Understanding and Generalizing Contrastive Learning from the Inverse Optimal Transport Perspective

Liangliang Shi [1] Gu Zhang [1] Haoyu Zhen [1] Jintao Fan [1] Junchi Yan [1]

Abstract

Previous research on contrastive learning (CL) has primarily focused on pairwise views to learn representations by attracting positive samples and repelling negative ones. In this work, we aim to understand and generalize CL from a point set matching perspective, instead of the comparison between two points. Specifically, we formulate CL as a form of inverse optimal transport (IOT), which involves a bilevel optimization procedure for learning where the outter minimization aims to learn the representations and the inner is to learn the coupling (i.e. the probability of matching matrix) between the point sets. Specifically, by adjusting the relaxation degree of constraints in the inner minimization, we obtain three contrastive losses and show that the dominant contrastive loss in literature InfoNCE falls into one of these losses. This reveals a new and more general algorithmic framework for CL. Additionally, the soft matching scheme in IOT induces a uniformity penalty to enhance representation learning which is akin to the CL's uniformity. Results on vision benchmarks show the effectiveness of our derived loss family and the new uniformity term.

1. Introduction

Unsupervised/self-supervised learning of representation (Hu et al., 2021; Grill et al., 2020) has garnered increasing attention. With contrastive learning (CL) (Chen et al., 2020; Gao et al., 2021), representation is learned by selecting an anchor and then identifying its positive/negative samples. The contrastive loss based on feature similarity is then used to distinguish between positive and negative pairs. However, the comparison of positive and negative pairs is often empirical, and popular contrastive losses like InfoNCE (Oord et al., 2018) have some inconsistencies in interpreting the lower bound of mutual information (Tschannen et al., 2019). Maximizing a tighter bound can also lead to worse performance for downstream tasks, leaving the full understanding of CL still open.

In this paper, we aim to understand CL with a collective point set matching perspective, which differs from mainstream pairwise contrasting practices. As shown in Fig. 1, traditional methods (Chen et al., 2020; He et al., 2020) focus on improving the similarity for positive pairs and decreasing it for negative pairs, by considering only one anchor at a time to form positive/negative pairs in isolation. This approach can easily ignore the relationships and influences among different anchors.

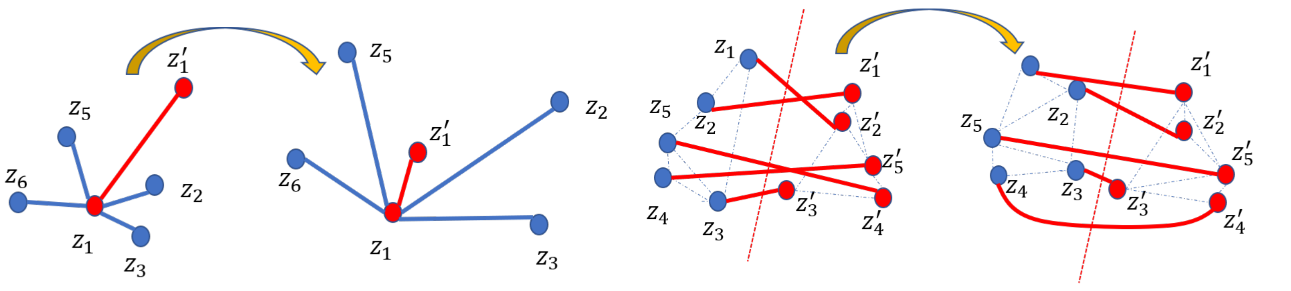
In contrast to this pairwise view, we consider the set of mini-batch samples as a whole and learn the representations by matching between two point sets, where each sample feature is regarded as a point as shown in Fig. 1(b). With this point set matching view, we propose to learn the representations with Inverse Optimal Transport (IOT) (Li et al., 2019; Stuart & Wolfram, 2020), which 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 bilevel optimization problem, where the inner minimization aims to learn the coupling matrix by varying the constraint relaxation and the outer minimization is to learn the representations by supervising the coupling calculated in inner minimization. Under this formulation, we can get some findings such as the equivalence between temperature coefficient in InfoCNE and the coefficient of entropic regularization in OT.

Furthermore, we propose a new penalty term based on the coupling matrix within our IOT framework, which encourages uniform matching probabilities among negative pairs. This is consistent with previous works (Wang & Isola, 2020; Wang & Liu, 2021) that have shown the importance of uniformity in contrastive learning. **The contribution of the paper can be summarized as follows:**

1) We propose a novel set matching view for contrastive learning, which jointly involves a collection matching of

[1] Department of Computer Science and Engineering, and MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University. Correspondence to: Junchi Yan <yanjunchi@sjtu.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).



(a) pairwise contrasting

(b) collective matching

Figure 1. Illustrative comparison between traditional contrastive learning and our point set matching framework. **(a)** In the traditional pairwise contrasting protocol, mainstream methods learn to attract the positive samples and repel the negative ones given one anchor $\mathbf{z}_1$. **(b)** In our point set 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.

points, rather than anchor-based pairwise comparison (i.e. positive/negative pairs) as done in previous CL works.

2) Based on the above perspective, we propose IOT-CL, which involves a bilevel optimization. It can be proved that the objective of minimization is a family of new contrastive loss functions by varying the degree of constraint relaxation in the coupling set: **i)** 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. **ii)** Other two kinds of contrastive losses are proposed by loosening and tightening the degree of constraint relaxation compared with the constraints of InfoNCE. The former loss enjoys a closed-form result and the latter one need iterative computing 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. The experimental results verify the effectiveness of our approach.

2. Background and Related Works

2.1. Optimal Transport and Entropic Regularization

As originally introduced by (Kantorovich, 1942), the discrete (in the sense of the matching target e.g. a point set) Kantorovich's 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 involves 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 $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$ are histograms (probability vectors), and $U(\mathbf{a},\mathbf{b})$ is the set of the couplings:

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

which is bounded and defined by $n+m$ equality constraints. When $n = m$ and $\mathbf{a} = \mathbf{b} = \mathbf{1}/n$ for each $i = 1, ..., n$, $j = 1, ..., m$, the OT is equivalent to solving a balanced matching problem, while unbalanced matching is 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 $H(\mathbf{P})$ and the $H(\mathbf{P})$, which can be specified as

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

The objective in Eq. 3 is an ϵ-strongly convex function, and thus the optimization has a unique solution, which can be solved 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.

