Your contrastive learning problem is secretly a distribution alignment problem

Zihao Chen*, **Chi-Heng Lin, Ran Liu, Jingyun Xiao, Eva L. Dyer***
School of Electrical & Computer Engineering
Georgia Tech, Atlanta, GA

Abstract

Despite the success of contrastive learning (CL) in vision and language, its theoretical foundations and mechanisms for building representations remain poorly understood. In this work, we build connections between noise contrastive estimation losses widely used in CL and distribution alignment with entropic optimal transport (OT). This connection allows us to develop a family of different losses and multistep iterative variants for existing CL methods. Intuitively, by using more information from the distribution of latents, our approach allows a more distribution-aware manipulation of the relationships within augmented sample sets. We provide theoretical insights and experimental evidence demonstrating the benefits of our approach for *generalized contrastive alignment*. Through this framework, it is possible to leverage tools in OT to build unbalanced losses to handle noisy views and customize the representation space by changing the constraints on alignment. By reframing contrastive learning as an alignment problem and leveraging existing optimization tools for OT, our work provides new insights and connections between different self-supervised learning models in addition to new tools that can be more easily adapted to incorporate domain knowledge into learning.

1 Introduction

In machine learning, the availability of vast amounts of unlabeled data has created an opportunity to learn meaningful representations without relying on costly labeled datasets [26, 52, 27]. Self-supervised learning has emerged as a powerful solution to this problem, allowing models to leverage the inherent structure in data to build useful representations. Among self-supervised methods, contrastive learning (CL) is widely adopted for its ability to create robust representations by distinguishing between similar (positive) and dissimilar (negative) data pairs. With success in fields like image and language processing [8, 46], contrastive learning now also shows promise in domains where cross-modal, noisy, or structurally complex data make labeling especially challenging [34, 56, 10].

Traditional contrastive learning methods primarily aim to bring positive pairs—often augmentations of the same sample—closer together in representation space. While effective, this approach often struggles with real-world challenges such as noise in views, variations in data quality, or shifts introduced by complex transformations, where positive pairs may not perfectly align. Additionally, in tasks requiring domain generalization, aligning representations across diverse domains (e.g., variations in style or sensor type) is critical but difficult to achieve with standard contrastive learning, which typically lacks mechanisms for incorporating domain-specific relationships. These limitations highlight the need for a more flexible approach that can adapt alignment strategies based on the data structure, allowing for finer control over similarity and dissimilarity among samples.

*Contact: {zchen959, evadyer}@gatech.edu

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

To address this challenge, we introduce a novel *generalized contrastive alignment* (GCA) framework, which reinterprets contrastive learning as a distributional alignment problem. Our method allows flexible control over the alignment of samples by defining a target transport plan, $\mathbf{P}_{tgt}$, that serves as a customizable alignment guide. For example, setting $\mathbf{P}_{tgt}$ to resemble a diagonal matrix encourages each positive to align primarily with itself or its augmentations, thereby reducing the effect of noise between views. Alternatively, we can incorporate more complex constraints, such as weighting alignments based on view quality or enforcing partial alignment structures where noise or data heterogeneity is prevalent. This flexibility enables GCA to adapt effectively to a wide range of tasks, from simple twin view alignments to scenarios with noisy or variably aligned data.

Our approach also bridges connections between GCA and established methods, such as InfoNCE (INCE) [38], Robust InfoNCE (RINCE) [12], and BYOL [22], demonstrating that these can be viewed as iterative alignment objectives with Bregman projections [6, 21]. This perspective allows us to systematically analyze and improve uniformity within the latent space, a property that enhances representation quality and ultimately boosts downstream classification performance.

We validate our method through extensive experiments on both image classification and noisy data tasks, demonstrating that GCA's unbalanced OT (UOT) formulations improve classification performance by relaxing our constraints on alignment. Our results show that GCA offers a robust and versatile framework for contrastive learning, providing flexibility and performance gains over existing methods and presenting a promising approach to addressing different sources of variability in self-supervised learning.

The contributions of this work include:

- A new framework called *generalized contrastive alignment* (GCA), which reinterprets standard contrastive learning as a distributional alignment problem, using optimal transport to provide flexible control over alignment objectives. This approach allows us to derive a novel class of contrastive losses and algorithms that adapt effectively to varied data structures and build customizable transport plans.
- We present GCA-UOT, a contrastive learning method that achieves strong performance on standard augmentation regimes and excels in scenarios with more extreme augmentations or data corrupted by transformations. GCA-UOT leverages unbalanced transport to adaptively weight positive alignments, enhancing robustness against view noise and cross-domain variations.
- We provide theoretical guarantees for the convergence of our GCA-based methods and show that our alignment objectives improve representation quality by enhancing the uniformity of negatives and strengthening alignment within positive pairs. This leads to more discriminative and resilient representations, even in challenging data conditions.
- Empirically, we demonstrate the effectiveness of GCA in both image classification and domain generalization tasks. Through flexible, unbalanced OT-based losses, GCA achieves superior classification performance and adapts alignment to include domain-specific information where relevant, without compromising classification accuracy in domain generalization.

2 Background

2.1 Contrastive learning

Contrastive learning (CL) is a representation learning methodology that uses positive and negative pairs to define similarity in the latent space. Let $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$ denote our dataset. For each sample $\mathbf{x}_i$ in a batch of training data with size B, we create two augmented copies $\mathbf{x}'_i$ and $\mathbf{x}''_i$ independently, i.e., $\mathbf{x}'_i = \psi(\mathbf{x}_i)$ where ψ is a randomly drawn augmentation function from some augmentation class $\mathcal{A}$ and likewise for $\mathbf{x}''_i$. The $(\mathbf{x}'_i, \mathbf{x}''_i)$ is called a positive pair of $\mathbf{x}_i$ while $(\mathbf{x}'_i, \mathbf{x}''_j)$ is treated as a negative pair for any $j \neq i$. One of the most widely used formulations of the CL problem, InfoNCE (INCE) [8], seeks to maximize the negative log probability that a sample is correctly classified as

$$\mathcal{L}_{\text{INCE}} = -\log\left(\frac{e^{s_{ii}}}{e^{s_{ii}} + \sum_{i\neq j} e^{s_{ij}}}\right), \tag{1}$$

where $s_{ij} = \varepsilon^{-1} f_\theta(\mathbf{x}'_i)^\top f_\theta(\mathbf{x}''_j)/\|f_\theta(\mathbf{x}'_i)\|\|f_\theta(\mathbf{x}''_j)\|$ is the score between augmented samples.

