbf{P} > -\epsilon H(\mathbf{P}), \tag{32}$$

Introducing dual variable $\lambda \in R$, the Lagrangian of the above equation reads:

$$L(\mathbf{P}, \lambda) =< \mathbf{C}^\theta, \mathbf{P} > -\epsilon H(\mathbf{P}) - \lambda(\sum_{i,j} P_{ij} - 1) \tag{33}$$

First order conditions then yield by:

$$\frac{\partial L(\mathbf{P}, \lambda)}{\partial \mathbf{P}_{ij}} = \mathbf{C}_{ij}^\theta + \epsilon \log \mathbf{P}_{ij} - \lambda = 0 \tag{34}$$

which result, for an optimal P coupling to the regularized problem, in the expression $\mathbf{P}_{ij} = e^{(\lambda - \mathbf{C}_{ij})/\epsilon}$. Due to $\sum_{ij} \mathbf{P}_{ij} = 1$, we can calculate the λ and the solution Coulping can be wriiten as

$$\mathbf{P}_{ij} = \frac{\exp\left(-\mathbf{C}_{ij}^\theta/\epsilon\right)}{\sum_{st} \exp\left(-\mathbf{C}_{st}^\theta/\epsilon\right)} \tag{35}$$

Then in outer minimization, if we set $\tilde{P}_{ii} = \frac{1}{n}$ for each i and $\tilde{P}_{ij} = 0$ when $i \neq j$, we get the contrastive loss under $U(1)$:

$$\mathcal{L}_{\text{IOT-CL}} = -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{\exp(-\mathbf{C}_{ii}^\theta/\epsilon)}{\sum_{t=1}^{n} \sum_{s=1}^{m} \exp(-\mathbf{C}_{ts}^\theta/\epsilon))} \right) \tag{36}$$

We have therefore got the loss of IOT-CL under $U(1)$.

B ALGORITHM

With SimCLR structure, the algorithm of Sinkhorn is shown in Algorithm 1.

Algorithm 1: Sinkhorn Algorithm for IOT-CL

feature similarity matrix Sim $i.e.$ $-\mathbf{C}^{\theta}$; the iterative number N for Sinkhorn Surrogate for similarity based on Sinkhorn

Sinkhorn *(Sim, N)*:

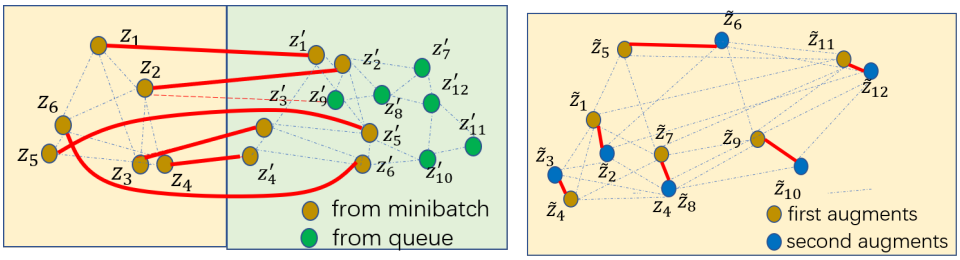
```
    #Let the diag of Sim tend to −∞, i.e. make self-matching case unavailable
    eye = torch.eye(Sim.shape[0])
    diags = -100 * eye * Sim; 100 is large enough
    off_diags = (1-eye) * Sim;
    Sim = diags + off_diags;
    Sinkhorn iterations
    texp = torch.exp(Sim); Initialization of P^θ
    for i = 1, ..., N do
        texp = torch.div(texp, torch.sum(texp, dim=0)).T; row normalization
        texp = torch.div(texp, torch.sum(texp, dim=0)).T; column normalization
    end
    return Similarity Surrogate: torch.log(texp)
```

C UNBALANCED MATCHING FOR CL.