2.2. Inverse Optimal Transport

Inverse Optimal Transport (IOT) has been studied (Dupuy et al., 2016; Li et al., 2019; Stuart & Wolfram, 2020) which aims to infer the unknown cost $\mathbf{C}$ that gives rise to an observation on the coupling. (Stuart & Wolfram, 2020) proposes a systematic approach to infer unknown costs and (Chiu et al., 2022) develops the mathematical theory behind IOT. (Li et al., 2019) shows that IOT can not only predict potential matching, but is also able to explain what leads to empirical matching and quantifies the impact of changes in matching factors. The IOT problem can be formulated as:

$$\begin{aligned} &\min_{\theta} KL(\tilde{\mathbf{P}}|\mathbf{P}^{\theta}) \\ \text{where} \quad &\mathbf{P}^{\theta} = \arg \min_{\mathbf{P}\in U(\mathbf{a},\mathbf{b})} < \mathbf{C}^{\theta}, \mathbf{P} > -\epsilon H(\mathbf{P}). \end{aligned} \tag{5}$$

Indeed in many previous IOT works, they directly optimize over C^{θ} which usually only involves a learnable distance between samples rather than the sample features. In the context of CL, it is more aimed to learn the feature representation hence the implication is different and the resulting method is also different.

2.3. Contrastive Learning Loss

Recently, self-supervised methods based on contrastive learning have drawn increasing attention (Logeswaran & Lee, 2018), which enables effective label-free pretraining. (Wu et al., 2018) 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. We revisit typical contrastive losses to better position our matching based methods especially for its ability for understanding and rethinking the following two main classes of CL forms.

InfoNCE Loss and Mutual Information Maximization View. Given two unlabeled data point sets $\{\mathbf{x}_i\}_{i=1}^m$ and $\{\mathbf{y}_j\}_{j=1}^n$ where $(\mathbf{x}_i, \mathbf{y}_i)$ is semantically related e.g. the raw image and its rotated version, the popular InfoNCE loss for CL is specified as (Oord et al., 2018):

$$\mathcal{L}_{\text{InfoNCE}} = -\sum_{i=1}^{n} \log \left(\frac{\exp(s_{ii}/\tau)}{\sum_{k\neq i} \exp(s_{ik}/\tau) + \exp(s_{ii}/\tau)} \right), \tag{6}$$

Here $\mathbf{s}_{ij} = s(\mathbf{z}_i, \mathbf{z}'_j)$ is a similarity (e.g. cosine) between the feature $\mathbf{z}_i$ and $\mathbf{z}'_j$, where $\mathbf{z}_i = f(\mathbf{x}_i)$ and $\mathbf{z}'_j = g(\mathbf{y}_j)$ with two feature extractors $f(\cdot)$ 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. (Chen et al., 2021a) also generalized the InfoNCE loss by adding alignment and distribution losses, and employed the sliced Wasserstein distance to support diverse prior distributions. However, it still interprets the contrastive loss with mutual information, and some works, such as (Tschannen et al., 2019), disagree with the lower bound interpretation. It has been empirically observed that maximizing a tighter bound can lead to worse performance in downstream tasks.

We will give a new interpretation with the matching view in this paper, which in fact also well fits with our empirical study.

Contrastive Loss based on Alignment and Uniformity. (Wang & Isola, 2020) views CL as enforcing two properties: alignment and uniformity of feature distributions on the output unit hypersphere:

$$\mathcal{L}_{\text{align}} = \sum_i ||\mathbf{z}_i - \mathbf{z}'_i||_2^2 \quad \text{and} \quad \mathcal{L}_{\text{uniform}} = \log \sum_{i,j} e^{2||\mathbf{z}_i - \mathbf{z}'_j||_2^2}. \tag{7}$$

where all features $\{\mathbf{z}_i\}$ and $\{\mathbf{z}'_i\}$ are ℓ_2 normalized (i.e. $||\mathbf{z}_i||_2 = ||\mathbf{z}'_i||_2 = 1$). 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 restricting the output space to the unit hypersphere. In this paper, instead of reducing their L2 norm, alignment in our approach refers to improving the chance of matching of the positive pairs while uniformity refers to guiding the probability values of all negative pairs to be close to each other, which generalizes the implication of alignment and uniformity from the probablistic matching perspective.

3. Set Matching with IOT for Deriving New Family of Contrastive Learning Losses

Overview. We give an overview to our loss family. By varying the relaxation in U of Eq. 8 in Sec. 3.1, we can get different contrastive losses. When $U = U(\mathbf{a})$ in Sec. 3.2, our loss from Eq. 8 can be simplified to InfoNCE as shown in Eq. 13 or Eq. 18. While setting $U = U(1)$ or $U = U(\mathbf{a}, \mathbf{b})$ in Sec. 3.3, we can get the new losses as shown in Eq. 20 and Eq. 23, respectively. At last, akin to the CL's uniformity in (Wang & Liu, 2021), a new uniformity penalty is devised with probabilistic matching by Eq. 24 in Sec. 3.4

3.1. Formulating IOT induced CL as Set Matching

We propose a point set matching approach with Inverse Optimal Transport for CL (IOT-CL), which studies contrastive learning by matching two feature point sets $\{\mathbf{z}_i\}_{i=1}^n$ and $\{\mathbf{z}'_j\}_{j=1}^m$ where $\mathbf{z}_i = f(\mathbf{z}_i)$ and $z'_j = g(y_j)$. Different from previous formulation of IOT, we do two generalizations: 1) cost parameterization with two feature point sets to learn the feature extractor; 2) constraint relaxation of the coupling in IOT. The cost matrix $\mathbf{C}^{\theta} \in \mathbb{R}_+^{n\times m}$ is designed with features $\{\mathbf{z}_i\}_{i=1}^n$ and $\{z'_j\}_{j=1}^m$ with parameters θ from the networks f and g. Without loss of generality, we set $\mathbf{C}_{ij}^{\theta}$ as the cosine distance between the vectors $\mathbf{z}_i$ and $\mathbf{z}'_j$, then the IOT-CL