Building upon the principles of INCE, SimCLR [8] and MoCo [24] are two representative works that form the foundation of contrastive learning methods for visual representation tasks. Alternatively, BYOL [22] and SimSiam [9] discard the use of negative samples to avoid large batch size and instead use exponential moving average-based updates to avoid representational collapse. Recent contrastive methods have focused on improving the tolerance to noise in samples to enhance robustness in diverse scenarios [13]. Among them, Robust INCE (RINCE) is a robust contrastive loss function characterized by its symmetric properties and theoretical resistance to noisy labels [47, 12]. Specifically, RINCE provides robustness to noisy views by introducing adjustable parameters λ and q [12] which rebalance the cost of positive and negative views, resulting in the following loss:

$$\mathcal{L}_{\text{RINCE}}^{\lambda,q} = \frac{1}{q}\big(- e^{q s_{ii}} + \lambda^q (e^{s_{ii}} + \sum\nolimits_{i \neq j} e^{s_{ij}})^q \big) \tag{2}$$

By optimizing the above loss functions, the encoder f is trained to construct a semantically coherent representation space where positive pairs of samples are positioned nearby, while those negative pairs with divergent semantic attributes are separated [57].

2.2 Proximal Operators and Projections

To make the connections between different CL losses clearer later, we use the notion of proximal operators. In words, the proximal operator will provide a way to find the closest point in some closed convex set. Formally, we can define the proximal operator as follows.

Definition 1 (Proximal Operator)**.** *Let $d_\Gamma(\mathbf{x}, \mathbf{v}) = \Gamma(\mathbf{x}) - \Gamma(\mathbf{v}) - \langle \nabla\Gamma(\mathbf{v}), \mathbf{x} - \mathbf{v}\rangle$ be a Bregman divergence with a convex function Γ. The proximal operator of $h : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is defined for a point $\mathbf{v} \in \mathcal{X}$ with a closed convex set $\mathcal{B} \subseteq \mathcal{X}$:*

$$\mathit{Prox}_{h,\mathcal{B}}^{d_\Gamma}(\mathbf{v}) = \arg\min_{\mathbf{x} \in \mathcal{B}} \{h(\mathbf{x}) + d_\Gamma(\mathbf{x}, \mathbf{v})\}\,.$$

Moreover, we can define the concept of a projection as a special case of the proximal operator when we let $h(\mathbf{x})$ be an indicator function $h_{\mathcal{B}}(x) = \{0, \text{if } x \in \mathcal{B}; \infty, \text{if } x \notin \mathcal{B}\}$ on constraint set $\mathcal{B}$. See Appendix A.2 for more details.

2.3 Solving Optimal Transport Through Proximal Point Methods

Optimal transport (OT) is widely used in characterizing the distance between two collections of samples $\{\mathbf{x}_i\}_{i=1}^B$ and $\{\mathbf{y}_j\}_{j=1}^B$ with associated measures $\mu = \sum_{i=1}^B \delta_{\mathbf{x}_i} p_i$ and $\nu = \sum_{j=1}^B \delta_{\mathbf{y}_j} q_j$ with Dirac delta function $\delta_{\mathbf{x}}$ and $\delta_{\mathbf{y}}$ on finite support [43]. Here, p and q are vertices of the $\mathbb{R}^B$ simplex defined as $\Delta_B := \{v \in \mathbb{R}^B : v_i \geq 0, \sum_{i=1}^B v_i = 1\}$. OT aims to learn a joint coupling matrix, or transport plan $\mathbf{P} \in \mathbb{R}_+^{B \times B}$ that minimizes the cost of transporting mass encoded by cost matrix $\mathbf{C} \in \mathbb{R}_+^{B \times B}$, from one distribution to another. In practice, entropy regularization is used to solve the OT objective, resulting in the following entropy-regularized OT (EOT) objective:

$$\min_{\mathbf{P} \in \mathcal{B}} \langle \mathbf{P}, \mathbf{C}\rangle - \varepsilon H(\mathbf{P}), \quad \text{where } H(\mathbf{P}) = -\sum_{ij} \mathbf{P}_{ij} \log(\mathbf{P}_{ij}), \tag{3}$$

where ε is a user specified parameter that controls the amount of smoothing in the transport plan, and $\mathbf{C}(\mathbf{x}, \mathbf{y}) = 1 - \langle \mathbf{x}, \mathbf{y}\rangle / \|\mathbf{x}\|\|\mathbf{y}\|$ is often set to encode the cosine similarity between pairs of samples.

The Sinkhorn Algorithm and its Interpretation as a Bregman Projection. Solving Equation (3) could be interpreted as iterative alignment problem on a Hilbert space generated from the kernel $\mathbf{K}_{ij} = \exp(-\mathbf{C}_{i,j}/\varepsilon)$. This alignment problem can be solved through iterative Bregman projections onto the two constraints sets that encode the marginals along the rows and columns [3, 5, 43]:

$$C_1^\mu \coloneqq \{\mathbf{P} : \mathbf{P}\mathbb{1}_B = \mu\}, C_2^\nu \coloneqq \{\mathbf{P} : \mathbf{P}^\top \mathbb{1}_B = \nu\} \tag{4}$$

The first step of Bregman projection is to find the minimizer $\mathbf{P}^{(1)} = \arg\min\{\varepsilon \text{KL}(\mathbf{P} \| \mathbf{K}) : \mathbf{P}\mathbb{1}_B = \mu\}$ by the proximal operator $\text{Prox}_{C_1^\mu}^{\text{KL}}(\mathbf{K})$ with Lagrange multiplier f on the row constraint set $\mathcal{C}_1^\mu$, and compute its derivatives with respect to $\mathbf{P}$ with $\mathbf{u} = e^{f/\varepsilon} > 0$:

$$\varepsilon \log(\mathbf{P}^{(1)}/\mathbf{K}) - f\mathbb{1} = 0 \Rightarrow \mathbf{P}^{(1)} = \mathbf{u}\mathbf{K}, \quad \langle \mathbf{P}^{(1)}, \mathbb{1}\rangle = \mu \Rightarrow \langle \mathbf{u}\mathbf{K}, \mathbb{1}\rangle = \mu, \mathbf{u} = \frac{\mu}{\mathbf{K}\mathbb{1}} \tag{5}$$

Next, we project $\mathbf{P}^{(1)}$ onto the column constraint set C_2^ν, resulting in $\mathbf{P}^{(2)} \coloneqq \text{Prox}_{C_2^\nu}^{\text{KL}}(\mathbf{P}^{(1)}) = \mathbf{P}^{(1)}\text{diag}(\frac{\nu}{\mathbf{P}^{(1)\top}\mathbb{1}_B})$. The iterative updates can be succinctly expressed as the Sinkhorn iterations:

$$\mathbf{P}^{(2t+1)} = \text{diag}(\mathbf{u}^{(t+1)})\mathbf{K}\,\text{diag}(\mathbf{v}^{(t)}), \quad \mathbf{P}^{(2t+2)} = \text{diag}(\mathbf{u}^{(t+1)})\mathbf{K}\,\text{diag}(\mathbf{v}^{(t+1)}), \tag{6}$$

with the scaling vectors $\mathbf{u}^{(t)}$ and $\mathbf{v}^{(t)}$ updated according to:

$$\mathbf{u}^{(t+1)} \overset{\text{def}}{=} \frac{\mu}{\mathbf{K}\mathbf{v}^{(t)}}, \quad \mathbf{v}^{(t+1)} \overset{\text{def}}{=} \frac{\nu}{\mathbf{K}^{\mathrm{T}}\mathbf{u}^{(t)}}. \tag{7}$$

Here, iterations converge to a stable transport plan $\mathbf{P}^{(\infty)}$as the optimal solution of Equation (3), which provides the minimum cost matching between two distributions. The convergence and dynamics of OT and its dual formulation have been studied extensively in [4, 43, 19, 1]. Thus, these results guarantee that the iterates will converge to the optimal solution of the EOT objective, or that $\mathbf{P}^{(t)} \to \mathbf{P}^{(\infty)}$ with $t \to \infty$. See Appendix A.3 for more details on both the continuous and discrete formulations of OT.