In memory bank based methods such as MoCo (He et al., 2020), negative samples are selected from the stored sample features. In this case, the matching is usually unbalanced. Specifically, assume that $\{z_i\}_{i=1}^n$ and $\{z'_j\}_{j=1}^n$ are features extracted by encoder f and momentum encoder g from the same mini-batch samples, while $\{z'_j\}_{j=n+1}^m$ are features extracted by g from the memory bank. Then we can get two feature sets $\{z_i\}_{i=1}^n$ and $\{z'_j\}_{j=1}^m$ where m is usually much larger than n. We can find the unbalanced matching for the memory bank based methods as shown in Fig. 6(a). Since MoCo select the negative samples in the memory bank, which contains the features of previous mini-bath data instead of the same mini-batch samples, the cost matrix is designed as

$$\mathbf{C}_{ij}^{\theta} = \begin{cases} +\infty, & i \neq j \text{ and } 1 \leq j \leq n \\ 1 - s_{ij}, & \text{else} \end{cases} \tag{37}$$

where s_{ij} is a similarity (e.g. cosine) between feature z_i and z'_j. Here $\mathbf{C}_{ij}^{\theta} \to +\infty$ implies $\exp(-\mathbf{C}_{ij}^{\theta}/\epsilon) \to 0$, which means that $\mathbf{P}_{ij}^{\theta} \to 0$ and the features z_i and z'_j will not be matched.



(a) Unbalanced Matching (b) Balanced Matching in one Space

Figure 6: Interpreting MoCo and SimCLR by point set matching. **(a)** the left part is extracted by f, while the right is extracted by the moment encoder g. In MoCo, in addition to the representations from minibatch (brown points), representations from moment queue are also in the right part, which is an unbalanced matching problem in our matching view ($n < m$); **(b)** the space is extracted by f (note $f = g$ in SimCLR) with two augmentations. Viewing the SimCLR framework from our perspective, the matching is not simply between two different sets but a balanced matching in the same space, and the feature points try to match their neighborhood with minimal total cost.

Table 5: ACC results (%) evaluated by Linear network (Lin.) and k-NN method for CIFAR-10and CIFAR-100 based on 100 epochs training. We mainly follow the MoCo framework in this case.

Method	CIFAR-10		CIFAR-100	
	Lin.	k-NN	Lin.	k-NN
InfoNCE (Oord et al., 2018)	72.11	61.66	47.71	28.23
$\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{uniform}}$ (Wang & Isola, 2020)	74.77	65.92	49.48	30.87
HNSampling (Wang & Liu, 2021)	72.18	60.73	47.88	28.06
IOT-CL (ours)	73.28	61.97	48.21	29.64

The condition $i \neq j$ and $1 \leq j \leq n$ represents that (z_i, z'_j) are negative pair from the same mini-batch, which is not adopt as negative pair in memory based methods.

With the cost matrix $\mathbf{C}^\theta$, the ground truth can be simply set as $\tilde{P}_{ii} = \frac{1}{n}$ for $i = 1, \ldots, n$ and $\tilde{P}_{ij} = 0$ when $i \neq j$. Then we can find that our IOT-CL does not contradict the memory bank based frameworks e.g. MoCo, which can be interpreted from our perspective as shown in Fig. 6(a).

Experiments under MoCo-based framework In addition to the SimCLR framework, we also test the loss of our model in Memory based framework (i.e. MoCo), which can be understood as an unbalanced matching in this paper. For the pretraining stage, all models are trained with SGD for 100 epochs by 0.03 learning rate with batch size being 128, the momentum and weight decay of SGD are 0.9 and $1e-4$ respectively. And the temperature τ is set to 0.07 for softmax based methods. We set the size of memory bank to 4096. The feature dimension is 128 and the momentum of updating the key encoder is 0.999. For Linear evaluation, we train for 100 epoch still with SGD except that the learning rate is 30 and weight decay is set to 0. Batch size for linear evaluation is 256. As shown in Table 5, We can find that the model with our loss works efficiently in MoCo-based framework.

D Connecting to Softmax Cross-Entropy Loss with IOT.

We can also view the classification under our collective matching perspective based on entropic regularized OT. In this case, assume that the feature $z_i = f(x_i)$ can be the logit vector for sample x_i and y_i is the corresponding one-hot label for $m-$ classification with n samples in a mini-batch. Then by defining $\mathbf{C}^\theta_{ij} = c - z_i \cdot y_j$ where c is large enough, we can get the loss

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} \log \frac{e^{-z_{ij} y_{ij}}}{\sum_{k=1}^{m} e^{-z_{ik} y_{ik}}} \tag{38}$$

where $y_{ij} \in \{0, 1\}$ and z_{ij} are the j dimension values of one-hot label y_i and logit vector z_i. The above loss is exactly the softmax cross-entropy loss, which verifies that the classification problem can also be viewed as a (soft) matching problem. Besides, InfoNCE and softmax cross-entropy loss can achieve theoretical unity under this matching view with aid of regularized OT.