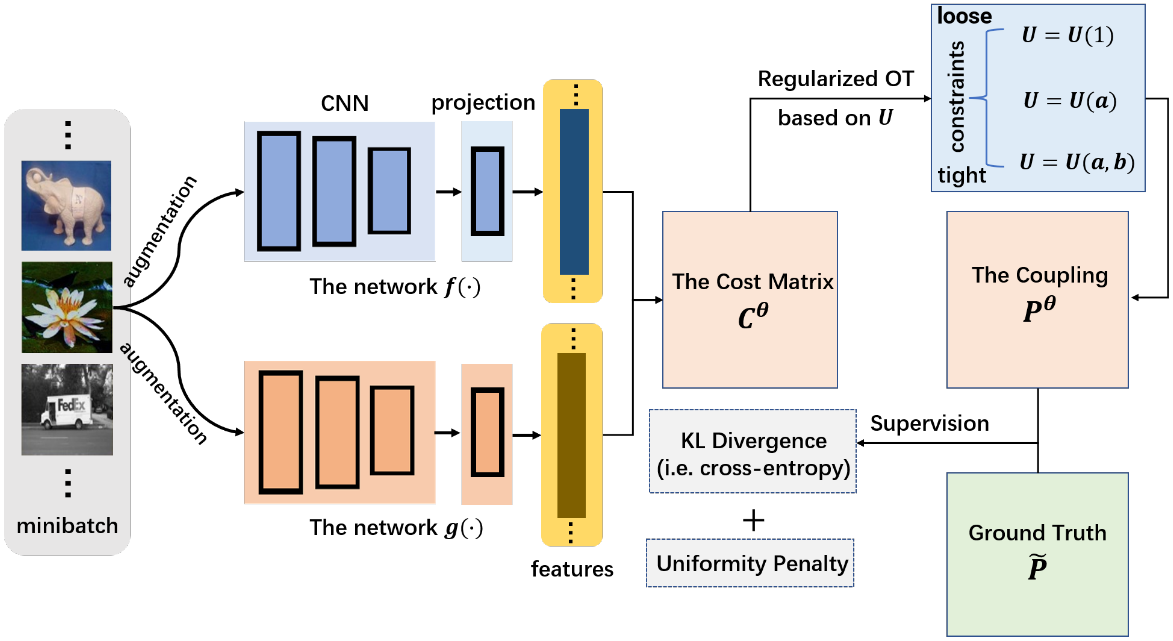


Figure 2. The overview of our approach for CL. The regularized OT is used to analyze and estimate the coupling (matching), which is supervised with ground truth matching for representation learning. Given a minibatch samples, features $\{\mathbf{z}_i\}_{i=1}^{n}$ and $\{\mathbf{z}'_j\}_{j=1}^{m}$ are extracted from the neural networks $f(\cdot)$ and $g(\cdot)$ and the cost matrix C^{θ} can be calculated to evaluate the distance among features. Then under different constraints in U, we can get the coupling with Eq. 19, Eq. 12 or Eq. 22 and supervise the coupling with KL divergence (i.e. cross-entropy loss). Besides, the uniformity penalty is also used here to improve the contrastive learning.

involves bilevel optimization as:

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

Here $H(\mathbf{P})$ is the entropic regularization as defined in Eq. 4. Different from previous works setting $U = U(\mathbf{a}, \mathbf{b})$, we think U can be designed according to the specific circumstances of the problem, especially in the case of CL. We will discuss it in detail in the next subsection. In the outer minimization, the coupling $\mathbf{P}^{\theta}$ is calculated by OT's inner minimization and $\tilde{\mathbf{P}}$ is the ground truth for supervision depending on the positive/negative pairs. For example, we can set $\tilde{\mathbf{P}}_{ij} = 1$ when $\mathbf{z}_i$ and $\mathbf{z}'_j$ are positive pairs and 0 otherwise.

The aim of outer minimization is to supervise the soft matching with the ground truth to learn the feature extractor (i.e. representation learning) parameterized by θ. In the 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 $\mathbf{a} = \mathbf{1}/n$ and $\mathbf{b} = \mathbf{1}/m$, we can relax the constraints in U as:

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

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

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

which involves basic probability requirements for coupling. In practical sense, $U(\mathbf{a}, \mathbf{b})$ asks the equal contribution of each sample in a and b. And $U(a)$ only makes the sample equality in a, while the contribution in b can be different. And when $U = U(\mathbf{a}, \mathbf{b})$, we do not ask for any contribution equality in $\mathbf{a}$ and $\mathbf{b}$, and it depends on the training of neural networks. 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 $U(\mathbf{a}, \mathbf{b})$ in detail and show the generality of our contrastive loss.

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 with rewriting the inner minimization for solving the coupling $\mathbf{P}^{\theta}$:

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

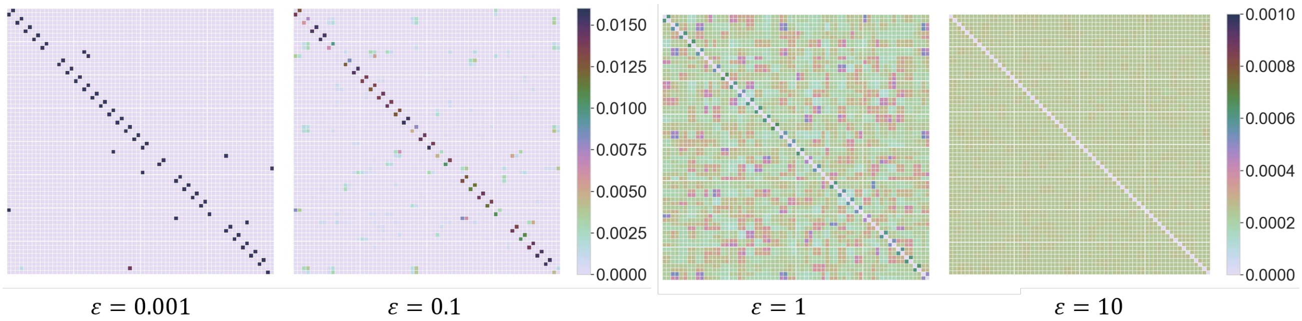


Figure 3. Results of couplings $\mathbf{P}^\theta$ by varying ϵ given 64 trained features on CIFAR-10 based on the SimCLR framework (Chen et al., 2020). When $\epsilon \to 0$, $\mathbf{P}^\theta$ becomes sharper for probability prediction. With the increment of ϵ, $\mathbf{P}^\theta$ becomes more uniform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

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

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

The proof is given in Appendix B.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 $\tilde{\mathbf{P}}_{ii} = \frac{1}{n}$ for each i and $\tilde{\mathbf{P}}_{ij} = 0$ when $i \neq j$, under $U(\mathbf{a})$ the outer minimization in fact leads to the following contrastive loss which we introduce in this paper as our IOT-CL loss (in fact a family of losses as will be shown later):

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

where con is a constant. We can easily find the equivalence between Eq. 13 and the InfoNCE loss in Eq. 6 if we set $\mathbf{C}^\theta_{ij} = c - s_{ij}$ (given c is large enough). It shows that we can understand the InfoNCE with the OT-based soft matching view and existing well-established theoretical results in entropic regularized OT may help better understand 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:

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

where s^i_{max} is the maximum of $\{\mathbf{s}_{i,:}\}$ with the anchor feature $\mathbf{z}_i$. 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.