2.4 Wasserstein Dependency Measure

The Wasserstein Dependency Measure (WDM) is a measure of deviation between two probability measures. We will use this later and thus provide the formal definition here [39].

Definition 2 (Wasserstein Dependency Measure)**.** *Define the WDM as the Wasserstein distance (W_1) between the joint distribution $\pi(x, y)$ and the product of marginal distributions $\mu \otimes \nu(x, y)$ of two random variables x and y. $W_1(\pi, \mu \otimes \nu) = \sup_{f \in \mathcal{C}(\mathcal{X}\times\mathcal{Y})} \big(\mathbb{E}_{\pi(x,y)}[f(x, y)] - \mathbb{E}_{\mu\otimes\nu(x,y)}[f(x, y)]\big)$, where $\mathcal{C}(\mathcal{X} \times \mathcal{Y})$ denotes the set of all 1-Lipschitz functions from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$.*

2.5 Optimal Transport and Alignment in Representation Learning

Distribution alignment and OT have been widely used for domain adaptation [33, 14, 30, 59], and in generative modeling [2, 55, 49, 58]. The connections between distribution alignment and contrastive learning, however, are still nascent. In [51], the authors explore the connection between inverse OT (IOT) [32, 53, 18] and INCE. Our work builds on this connection to OT to build robust divergences (RINCE) and to build a novel unbalanced optimal transport (UOT) method (Section 3.3). Additionally, we show how our framework can be used to build flexible methods for encouraging contrast at multiple levels. We use this concept of hierarchical contrast and show that it can be used in domain generalization settings (Section 6.2). It is of note that GCA-UOT focuses on relaxing the hard constraints on the row and columns into the soft penalties, which is different with the idea of "unbalanced matching" in [51] which considers the case where the encoders may not have the same weights.

3 Generalized Contrastive Alignment (GCA)

In this section, we will introduce a new framework for *generalized contrastive alignment* and demonstrate the connections between contrastive learning and optimal transport.

3.1 Problem Formulation

Traditional contrastive learning methods focus on bringing positive examples, such as augmentations of the same sample, closer together in representation space. In contrast, our approach reframes contrastive learning as a distributional alignment problem, allowing flexible control over how pairs are matched by imposing specific constraints on the target transport plan, $\mathbf{P}_{tgt}$.

Our objective is to learn an encoder f_θ that minimizes the *transport cost* between positive samples. By defining $\mathbf{P}_{tgt}$ with specific alignment rules, such as domain-specific or hierarchical constraints, we can influence how samples are organized in the latent space. For instance, setting $\mathbf{P}_{tgt}$ to resemble a diagonal matrix encourages each positive to align primarily with itself or its augmentations, minimizing $\text{div}(\mathbf{I}||\mathbf{P}) \approx 0$, where div measures the deviation from an identity matrix (e.g., KL-divergence).

This flexibility allows us to encode more nuanced forms of similarity, adapting to tasks where alignment structure varies based on domain, class, or other high-level constraints. By expanding contrastive learning in this way, our method enhances separation of negatives while addressing complex relational patterns, making it suitable for a wider range of learning tasks.

Defining the Kernel Space. Before formally stating our objective, we first need to define the concept of an augmentation kernel for our positive and negative examples.

Definition 3 (Augmentation Kernel)**.** *Let f_θ denote an encoder with parameters θ and let $(\mathbf{x}'_i, \mathbf{x}''_j) \sim \mathcal{A}$ be two views drawn from the family of augmentations $\mathcal{A}$. The augmentation kernel for the encoder θ is defined as $\mathbf{K}_\theta(\mathbf{x}'_i, \mathbf{x}''_j) = \exp(-dist(\widetilde{f}_\theta(\mathbf{x}'_i), \widetilde{f}_\theta(\mathbf{x}''_j))/\varepsilon)$, where $dist(\cdot)$ can be an arbitrary distance metric, and $\widetilde{f}_\theta(\mathbf{x}'_i)$ is the normalized output of f_θ, and ε is the regularization parameter.*

Main Objective. With this definition in hand, we can now formalize our objective as follows:

$$\min_{\theta} \; d_M\big(\mathbf{P}_{\text{tgt}} || \mathbf{P}_\theta\big), \text{ with } \quad \mathbf{P}_\theta = \arg\min_{\mathbf{P} \in \mathcal{B}} \{h(\mathbf{P}) + d_\Gamma(\mathbf{P} || \mathbf{K}_\theta)\}, \tag{8}$$

where $\mathbf{K}_\theta$ is the augmentation kernel defined in Definition (3), $h(x)$ is a convex function (typically an indicator function), $\mathcal{B}$ is a closed convex constraint set (i.e. Birkhoff polytope) that defines the constraints of proximal operators, d_Γ is a Bregman divergence that is used to find the nearest points $\mathbf{P}_\theta$ on the constraint set $\mathcal{B}$ of $\mathbf{K}_\theta$, d_M is a convex function (e.g., KL-divergence) that measures divergence between $\mathbf{P}_\theta$ and the target coupling plan $\mathbf{P}_{\text{tgt}}$.

Our objective is a bi-level optimization problem which aims to learn a representation that minimizes the divergence between the transport plan $\mathbf{P}_\theta$ with the target alignment plan $\mathbf{P}_{\text{tgt}}$ that encodes the matching constraints. When we consider a standard contrastive learning setup where we have pairs of positive examples the source and target distribution, then the target $\mathbf{P}_{\text{tgt}}$ is the identity matrix $\mathbf{I}$. However, we will show later that other alignment constraints can be considered. Moreover, when $\mathcal{B}$ is the intersection of more constraint sets like $C_1^\mu \cap C_2^\nu$ in Equation (4), a nature way to get the approximation of the nearest points $\mathbf{P}_\theta$ of $\mathbf{K}_\theta$ is to run iterative projections algorithm [3], which could be extended into the intersection of several constraint sets like $\{\cap_{i=1}^n C_i\}$, resulting in a multi-marginal problem [41].

3.2 A Proximal Point Algorithm for GCA

In practice, we can solve the alignment problem above by iteratively updating the two main components in our bi-level objective. First, for a fixed encoder parameters θ, we obtain the transport coupling $\mathbf{P}_\theta$ through our corresponding proximal operator. Second, we measure the deviation between the transport plan $\mathbf{P}_\theta$ with the target $\mathbf{P}_{\text{tgt}}$ that encodes our matching constraints, which denotes the ideal alignment plan on the intersection of the constraint sets. We provide pseudocode for this iterative approach in Algorithm 1, which we refer to as generalized contrastive alignment or GCA. The implementation of our methods is in https://github.com/nerdslab/gca.

Algorithm 1 Proximal-Point Algorithm for Generalized Contrastive Alignment (GCA)

1: **Initialization:** Initial encoder parameters θ, target transport plan $\mathbf{P}_{\text{tgt}}$, kernel function $\mathbf{K}_\theta$, the function $h(x)$, divergences d_Γ and d_M (KL or W_1). Initialize transport plan $\mathbf{P}_\theta$ based on θ.

2: **Compute the transport coupling $\mathbf{P}_\theta$:** Update $\mathbf{P}_\theta$ using the proximal operator scaling for fixed θ as described in Eq. (8):

$$\mathbf{P}_\theta = \arg\min_{\mathbf{P} \in \mathcal{B}} \{h(\mathbf{P}) + d_\Gamma(\mathbf{P} || \mathbf{K}_\theta)\}.$$

3: **Calculate the loss:** Calculate deviation between the target and current transport plans

$$\mathcal{L}_{GCA} = d_M(\mathbf{P}_\theta, \mathbf{P}_{\text{tgt}}).$$

Update networks f_θ (encoder) and g_θ (projector) to minimize $\mathcal{L}_{GCA}$.

4: **Repeat until convergence:** Repeat steps 2 and 3 until convergence.

Computing the transport coupling $\mathbf{P}_\theta$ [2] (forward-pass) in GCA algorithms could be treated as a specific type of Dykstra's projection algorithms [5], which computes the **iterative projection** on the intersection of affine convex sets [3, 42]. The proofs of convergence are provided in Appendix B.1.

[2]With a single constraint set like C_1^μ in Equation (4), computing the proximal point only involves a single projection. However, if there are intersecting constraint sets like $C_1^\mu \cap C_2^\nu$, solving for the proximal point requires multiple projections before we approach the nearest point on their intersection.

3.3 GCA-UOT Method

We can also benefit from the rich literature on optimal transport to build different relaxations of our objective [43, 7, 54, 33, 35]. In particular, we choose to leverage a formulation of *unbalanced optimal transport* (UOT) to further relax the marginal constraints [11] in our objective.

In this case, we can add the dual form of d_Γ to the Equation (8) and reformulate our objective as:

$$\min_{\theta} \; d_M(\mathbf{P}_{\text{tgt}} \| \mathbf{P}_\theta) + \lambda_{\mathbf{1}} \mathbf{h}_{\mathcal{F}}(\mathbf{P}_\theta \mathbb{1} || \mu) + \lambda_{\mathbf{2}} \mathbf{h}_{\mathcal{G}}(\mathbf{P}_\theta^\top \mathbb{1} || \nu) + \varepsilon \mathbf{H}(\mathbf{P}_\theta). \tag{9}$$

Here $h_{\mathcal{F}}$ and $h_{\mathcal{G}}$ can be different divergence measures (e.g., KL divergence) that penalize deviations from the desired marginals μ and ν, and λ_1 and λ_2 are regularization parameters that control the trade-off between the transport cost and the divergence penalties. This relaxation leads to different types of proximal operators which we outline in Appendix B.2. The impact of the entropy regularization parameter ε on the coupling matrix is studied in Figure A5, along with the number of iterations and corresponding sensitivity is provided in Figure A6.

3.4 Modifying the Target Transport Plan to Encode Matching Constraints

Contrastive learning objectives can be cast as a minimization of the deviations between the transport plan $\mathbf{P}_\theta$ and the identity matrix, i.e., $\mathbf{P}_{tgt} = \mathbf{I}$. However, our GCA formulation enables learning representations that extend beyond this one-to-one matching constraint. This flexibility allows us to incorporate additional matching constraints informed by domain-specific knowledge. For example, in domain generalization scenarios [23, 28], where each batch contains samples from multiple domains, the target alignment plan can be structured as:

$$\mathbf{P}_{\text{tgt}}[i, j] = \mathbf{I}[i, j] + \alpha \cdot \mathbb{I}(D_i = D_j,\, i \neq j) + \beta \cdot \mathbb{I}(D_i \neq D_j,\, i \neq j),$$

Where $\mathbb{I}(\cdot)$ is the indicator function, which equals 1 if the condition inside is true and 0 otherwise. D_i represents the domain of sample i, where $\alpha \geq 0$ and $\beta \geq 0$. In this case, we can improve the representation by building the block constraints which encode either class information (in supervised setting) or domain information (in across domain generalization, visualized in Figure 1).

3.5 Computational Complexity

The forward-pass only involves the scaling operations in Equation (7) and doesn't affect the complexity of the backward-pass. Therefore, GCA methods can be thought of as a form of batch normalization operations with adaptive scaling. An analysis of the complexity is provided along with experiments in Appendix B.1. Our results show that GCA iterations only slightly increase the computational complexity when compared with their single step equivalent (GCA-INCE vs. INCE). However, we found that GCA-UOT is faster than INCE due to the improved symmetry and smoothness of the loss. Moreover, we record the floating point operations per second (Flops) of running GCA methods. We find that GCA-INCE (6.65 MFlops) has 5% more Flops than INCE (6.31 MFlops), while GCA-UOT saves 30% Flops (4.54 MFlops). These results show that our GCA-UOT method is not only superior in terms of accuracy but also in speed.

4 Building Connections to Different CL Objectives

In this section, we show how the modification of the different parts of our main objective $(d_\Gamma, d_M, \mathcal{B}, \mathbf{K}_\theta)$ in Equation (8) can be connected to different contrastive losses. See Table 1 for a summary of how different losses can be mapped back to our formulation.

Table 1: *Comparison of different contrastive alignment objectives.* Here we have C_1^μ and C_2^ν as constraint sets (denoted as $\mathcal{B}$) defined in Equation (4) with their corresponding indicator function. "Iter" refers to iterative methods.

Methods	d_M	d_Γ	$\mathcal{B}$	Iter
INCE	KL	KL	C_1^μ	
GCA-INCE	KL	KL	$C_1^\mu \cap C_2^\nu$	✓
RINCE (q=1)	W1	KL	C_1^μ	
GCA-RINCE (q=1)	W1	KL	$C_1^\mu \cap C_2^\nu$	✓
BYOL	KL	L2	$R^{B \times B}$	

4.1 Connection to INCE

An interesting connection that we can make between GCA main objective and contrastive learning is that we can interpret INCE as a **single step** in a iterative GCA objective [51]. This connection can be further summarized through the following theorem.