Connecting to Softmax Cross-Entropy Loss with IOT. We can also view the classification under our point set matching perspective based on entropic regularized OT. Assume that the feature $\mathbf{z}_i = f(\mathbf{x}_i)$ 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 $\mathbf{C}^\theta_{ij} = c - \mathbf{z}_i\mathbf{o}_j = c - \mathbf{z}_{ij}$ where c is large enough and $\mathbf{o}_j$ is one-hot vector with one value on index j , then we can‘ obtain:

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^m \mathbf{y}_{ij}\log\frac{e^{\mathbf{z}_{ij}}}{\sum_{k=1}^m e^{\mathbf{z}_{ik}}}, \tag{15}$$

where $\mathbf{y}_{ij} \in \{0, 1\}$ (resp. $\mathbf{z}_{ij}$) is the j-th dimension value of one-hot vector $\mathbf{y}_i$ (resp. logits $\mathbf{z}_i$). The above loss is exactly the cross-entropy loss. This verifies that classification can also be viewed as a (soft) matching problem. Besides, InfoNCE and Softmax cross-entropy loss can be understood under this matching view with aid of regularized OT.

Learning with Matching. A popular way is to select the negative samples from their own minibatch samples or corresponding augmentations e.g. InvaSpread (Ye et al., 2019) and SimCLR (Chen et al., 2020), and in this case, we can find $m = n$ which means the balanced matching. We take SimCLR framework as an example with the matching view. Given N minibatch samples, the two point sets are in the same space within $2N$ data points and in this case, the matching is balanced as shown in Fig. 6(b). Specifically, with features $\{\mathbf{z}_i\}_{i=1}^N$ and $\{\mathbf{z}'_j\}_{j=1}^N$, we can reset the features as $\tilde{\mathbf{z}}_{2k-1} = \mathbf{z}_k$ and $\tilde{\mathbf{z}}_{2k} = \mathbf{z}'_k$ when $k = 1, 2, \ldots, N$. The new cosine similarity is specified as $\tilde{\mathbf{z}}_{ij} = \tilde{\mathbf{z}}_i \cdot \tilde{\mathbf{z}}_j/(||\tilde{\mathbf{z}}_i|| \cdot ||\tilde{\mathbf{z}}_j||)$. In this SimCLR case, the

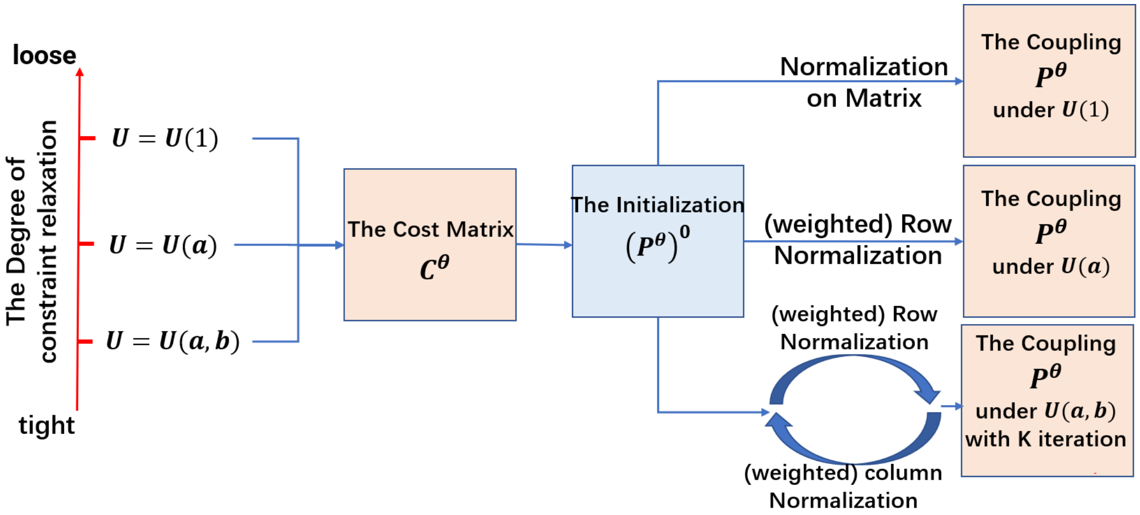

loose
The Degree of constraint relaxation
tight
$U = U(1)$
$U = U(a)$
$U = U(a, b)$
The Cost Matrix C^{θ}
The Initialization $(P^{\theta})^0$
Normalization on Matrix
(weighted) Row Normalization
(weighted) Row Normalization
(weighted) column Normalization
The Coupling P^{θ} under $U(1)$
The Coupling P^{θ} under $U(a)$
The Coupling P^{θ} under $U(a, b)$ with K iteration

Figure 4. Comparison of different losses proposed in the paper, with the coupling by varying the degree of constraint relaxation. With different degree of constraints of the coupling, we can get a family of contrastive loss uder $U = U(1), U(\mathbf{a})$ and $U(\mathbf{a}, \mathbf{b})$.

cost matrix $\mathbf{C}^{\theta}$ and ground truth $\tilde{\mathbf{P}}$ read

$$\mathbf{C}^{\theta}_{ij} = \begin{cases} +\infty, & i = j, \\ 1 - \tilde{\mathbf{z}}_{ij}, & \text{else.} \end{cases} \quad \text{and} \quad \tilde{\mathbf{P}}_{ij} = \begin{cases} \frac{1}{n}, & (i, j) \in S, \\ 0, & \text{else.} \end{cases} \tag{16}$$

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

$$\begin{aligned} S_1 &= \{(i, j) | i = 2k, j = 2k - 1, k = 1, \ldots, N\}, \\ S_2 &= \{(i, j) | i = 2k - 1, j = 2k, k = 1, \ldots, N\}. \end{aligned} \tag{17}$$

When $i = j$, we set $\mathbf{C}^{\theta}_{ij} \to +\infty$, to disallow self matching. Then we can get $\exp(-\mathbf{C}^{\theta}_{ii}/\epsilon) \to 0$. Then from Eq. 8, our IOT-CL based contrastive loss becomes:

$$\mathcal{L}^{\text{SimCLR}}_{\text{IOT-CL}} = -\frac{1}{2N} \sum_{(i,j) \in S} \log \left(\frac{\exp(-\mathbf{C}^{\theta}_{ij}/\epsilon)}{\sum_{s=1}^{2N} 1_{i \neq s} \exp(-\mathbf{C}^{\theta}_{is}/\epsilon))} \right), \tag{18}$$

which is exactly the contrastive loss in SimCLR (Chen et al., 2020). Thus SimCLR can be interpreted by the above balanced matching view as further illustrated in Fig. 6(b) in Appendix. When $m \neq n$, we discuss the unbalanced matching case in Appendix C. Note MoCo (He et al., 2020) can also be interpreted by the unbalanced matching view.

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 $\mathbf{U(1)}$: Analytical Solution. To propose the new contrastive loss, we first relax the constraints by setting $U = U(1)$. Then we get the coupling $\mathbf{P}^{\theta}$

$$\mathbf{P}^{\theta}_{ij} = \frac{\exp(-\mathbf{C}^{\theta}_{ij}/\epsilon)}{\sum_{t=1}^{n} \sum_{s=1}^{m} \exp(-\mathbf{C}^{\theta}_{ts}/\epsilon))}, \tag{19}$$

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

$$\mathcal{L}^{U(1)}_{\text{IOT-CL}} = -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{\exp(-\mathbf{C}^{\theta}_{ii}/\epsilon)}{\sum_{t=1}^{n} \sum_{s=1}^{m} \exp(-\mathbf{C}^{\theta}_{ts}/\epsilon))} \right), \tag{20}$$

Proof of Eq. 19, Eq. 20 are given in Appendix B.2. The main difference between these losses 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 involves all the elements in $\mathbf{C}^{\theta}$.

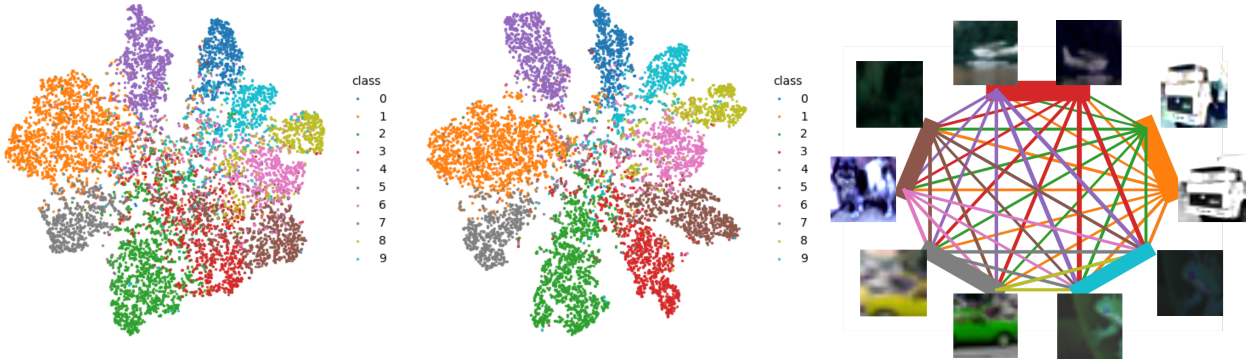
Contrastive Loss under $\mathbf{U}(\mathbf{a}, \mathbf{b})$: Numerical Solution. We propose another new loss for IOT-CL by tightening the constraint relaxation (i.e. fulfilling full constraints of matching in U). Similarly in Sec. C, we first solve the inner minimization in Eq. 8. As discussed in (Cuturi,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

$$\left(\mathbf{P}^{\theta}\right)^0 = \exp\left(-\mathbf{C}^{\theta}/\epsilon\right), \tag{21}$$

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

$$\begin{aligned} \left(\mathbf{P}^{\theta}\right)^k_{\text{temp}} &= \frac{1}{n} \left(\mathbf{P}^{\theta}\right)^{k-1} \oslash \left(\left(\mathbf{P}^{\theta}\right)^{k-1} \mathbf{1}_{m \times m} \right), \\ \left(\mathbf{P}^{\theta}\right)^k &= \frac{1}{m} \left(\mathbf{P}^{\theta}\right)^k_{\text{temp}} \oslash \left(\left(\mathbf{1}_{n \times n} \mathbf{P}^{\theta}\right)^k_{\text{temp}} \right). \end{aligned} \tag{22}$$

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. Ex-



(a) without uniformity penalty (b) with uniformity penalty (c) matching results with penalty

Figure 5. (a) and (b) are T-SNE visualizations of the embedding without and with the uniformity penalty on the coupling 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 augmentation, we visualize the matching by SimCLR. The thickness of lines is proportional to the value in $\mathbf{P}^\theta$.

Algorithm 1 Computing the IOT induced Contrastive loss (IOT-CL) under $U(\mathbf{a}, \mathbf{b})$ (under the framework of SimCLR)

```
Input: mini-batch features $\{\mathbf{z}_i\}_{i=1}^N$ and $\{\mathbf{z}'_j\}_{j=1}^N$ (see Sec. 3.2), regularization coefficient $\epsilon$, iterative number $K$
Output: IOT-CL loss value under $U(\mathbf{a},\ \mathbf{b})$
  z = Cat($\{\mathbf{z}_i\}$, $\{\mathbf{z}'_j\}$)
  Cost = $1-$CosSim($\mathbf{z}$, $\mathbf{z}^\top$)
  diags = $C$ * Eye(Sim.shape[0])        #C is large enough
  diags$_{\text{off}}$ = ($1-$eye) * Cost
  Cost = diags $+$ diags$_{\text{off}}$
  Probs = Exp($-$Cost/$\epsilon$)        #initialization of $\mathbf{P}^\theta$
  for $i = 1, \dots, K$ do
    Probs = Div(Probs,Sum(Probs, dim=0)).T        #row normalization
    Probs = Div(Probs, Sum(Probs, dim=0)).T        #column normalization
  end for
  Prob$_i$ = Diag(Probs, $N$)
  Prob$_j$ = Diag(Probs, $-N$)        #Selecting the positive pairs
  Prob$_+$ = Cat((Prob$_i$, Prob$_j$), dim=0)
  Loss = $-$log(Prob$_+$).sum()/$2N$
  return Loss
```

actly the Sinkhorn algorithm works iteratively by taking $1/n$ weighted row normalization and $1/m$ weighted column normalization according to Eq. 22 alternatively. By iterating over Eq. 22 for K rounds, we can get the coupling results. 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. When $K \to \infty$, $\left(\mathbf{P}^\theta\right)^K$ will converge to optimal solution under $U(\mathbf{a}, \mathbf{b})$. Thus increasing K is 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 $U(\mathbf{a}, \mathbf{b})$:

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

where K is iterative number and $\tilde{\mathbf{P}}$ is the ground truth. With SimCLR structure, Algorithm 1 shows computing the IOT induced Contrastive loss (IOT-CL) under $U(\mathbf{a}, \mathbf{b})$.

3.4. Enhancing Uniformity for IOT-CL

Following (Wang & Isola, 2020; Wang & Liu, 2021) emphasizing alignment and uniformity on the hypersphere for CL, we can rethink these two 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 is:

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

where $\mathcal{L}_{\text{IOT-CL}}$ can take any form of the aforementioned contrastive loss in our loss family, λ_p is the uniformity penalty coefficient and $\bar{\mathbf{Q}}^\theta$ reads:

$$\bar{\mathbf{Q}}^\theta_{ij} = \begin{cases} \mathbf{P}^\theta_{ij}, & (i,j) \in S, \\ \underset{(i',j')\notin S}{\text{mean}} \{\mathbf{P}^\theta_{i'j'}\}, & (i,j) \notin S. \end{cases} \tag{25}$$

The first term in Eq. 24 represents the matching alignment, which increases the probability of positive pair matching.