Theorem 1 (INCE Equivalence)**.** *Let $\mathbf{K}_\theta$ denote the augmentation kernel as in Definition* (3) *with cosine similarity, d_Γ and d_M equal to KL-divergence, and constraint set as C_1^μ in Equation* (4)*. The INCE objective in Equation* (1) *can be re-expressed as a GCA problem in Equation* (8) *as follows:*

$$\min_\theta KL\big(\mathbf{I}||Prox_{C_1^\mu}^{KL}(\mathbf{K}_\theta)\big). \tag{10}$$

The proof is contained in Appendix B.3. Theorem (1) shows that the INCE loss can be viewed as solving the matching problems in Equation (3) with row normalization constraints C_1^μ. This connection between GCA and INCE allows us to derive the iterative algorithm for GCA-INCE by running Bregman projection iteratively on both row and column normalization sets

4.2 Connection to RINCE

We introduce the following result to build the connection between our framework and RINCE [12].

Theorem 2 (RINCE Equivalence)**.** *Let $\mathbf{K}_\theta$ denote the augmentation kernel as in Definition* (3)*. Set target plan $\mathbf{P}_{tgt} = \mathbf{I}$, d_Γ equal to the KL-divergence, $d_M(\mathbf{I}\|\mathbf{P}) = -\frac{1}{q}(\frac{\mathrm{diag}(\mathbf{P}_\theta)}{\mathbf{u}})^q + \big(\frac{\lambda\mathbf{I}}{\mathbf{u}}\big)^q$ with λ, q, and $\mathbf{u} = \mathrm{diag}\big(\frac{\mu}{\mathbf{P}^{(0)}\mathbb{1}}\big)$, and constraint set C_1^μ defined in Equation* (4)*. The RINCE objective in Equation* (2) *can be re-expressed as a GCA problem as follows:*

$$\min_\theta d_M(\mathbf{I}\|\mathbf{P}_\theta), \;\; with \;\; \mathbf{P}_\theta = Prox_{C_1^\mu}^{KL}(\mathbf{K}_\theta), \tag{11}$$

The proof is provided in Appendix B.4.1. As we can see, RINCE introduces adjustable parameters q and λ, with λ controlling the weight of negative samples, while $q \in (0, 1]$ serves to switch between KL divergence and Wasserstein discrepancy. When $q = 1$, we have the following theorem:

Theorem 3 (W1 Equivalence)**.** *Let $\mathbf{K}_\theta$ denote the augmentation kernel as in Definition* (3) *with cosine similarity. Set target plan $\mathbf{P}_{tgt} = \mathbf{I}$, d_Γ equal to the KL-divergence, d_M equal to the 1-Wasserstein distance (W_1) in Definition* (2)*, and the constraint set as C_1^μ defined in Equation* (4)*. The RINCE object in Equation* (2) *with $q = 1$ can be re-expressed as a GCA problem as follows:*

$$\min_\theta W_1\big(\mathbf{P}_{tgt}||Prox_{C_1^\mu}^{KL}(\mathbf{K}_\theta)\big). \tag{12}$$

See Appendix B.5 for the proof.

This connection to RINCE suggests an extended iterative formulation to calculate the coupling plan as the projection point $\mathbf{P}^{(\infty)} = \mathrm{Prox}_{C_1^\mu \cap C_2^\nu}^{\mathrm{KL}}(\mathbf{K}_\theta)$ of $\mathbf{K}_\theta$ on the constraint set $C_1^\mu \cap C_2^\nu$. In this case, we can write an iterative algorithm for robust alignment called GCA-RINCE as follows:

$$L_{\text{GCA-RINCE}}^{\lambda,q} = \min_\theta -q^{-1}(\mathrm{diag}(\mathbf{P}_\theta^{(2t-1)})/\mathbf{u}^{(t)})^q + q^{-1}(\lambda\mathbf{P}_{tgt}/\mathbf{u}^{(t)})^q, \tag{13}$$

where λ and q are hyperparameters, $\mathbf{P}^{(1)} \coloneqq \mathrm{diag}(\mathbf{u}^{(1)})\mathbf{K}_\theta\,\mathrm{diag}(\mathbf{v}^{(0)})$, and t is the number of iterations.

4.3 Connection to BYOL

Our framework also allows us to make connections to BYOL [22]. BYOL learns by encouraging similarity between positive image pairs, without explicitly conditioning on negative examples. To build this connection, recall that BYOL has the online network parameterized by θ and target network parameterized by ξ, where $\mathbf{z}'_\theta = \widetilde{f}_\theta(\mathbf{x}')$ and $\mathbf{z}''_\xi = \widetilde{f}_\xi(\mathbf{x}'')$ are the normalized outputs of the online and target networks, respectively. A simplified version of the BYOL loss can be written as: $L_{\text{BYOL}} = \|\widetilde{q}_\theta(\mathbf{z}'_\theta) - \mathbf{z}''_\xi\|_2^2$, where $\widetilde{q}_\theta(\mathbf{z}'_\theta)$ is the normalized output after online network and q_θ is the predictor.[3] In this case, we can provide the following connection between GCAand BYOL as follows.

Theorem 4 (BYOL Equivalence)**.** *Let $\mathbf{S}_\theta(\mathbf{x}'_i, \mathbf{x}''_j) = \exp(-\|\widetilde{q}_\theta(\mathbf{z}'_i) - \mathbf{z}''_j\|)$ denote the augmentation kernel. Set the target plan $\mathbf{P}_{tgt} = \mathbf{I}$, d_Γ equal to the L2-distance, d_M equal to the KL-divergence, and constraint set as $R^{B\times B}$. The BYOL objective can be re-expressed as a GCA problem as follows:*

$$\min_\theta KL\big(\mathbf{I}||\mathbf{S}_\theta\big), \;\; with \;\; \mathbf{S}_\theta = Prox_{R^{B\times B}}^{\|\cdot\|}(\mathbf{S}_\theta). \tag{14}$$

See the proof in Appendix B.6.

[3] In practice, BYOL also switches the order of views to symmetrize the loss. For ease of discussion, we consider just one pair of views but the same could be argued for the full symmetric version.

5 Theoretical Analysis

In this section, we aim to show how the GCA-methods can improve alignment and uniformity in the latent space [57]. Here, *alignment* means that the features of the positive samples are as close as possible, while *uniformity* means that the features of negative samples are uniformly distributed on latent space (see Appendix C.1 for formal definitions). These quantities have been studied in a number of related works [57, 45], where one can show that improved alignment and uniformity can lead to different benefits in representation learning.

5.1 Improved alignment with GCA

Contrastive learning minimizes the deviation between the target alignment plan with the transport plan in Definition 3 through empirical risk minimization (ERM). Therefore, a tighter bound on the empirical risk corresponds to a smaller difference between the ideal alignment with the coupling matrix. We show that this in turn leads to better alignment of the positive views.

Analysis of INCE vs GCA-INCE. GCA-INCE ensures that the final transport plan $\mathbf{P}^{(\infty)}$ is closer to the ideal identity matrix compared to the INCE, as we show in the following theorem.

Theorem 5 (Improved Alignment with INCE)**.** *Let $\mathbf{K}_\theta$ denote the augmentation kernel as in Definition* (3)*. Set d_M and d_Γ to the KL-divergence, and $\mathbf{P}_{tgt} = \mathbf{I}$. The GCA-INCE loss with converged plan $\mathbf{P}_\theta^{(\infty)}$ is lower than the GCA-INCE loss with $\mathbf{P}_\theta^{(t)}$ in Equation* (6) *for all t.*

The full proof is provided in Appendix C.1.1. The above theorem tells us that solving Equation (8) with iterative projection will converge to a transport plans $\mathbf{P}_\theta^{(\infty)}$ with lower KL divergence than the one-step solution provided by INCE. We can establish the convergence of the $\mathbf{P}^{(t)} \to \mathbf{P}^{(\infty)}$, based on the convergence of Bregman projection.

Analysis of RINCE vs GCA-RINCE. GCA also benefits from other Bregman divergences, like the WDM in RINCE, which provides robustness against distribution shift compared to the KL-divergence in INCE. GCA-RINCE provides a lower bound on the RINCE loss in Equation (2), which allows us to develop a tighter bound with $\mathbf{P}^{(\infty)}$ obtained by several proximal steps with GCA.

Theorem 6 (Improved Alignment with RINCE)**.** *GCA-RINCE loss with $\mathbf{P}_\theta^{(t)}$ in Equation* (13) *is lower than the loss in the Theorem* (2) *as $L^{\lambda, q=1}_{\textit{GCA-RINCE}}(\mathbf{P}_\theta^{(t)}) \leq L^{\lambda, q=1}_{\textit{RINCE}}(\mathbf{P}_\theta^{(1)})$.*

See Appendix C.1.1 for the full proof and an analysis of GCA methods for different choices of d_M.

5.2 Improved Uniformity of Representations Through GCA

The improved alignment of GCA-methods comes from maximization of the uniformity under the constraint of intersection $C_1^\mu \cap C_2^\nu$ in Equation (4), rather than the constraint set C_1^μ in INCE (see Table 1). Finding the projection of $\mathbf{K}_\theta$ on set of $C_1^\mu \cap C_2^\nu$ through proximal steps is equivalent to solving the dual problem of EOT, which can be summarized through the following theorem.

Theorem 7 (Improved Uniformity)**.** *Given the constraint sets in Equation* (4)*, the optimal transport coupling upon convergence of Equation* (6)*, denoted as $\mathbf{P}^{(\infty)}$, achieves a higher uniformity loss compared to the single-step transport plan $\mathbf{P}^{(1)}$ obtained by INCE.*

The proof is provided in the Appendix C.2. Through loss propagation, we show that the alignment plan offered by $\mathbf{P}^{(\infty)}$ will guide the subsequent iterations towards more uniform representations.

5.3 Impacts of GCA on a downstream classification task

We take this one step further and examine the impact of GCA on a downstream classification task. For a classification task, using a labeled dataset $\mathcal{D} = \{(\bar{\mathbf{x}}_i, \mathbf{y}_i)\} \in \bar{\mathcal{X}} \times \mathcal{Y}$ where $\mathcal{Y} = [1, \ldots, M]$ with M classes, we consider a fixed, pre-trained encoder $f_\theta \in \mathcal{F} : \mathcal{X} \to \mathcal{S}$. Assume that positive and negative views of n original samples $(\bar{\mathbf{x}}_i)_{i \in [1..n]} \subset \bar{\mathcal{X}}$ are sampled from the data distribution $p(\bar{\mathbf{x}})$.

In this case, the uniformity loss is equivalent to optimizing the downstream supervised classification tasks with cross-entropy (CE) loss when the following two assumptions are satisfied [16].

Assumption 1 (Expressivity of the Encoder)**.** *Let us define $\mathcal{H}_{\bar{\mathcal{X}}}$ is the RKHS associated with the kernel $\mathbf{K}_{\bar{\mathcal{X}}}$ defined on $\bar{\mathcal{X}}$, and $(\mathcal{H}_{f_\theta}, \mathbf{K}_\theta)$ defined on $\mathcal{X}$ with augmentation kernel $\mathbf{K}_\theta = \langle f_\theta(\cdot), f_\theta(\cdot) \rangle_{\mathbb{R}^d}$ in Definition 3. And we assume that $\forall g \in \mathcal{H}_{f_\theta},\ \mathbb{E}_{\mathcal{A}(x|\cdot)} g(x) \in \mathcal{H}_{\bar{\mathcal{X}}}$.*

	Standard Setting				Noisy Setting			
Method	CIFAR-10	CIFAR-100	SVHN	ImageNet100	CIFAR-10 (Ex)	CIFAR-100 (Ex)	CIFAR-10C	CIFAR-10C (Ex)
INCE	92.01 ± 0.40	70.07 ± 0.42	90.60 ± 0.17	73.01 ± 0.61	82.03 ± 0.32	54.70 ± 0.43	87.20 ± 0.37	74.84 ± 0.21
GCA-INCE	<u>92.36 ± 0.24</u>	<u>70.11 ± 0.45</u>	90.40 ± 0.16	73.04 ± 0.76	82.18 ± 0.69	54.91 ± 0.56	87.34 ± 0.34	76.00 ± 0.17
Δ	+0.35	+0.04	-0.20	+0.03	+0.15	+0.21	+0.14	+1.16
RINCE	91.05 ± 0.50	69.06 ± 0.64	90.97 ± 0.19	71.91 ± 0.43	82.60 ± 0.63	55.43 ± 0.48	88.62 ± 1.33	77.05 ± 0.82
GCA-RINCE	92.09 ± 0.22	69.72 ± 0.27	<u>91.45 ± 0.41</u>	<u>73.44 ± 0.55</u>	<u>82.76 ± 0.49</u>	<u>55.90 ± 0.41</u>	<u>88.76 ± 0.72</u>	<u>77.23 ± 0.76</u>
Δ	+1.04	+0.66	+0.48	+1.53	+0.16	+0.47	+0.14	+0.18
SimCLR	92.16 ± 0.16	69.95 ± 0.14	90.24 ± 0.24	72.20 ± 0.78	81.87 ± 0.53	54.54 ± 0.79	86.98 ± 1.59	73.79 ± 0.32
BYOL	90.56 ± 0.59	69.75 ± 0.37	89.50 ± 0.46	69.75 ± 0.83	81.55 ± 0.50	54.18 ± 0.46	87.88 ± 1.02	69.40 ± 1.11
IOT [51]	90.99 ± 0.54	67.19 ± 0.21	90.15 ± 0.21	72.27 ± 0.53	80.59 ± 0.64	52.40 ± 0.48	67.36 ± 1.97	58.75 ± 1.96
IOT-uni [51]	90.89 ± 0.57	67.03 ± 0.40	90.54 ± 0.20	72.88 ± 0.71	80.79 ± 0.24	53.04 ± 0.52	69.58 ± 1.25	59.05 ± 1.86
GCA-UOT	**92.61 ± 0.32**	**71.45 ± 0.37**	**91.96 ± 0.15**	**74.09 ± 0.40**	**83.18 ± 0.44**	**56.30 ± 0.51**	**89.61 ± 0.30**	**77.60 ± 0.54**

Table 2: *Test accuracy (%) on a downstream classification task after pretraining.* Results are provided for CIFAR-10 (ResNet18), CIFAR-100 (ResNet18), SVHN (ResNet50), and ImageNet100 (ResNet50) under standard and extreme (Ex) augmentation conditions (averaged over 5 seeds). The top model is bold and the second-place model is underlined. For INCE and RINCE, we also provide the improvement Δ by adding GCA to each method.