Table 1. Top-1 classification accuracy (%) of using the proposed IOT-CL loss (without uniformity penalty) evaluated by linear networks when varying the relaxation of constraints in U with 100/200 epochs of 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 (see Sec. 3.3).

Variants of $\mathcal{L}_{\text{IOT-CL}}$	**CIFAR-10**		**CIFAR-100**		**SVHN**	
	100	200	100	200	100	200
$U(1)$	90.53	90.49	**67.15**	66.97	91.50	91.70
$U(\mathbf{a})$	90.61	90.57	66.34	66.75	**91.64**	**91.85**
$U(\mathbf{a}, \mathbf{b})$ w/ $K = 1$	90.55	90.56	66.74	66.91	91.56	91.77
$U(\mathbf{a}, \mathbf{b})$ w/ $K = 2$	**91.06**	90.99	66.75	**67.23**	91.19	91.26
$U(\mathbf{a}, \mathbf{b})$ w/ $K = 4$	90.66	90.49	66.87	67.07	91.12	91.33
$U(\mathbf{a}, \mathbf{b})$ w/ $K = 8$	90.98	**91.02**	66.61	66.73	91.14	91.28

Table 2. Top-1 accuracy (%) by linear classifier network/k-NN based on 100 training epochs for different contrastive losses.

CL loss form	**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)	90.80	86.68	66.71	53.25	91.72	77.33
InfoNCE (Chen et al., 2020)	90.61	86.79	66.34	52.98	91.64	75.25
$\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{uniform}}$ (Wang & Isola, 2020)	90.84	86.53	66.28	55.10	91.90	82.56
HNSampling (Wang & Liu, 2021)	90.26	86.88	65.81	52.84	92.37	80.47
$\mathcal{L}_{\text{IOT-CL}}$ (w/o penalty)	**91.06**	87.07	67.15	53.76	91.64	77.66
$\mathcal{L}_{\text{IOT-CL}}^{\text{uniform}}$ $(U - U(\mathbf{a}))$	90.98	**87.58**	**67.59**	**55.72**	**93.22**	**84.15**

The second term encourages the uniformity, where $\bar{\mathbf{Q}}_{\mathbf{ij}}^{\theta}$ is the mean of matching probability for negative pairs when $(i, j) \notin S$. We let $\mathbf{P}^{\theta}$ approximate $\bar{\mathbf{Q}}^{\theta}$ by using KL divergence, which decreases the volatility as well as uniformity.

3.5. Potential Impact to IOT Itself

Beyond contrastive learning, we show that one can relax the constraints (or its relaxation degree by varying K) for learning the cost matrix in IOT. The rationale behind our relaxation scheme is that we find it can reduce the need for large iteration number K and still achieves even better results than large K with full convergence. This is empirically supported by Table 1 with $U(1)$, $U(a)$ i.e. with constraint relaxation and different K values from 1 to 8 with $U(\mathbf{a}, \mathbf{b})$.

We also try to connect OT and Softmax-based losses. We think it is helpful for the application of IOT and one can naturally link OT with the research direction corresponding to Softmax such as classification, long-tailed recognition, adversarial training etc, which we leave for future work.

4. Experiments

4.1. Evaluation Protocols

Experiments run on a single RTX-3090 (24GB) GPU and 128G memory, 24 physical CPU with 3.50GHz.

Datasets and Pretraining. We test our loss 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 pretraining, we adopt Resnet50 (He et al., 2016) as the backbone. The augmentations follow (Chen et al., 2020) with random color distortions, Gaussian blur, and cropping. 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. All the models are pretrained on CIFAR-10/200 and SVHN with Adam (Kingma & Ba, 2014) for 500 epochs by 3e-4 learning rate with a mini-batch size of 128[1], while the pretraining epoch is reduced to 100 for ImageNet-100. We set $\tau = 0.5$ 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 200 epochs with 256 mini-batch sizes. We also compare a k-nearest neighbors classifier (k-NN, k = 5 here) with our linear evaluation. k-NN does not require additional parameters, which is applicable without training.

Compared Methods. In addition to the contrastive losses in Sec. 2.3 (i.e. InfoNCE (Oord et al., 2018), HNSamp-

[1] Exception is made to Triplet (Schroff et al., 2015) which only takes 200 epochs in Table 2 and more training hurts performance.

Table 3. Top-1 accuracy (%) of IOT-CL loss $(U - U(\mathbf{a}))$ with uniformity penalty by linear networks with different uniformity coefficient λ_p and different training epochs.

λ_p	**CIFAR-10**		**CIFAR-100**		**SVHN**	
	100	200	100	200	100	200
0	90.61	90.57	66.34	66.75	91.64	91.85
0.1	90.80	90.47	67.47	**67.21**	91.15	91.33
0.5	90.98	**90.99**	67.56	67.17	92.49	92.39
1	90.56	90.68	**67.59**	67.09	**93.22**	**93.05**
5	90.84	90.66	67.03	66.67	92.93	92.91
10	**90.99**	90.96	66.01	65.64	92.34	92.38

ing (Wang & Liu, 2021), and $\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{uniform}}$ (Wang & Isola, 2020)), we compare our losses of IOT-CL with the triplet loss (Schroff et al., 2015) and the loss in InvaSpread (Ye et al., 2019). The triplet loss can be understood as a special case of InfoNCE when the temperature $\tau \to 0$, while the loss in InvaSpread is similar with the perturbation loss of graph matching (Wang et al., 2019b), which learns the probability of positive/negative pairs with Bernoulli distribution.

4.2. Main Results

Impact of Constraint Relaxation. Table 1 show the results as evaluated by liner network on CIFAR-10, CIFAR100 and SVHN. To align the experimental setting for fairness, we re-run the compared methods and obtain their results in Table 2. We can find that the loss of IOT-CL (without uniformity penalty) can increase the performance as show in Table 2 under the simclr framework.

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 (Wang & Isola, 2020) ($\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{uniform}}$) 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. Compared with InfoNCE and its variants, the accuracy gets improved in all the datasets for our method with uniformity penalty on coupling. Thus, our method outperforms state-of-the-art methods in most cases.

Besides, as shown in Fig. 5, by adding uniformity penalty based on the loss of IOT-CL, the representation 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 and 200 epochs.

Table 4. Top-1 classification accuracy of linear classifier and k-NN for ImageNet-100 with 100 pre-training epochs.