Assumption 2 (Small Intra-Class Variance)**.** *For $y \neq y'$, the intra-class variance δ_i, δ_j are negligible compared to the distance among different class centroids, $\mu_y, \mu_{y'}$ as $\|\mu_y - \mu_{y'}\| \gg \|\delta_i - \delta_j\|$.*

Claim 1. *If Assumption 1 and Assumption 2 hold, then maximizing the uniformity is equivalent to minimizing the downstream CE loss.*

The proof is provided in Appendix C.2. Optimizing the self-supervised loss under ideal conditions improves downstream CE tasks and helps to explain why maximizing uniformity aids classification.

Remark.. Maximizing uniformity can enhance downstream classification but risks "feature suppression" by encouraging shortcut features that harm generalization [48]. In GCA-UOT, adding penalties modifies the transport plan from that of a pure uniformity loss, helping to avoid feature suppression. We find empirical evidence that UOT provides a more robust transport plan which appears to circumvent some of these shortcut features from being learned (Figure A4 in Appendix C.3).

6 Experiments

In this section, we conduct empirical evaluations to study the performance of our approach in both handling noisy and corrupted views and in domain generalization tasks.

6.1 Comparison with CL Baselines

Experiment setup. To examine the robustness of our framework, we trained INCE and RINCE as baselines, and developed their GCA-based alternatives (+GCA). In addition, we also compared with our novel GCA-UOT method, two variants of IOT established in [51], and other CL baselines, including BYOL and SimCLR. For experiments with SVHN [36] and ImageNet100 [15] we use the ResNet-50 encoder as the backbone and use a ResNet-18 encoder as the backbone for CIFAR-10, CIFAR-100 [29] and a corrupted version of CIFAR called CIFAR-10C [25].

In all of these cases, we follow the standard self-supervised learning evaluation protocol [8], where we train the encoder on the training set in an unsupervised manner and then train a linear layer on top of the frozen representations to obtain the final accuracy on the test set. In addition to standard data augmentation policies commonly used [12], we also apply three different extreme augmentation policies to examine the robustness of GCA towards noisy views (details in Appendix D.2). Learning rates and other training details for CIFAR-10, CIFAR-100, SVHN, and ImageNet100 are provided in Appendix D.1, while specific training details for CIFAR-10C are included in Appendix D.2.

Results on Standard Augmentations. First, we performed experiments on CIFAR-10, CIFAR-100, SVHN, and ImageNet100 using standard sets of augmentations that are applied to achieve state-of-the-art performance (Table 2, Standard Setting). We found the +GCA versions of INCE and RINCE exhibit performance gains in almost all settings except for SVHN, with bigger gains observed when adding GCA to RINCE. Additionally, we find that our unbalanced OT method, GCA-UOT, achieves the top performance across the board, on all four datasets tested. The transport plans obtained by each methods are provided in Figure A4 along with a study of the sensitivity of the methods to hyperparameters (Appendix A7).

Results on Corrupted Data and Extreme Augmentations. Next, we tested the methods in two noisy settings. In the first set of experiments, we apply extreme augmentations to CIFAR-10 (Ex) and

CIFAR-100 (Ex) (see Appendix D.2) to introduce noisy views during training. In the second set of experiments, we used the CIFAR-10C to further test the ability of our method to work in noisy settings. Our experimental results demonstrate that the GCA-based strategy effectively enhances the model's generalization ability and adaptability to aggressive data augmentations. In addition to improving classification accuracy, the GCA-based methods also improve the representational alignment and uniformity, as shown in Appendix E.2. This observation is in line with our theoretical analysis in Section 5.2, where we show that the obtained representations provide better overall alignment of positive views and better spread in terms of uniformity [57].

6.2 Block Diagonal Transport in Domain Generalization

In a final experiment, we aimed to demonstrate the flexibility and robustness of our framework by applying it to a domain generalization task, where samples originate from different domains (e.g., Photo, Cartoon, Sketch, Art). We explored the effects of introducing domain-specific alignment constraints in our transport plan, hypothesizing that this could enhance the latent space organization to capture more nuanced domain similarities.

Our approach enables additional contextual information to be seamlessly integrated into the transport process. In this case, domain information was incorporated to distinguish the alignment of samples from the same versus different domains. To achieve this, we adjusted the target transport plan $\mathbf{P}_{tgt}$, selectively modifying parameters (α, β) to vary the influence of domain-based alignment constraints as shown in Figure 1(A). Specifically, we set $\{\alpha = 0, \beta > 0\}$ to prioritize cross-domain alignment and $\{\alpha > 0, \beta = 0\}$ to focus on intra-domain alignment.

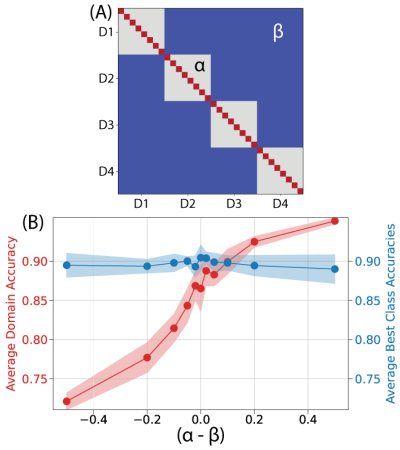

Figure 1: *Incorporating different priors into learning across multiple domains.* (A) Example target alignment plan $\mathbf{P}_{tgt}$, where the target over all samples from the same domain are set to α, the diagonal values are set to 1, and across-domain samples are set to β. (B) The domain classification accuracy (red) and overall class accuracy (blue) with $(\alpha - \beta)$ increases.

The training was conducted on the PACS dataset [31] using a ResNet-18 encoder with the GCA-INCE objective. After training the encoder in an unsupervised manner, we freeze the encoder and then train a linear readout layer to predict either the sample's class or the domain it belonged to. This setup allowed us to isolate the effect of our transport adjustments on the latent space's capacity to encode both class and domain information.

The results, displayed in Figure 1(B), revealed that increasing the domain alignment weight enhances the accuracy of domain classification (from 72.11% to 95.16%) without diminishing classification performance. This outcome suggests that GCA can effectively encode both domain and class information in a single latent representation. The ability to adjust alignment constraints provides a powerful tool for domain generalization tasks, enabling multiple types of similarity to be jointly encoded. This flexibility can potentially alleviate issues related to information loss from data augmentation, especially in fine-grained classification settings, by retaining essential domain-specific characteristics across transformations.