Method	Lin.	k-NN
InfoNCE (Chen et al., 2020)	67.35	53.50
HNSampling (Wang & Liu, 2021)	65.89	51.95
InvaSpread (Ye et al., 2019)	67.78	53.64
$\mathcal{L}_{\text{IOT-CL}}^{\text{uniform}}$ $(U - U(\mathbf{a}))$	**68.32**	**55.28**

Besides, more experiments are presented in Appendix, including those based on the MoCo framework which further verify the wide effectiveness of our methods in Appendix C.

Experiments on ImageNet-100. We further test our model on ImageNet-100 and using Resnet50 as the backbone. The pretraining protocol is given in Sec. 4.1. Table 4 shows that our model with uniformity penalty ($\lambda_p = 0.5$) greatly improves the performance.

5. Conclusion and Outlook

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 losses by introducing inverse optimal transport techniques. 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, with verified effectiveness on public vision datasets.

Many directions can be further studied based on the matching perspective: 1) Consider the non-entropic regularization in OT, and we may get another family of new contrastive loss for CL; 2) Consider the matching on the unit hypersphere. Some works of optimal transport on sphere (McRae et al., 2018; Hamfeldt & Turnquist, 2021; Cui et al., 2019) or hypersphere (Tu et al., 2020) may enhance the theoretical study of CL under hypersphere space.

It is also interesting to observe that the cross-entropy loss has been well adopted in the graph matching literature (Wang et al., 2019a) beyond the regression loss as adopted in earlier literature (Zanfir & Sminchisescu, 2018), including both matching with two explicit graphs (Wang et al., 2022) and the more general Lawler's QAP (Wang et al., 2023), as well as the self-supervised setting (Liu et al., 2022). We believe these works may also bear some connection to optimal transport, which we leave for future work.

Acknowledgments

The work was supported in part by NSFC (62222607) and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

References

Adams, R. P. and Zemel, R. S. Ranking via sinkhorn propagation. *arXiv preprint arXiv:1106.1925*, 2011.

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

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

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

Chen, T., Luo, C., and Li, L. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34:11834–11845, 2021a.

Chen, X., Xie, S., and He, K. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021b.

Chiu, W.-T., Wang, P., and Shafto, P. Discrete probabilistic inverse optimal transport. In *International Conference on Machine Learning*, pp. 3925–3946, 2022.

Cui, L., Qi, X., Wen, C., Lei, N., Li, X., Zhang, M., and Gu, X. Spherical optimal transportation. *Computer-Aided Design*, 115:181–193, 2019.

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

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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

Gao, T., Yao, X., and Chen, D. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

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

Hamfeldt, B. F. and Turnquist, A. G. A convergence framework for optimal transport on the sphere. *arXiv preprint arXiv:2103.05739*, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Hu, Q., Wang, X., Hu, W., and Qi, G.-J. 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.

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

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

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

Li, R., Ye, X., Zhou, H., and Zha, H. Learning to match via inverse optimal transport. *Journal of machine learning research*, 20, 2019.

Liu, C., Zhang, S., Yang, X., and Yan, J. Self-supervised learning of visual graph matching. 2022.

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

McRae, A. T., Cotter, C. J., and Budd, C. J. Optimal-transport–based mesh adaptivity on the plane and sphere using finite elements. *SIAM Journal on Scientific Computing*, 40(2):A1121–A1148, 2018.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Schroff, F., Kalenichenko, D., and Philbin, J. 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.

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

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

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020.

Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

Tu, Y., Mi, L., Zhang, W., Zhang, H., Zhang, J., Fan, Y., Goradia, D., Chen, K., Caselli, R. J., Reiman, E. M., et al. Computing univariate neurodegenerative biomarkers with volumetric optimal transportation: a pilot study. *Neuroinformatics*, 18(4):531–548, 2020.

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

Wang, R., Yan, J., and Yang, X. Learning combinatorial embedding networks for deep graph matching. pp. 3056–3065, 2019a.

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

Wang, R., Yan, J., and Yang, X. Neural graph matching network: Learning lawler's quadratic assignment problem with extension to hypergraph and multiple-graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Wang, R., Yan, J., and Yang, X. Combinatorial learning of robust deep graph matching: an embedding based approach. *IEEE TPAMI*, 2023.

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

Wang, W., Slepčev, D., Basu, S., Ozolek, J. A., and Rohde, G. K. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101(2):254–269, 2013.

Wilson, A. G. 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.

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

Ye, M., Zhang, X., Yuen, P. C., and Chang, S.-F. 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.

Zanfir, A. and Sminchisescu, C. Deep learning of graph matching. 2018.

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

A. Description of Other Peer CL Methods

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

$$\mathcal{L}_{\text{hard}} = -\sum_{i=1}^{n} \log \left(\frac{\exp(s_{ii}/\tau)}{\sum_{k:s_{ik}>s_{\alpha}^{i}} \exp(s_{ik}/\tau) + \exp(s_{ii}/\tau)} \right) \tag{26}$$

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.

Triplet loss. The triplet loss is a famous and popular for contrastive learning (Schroff et al., 2015). We use the triplet loss in both MoCo and SimCLR framework to compare with our methods.

B. Lagrangian for Regularized OT

Our proof is as follows, which is technically akin to (Cuturi, 2013) in terms of using Lagrangian duals for Regularized OT, which in fact has been well adopted in OT literature.

B.1. Lagrangian under $U(a)$

Now we show the point set matching framework for CL with the simplified constraints:

$$U(\mathbf{a}) = \{\mathbf{P} \in \mathbb{R}_{+}^{n\times m} | \mathbf{P}\mathbf{1}_m = \mathbf{a}\} \tag{27}$$

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

$$\mathbf{P}^{\theta} = \arg\min_{P\in U(\mathbf{a})} < \mathbf{C}^{\theta}, \mathbf{P} > -\epsilon H(\mathbf{P}), \tag{28}$$

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

$$L(\mathbf{P}, \mathbf{f}) =< \mathbf{C}^{\theta}, \mathbf{P} > -\epsilon H(\mathbf{P}) - \sum_{i=1}^{n} \mathbf{f}_i \cdot \left(\sum_{j=1}^{m} \mathbf{P}_{ij} - \frac{1}{n} \right) \tag{29}$$

The first order conditions then yield by:

$$\frac{\partial L(\mathbf{P}, \mathbf{f})}{\partial \mathbf{P}_{ij}} = \mathbf{C}_{ij}^{\theta} + \epsilon \log \mathbf{P}_{ij} - \mathbf{f}_i = 0 \tag{30}$$

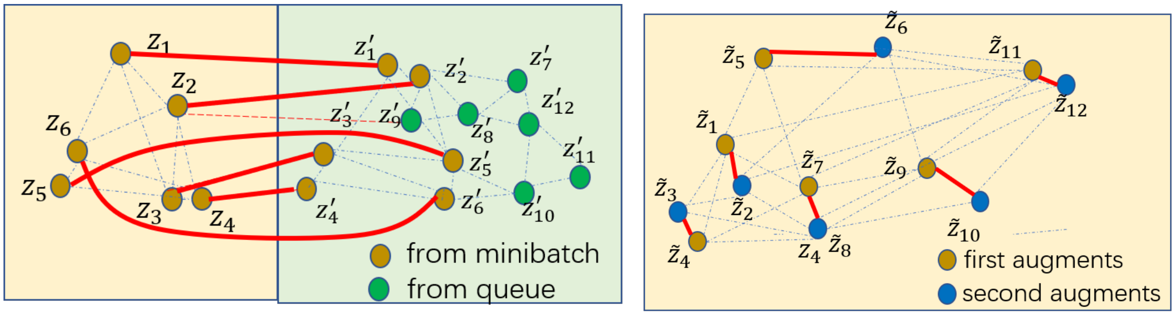
Thus we have $\mathbf{P}_{ij} = e^{(\mathbf{f}_i - C_{ij}^{\theta})/\epsilon}$ for every i and j, for optimal $\mathbf{P}$ coupling to the regularized problem. Due to $\sum_j \mathbf{P}_{ij} = 1/n$ for every i, we can calculate the Lagrangian parameter $\mathbf{f}_i$ and the solution of the coupling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

$$\mathbf{P}_{ij} = \frac{\exp\left(-\mathbf{C}_{ij}^{\theta}/\epsilon\right)}{n \sum_{t=1}^{m} \exp\left(-\mathbf{C}_{it}^{\theta}/\epsilon\right)} \tag{31}$$

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 $U(\mathbf{a})$

$$\mathcal{L}_{\text{IOT-CL}} = -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{\exp(-\mathbf{C}_{ii}^{\theta}/\epsilon)}{\sum_{j=1}^{m} \exp(-\mathbf{C}_{ij}^{\theta}/\epsilon))} \right) + \text{Constant} \tag{32}$$

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



(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, it is more than a simple matching between two sets but a balanced matching in the same space, and the feature points try to match their neighbors with minimal total cost.

B.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

$$\mathbf{P}^\theta = \arg\min_{\mathbf{P}\in U(1)} < \mathbf{C}^\theta, \mathbf{P} > -\epsilon H(\mathbf{P}), \tag{33}$$

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

$$L(\mathbf{P}, \lambda) =< \mathbf{C}^\theta, \mathbf{P} > -\epsilon H(\mathbf{P}) - \lambda(\sum_{i,j} \mathbf{P}_{ij} - 1) \tag{34}$$

First order conditions then yield by:

$$\frac{\partial L(\mathbf{P}, \lambda)}{\partial \mathbf{P}_{ij}} = \mathbf{C}^\theta_{ij} + \epsilon \log \mathbf{P}_{ij} - \lambda = 0 \tag{35}$$

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 λ and the coupling solution is written as:

$$\mathbf{P}_{ij} = \frac{\exp\left(-\mathbf{C}^\theta_{ij}/\epsilon\right)}{\sum_{st} \exp\left(-\mathbf{C}^\theta_{st}/\epsilon\right)} \tag{36}$$

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 $U(1)$:

$$\mathcal{L}_{\text{IOT-CL}} = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\exp(-\mathbf{C}^\theta_{ii}/\epsilon)}{\sum_{t=1}^{n}\sum_{s=1}^{m} \exp(-\mathbf{C}^\theta_{ts}/\epsilon))}\right) \tag{37}$$

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

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 $\{\mathbf{z}_i\}_{i=1}^n$ and $\{\mathbf{z}'_j\}_{j=1}^n$ are features

Table 5. ACC (%) by Linear network (Lin.) and k-NN for CIFAR-10/100 with 100 epochs of training, under the MoCo framework.

CL loss form	**CIFAR-10**		**CIFAR-100**	
	Lin.	k-NN	Lin.	k-NN
InfoNCE (Oord et al., 2018) (i.e. MoCo (He et al., 2020))	72.11	61.66	47.71	28.23
HNSampling (Wang & Liu, 2021)	72.18	60.73	47.88	28.06
$\mathcal{L}^{\text{uniform}}_{\text{IOT-CL}}$ (using $\mathcal{L}^{U(\mathbf{a},\mathbf{b})}_{\text{IOT-CL}}$ with $K=1$)	73.28	61.97	48.21	29.64

extracted by encoder f and momentum encoder g from the same mini-batch samples, while $\{\mathbf{z}'_j\}_{j=n+1}^{m}$ are features extracted by g from the memory bank. Then we can get two feature sets $\{\mathbf{z}_i\}_{i=1}^{n}$ and $\{\mathbf{z}'_j\}_{j=1}^{m}$ 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 selects 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

$$\mathbf{C}^{\theta}_{ij} = \begin{cases} +\infty, & i \neq j \text{ and } 1 \leq j \leq n \\ 1 - \mathbf{z}_{ij}, & \text{else} \end{cases} \tag{38}$$

where $\mathbf{z}_{ij}$ is a similarity (e.g. cosine) between feature $\mathbf{z}_i$ and $\mathbf{z}'_j$. Here $\mathbf{C}^{\theta}_{ij} \to +\infty$ implies $\exp(-\mathbf{C}^{\theta}_{ij}/\epsilon) \to 0$, which means that $\mathbf{P}^{\theta}_{ij} \to 0$ and the features $\mathbf{z}_i$ and $\mathbf{z}'_j$ will not be matched. The condition $i \neq j$ and $1 \leq j \leq n$ represents that $(\mathbf{z}_i, \mathbf{z}'_j)$ are negative pairs from the same mini-batch, which is not adopted as negative pair in memory based methods.

With the cost matrix $\mathbf{C}^{\theta}$, the ground truth can be simply set as $\tilde{\mathbf{P}}_{ii} = \frac{1}{n}$ for $i = 1, \dots, n$ and $\tilde{\mathbf{P}}_{ij} = 0$ 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 epochs 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.

C.1. Remarks on the Limitation of Our Approach

Note that in our SimCLR-based version (results in the main paper), both IOT and SimCLR share the similar computational overhead in terms of the feature dimension. In another word, our new formulation does not increase the complexity.

While in a Moco-based version which involves memory bank (results in Table 5), there would incur a skewed setting: few-matching-many e.g. 256-vs-1024 in contrast to the 256-vs-256 matching in the SimCLR-based version. In our experiments for Table 5, we enforce the uniform penalty to mitigate this issue. In theory our method is less suitable for the Moco framework due the memory bank scheme. Fortunately, moco v3 (Chen et al., 2021b) has dismissed the memory bank scheme.