7 Conclusion

In this work, we introduced *generalized contrastive alignment* (GCA), a flexible framework that redefines contrastive learning as a distributional alignment problem using optimal transport to control alignment. By allowing targeted control over alignment objectives, GCA demonstrates strong performance across both standard and challenging settings, such as noisy views and domain generalization tasks. This work opens up broader possibilities for learning robust representations in real-world scenarios, where data is often diverse, noisy, or comes from multiple domains.

Future work includes applications of GCA to graphs and time series data, as well as multi-modal settings where our approach can integrate various forms of similarity. As alignment strategies become integral to contrastive learning, GCA offers a promising foundation for more adaptive and expressive self-supervised models.

Acknowledgements

We would like to thank Mehdi Azabou, Divyansha, Vinam Arora, Shivashriganesh Mahato, and Ian Knight for their valuable feedback on the work. This work was funded through NSF IIS-2212182, NSF IIS-2039741, and the support from the Canadian Institute for Advanced Research (CIFAR). We would also like to acknowledge the use of ChatGPT for providing useful feedback and suggestions on the writing of the paper.

References

[1] Dongsheng An, Na Lei, Xiaoyin Xu, and Xianfeng Gu. Efficient optimal transport algorithm by accelerated gradient descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10119–10128, 2022.

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[3] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[4] Robert J Berman. The sinkhorn algorithm, parabolic optimal transport and geometric monge–ampère equations. *Numerische Mathematik*, 145(4):771–836, 2020.

[5] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[6] Xing-Ju Cai, Ke Guo, Fan Jiang, Kai Wang, Zhong-Ming Wu, and De-Ren Han. The developments of proximal point algorithms. *Journal of the Operations Research Society of China*, 10(2):197–239, 2022.

[7] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16296–16305, 2021.

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

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[10] Zining Chen, Weiqiu Wang, Zhicheng Zhao, Fei Su, Aidong Men, and Yuan Dong. Instance paradigm contrastive learning for domain generalization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(2):1032–1042, 2023.

[11] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

[12] Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. Robust contrastive learning against noisy views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16670–16681, 2022.

[13] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

[14] Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14*, pages 274–289. Springer, 2014.

[15] 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*, pages 248–255. Ieee, 2009.

[16] Benoit Dufumier, Carlo Alberto Barbano, Robin Louiset, Edouard Duchesnay, and Pietro Gori. Integrating prior knowledge in contrastive learning with kernel. In *International Conference on Machine Learning*, pages 8851–8878. PMLR, 2023.

[17] Marvin Eisenberger, Aysim Toker, Laura Leal-Taixé, Florian Bernard, and Daniel Cremers. A unified framework for implicit sinkhorn differentiation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 509–518, 2022.

[18] Zhenghan Fang, Sam Buchanan, and Jeremias Sulam. What's in a prior? learned proximal networks for inverse problems. *arXiv preprint arXiv:2310.14344*, 2023.

[19] Promit Ghosal and Marcel Nutz. On the convergence rate of sinkhorn's algorithm. *arXiv preprint arXiv:2212.06000*, 2022.

[20] Aritra Ghosh, Naresh Manwani, and PS Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107, 2015.

[21] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

[22] 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.

[23] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

[24] 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*, pages 9729–9738, 2020.

[25] Dan Hendrycks and Thomas G Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

[26] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

[27] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.

[28] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9619–9628, 2021.

[29] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[30] John Lee, Max Dabagia, Eva Dyer, and Christopher Rozell. Hierarchical optimal transport for multimodal distribution alignment. *Advances in neural information processing systems*, 32, 2019.

[31] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[32] Ruilin Li, Xiaojing Ye, Haomin Zhou, and Hongyuan Zha. Learning to match via inverse optimal transport. *Journal of machine learning research*, 20(80):1–37, 2019.

[33] Chi-Heng Lin, Mehdi Azabou, and Eva L Dyer. Making transport more robust and interpretable by moving data through a small number of anchor points. *Proceedings of machine learning research*, 139:6631, 2021.

[34] Ran Liu, Mehdi Azabou, Max Dabagia, Chi-Heng Lin, Mohammad Gheshlaghi Azar, Keith Hengen, Michal Valko, and Eva Dyer. Drop, swap, and generate: A self-supervised approach for generating neural activity. *Advances in neural information processing systems*, 34:10587–10599, 2021.

[35] Eduardo Fernandes Montesuma, Fred Ngole Mboula, and Antoine Souloumiac. Recent advances in optimal transport for machine learning. *arXiv preprint arXiv:2306.16156*, 2023.

[36] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, volume 2011, 2011.

[37] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

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

[39] Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron Van den Oord, Sergey Levine, and Pierre Sermanet. Wasserstein dependency measure for representation learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[40] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014.

[41] Brendan Pass. Multi-marginal optimal transport: theory and applications. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1771–1790, 2015.

[42] Gabriel Peyré. Entropic approximation of wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

[43] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[44] Khiem Pham, Khang Le, Nhat Ho, Tung Pham, and Hung Bui. On unbalanced optimal transport: An analysis of sinkhorn algorithm. In *International Conference on Machine Learning*, pages 7673–7682. PMLR, 2020.

[45] Shi Pu, Kaili Zhao, and Mao Zheng. Alignment-uniformity aware representation learning for zero-shot video classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19968–19977, 2022.

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[47] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

[48] Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986, 2021.

[49] Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative modeling with optimal transport maps. *arXiv preprint arXiv:2110.02999*, 2021.

[50] Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pages 19250–19286. PMLR, 2022.

[51] Liangliang Shi, Gu Zhang, Haoyu Zhen, Jintao Fan, and Junchi Yan. Understanding and generalizing contrastive learning from the inverse optimal transport perspective. In *International Conference on Machine Learning*, pages 31408–31421. PMLR, 2023.

[52] Saeed Shurrab and Rehab Duwairi. Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science*, 8:e1045, 2022.

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

[54] Fariborz Taherkhani, Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, and Nasser M Nasrabadi. Self-supervised wasserstein pseudo-labeling for semi-supervised image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12267–12277, 2021.

[55] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

[56] Ankit Vishnubhotla, Charlotte Loh, Akash Srivastava, Liam Paninski, and Cole Hurwitz. Towards robust and generalizable representations of extracellular data using contrastive learning. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[58] Yule Wang, Chengrui Li, Weihan Li, and Anqi Wu. Exploring behavior-relevant and disentangled neural dynamics with generative diffusion models. *arXiv preprint arXiv:2410.09614*, 2024.

[59] Yule Wang, Zijing Wu, Chengrui Li, and Anqi Wu. Extraction and recovery of spatio-temporal structure in latent dynamics alignment with diffusion model. *Advances in Neural Information Processing Systems*, 36, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: This paper is a theoretical paper which discusses the a generalized framework for contrastive learning, which involves to convert them into a series of proximal algorithms. The abstract and instruction illustrate this point properly.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Contrastive learning is still used in an ad hoc manner, with augmentations often causing harmful effects. This in turn can introduce bias or hallucinations which can negative societal impacts. We will expand on these topics in a future revision. Since we disscuss the limitations and will introduce future works in the conclusion, the answer should be yes.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Please check the theoretical part in the main text, which is mainly contained in section 3, 4 and 5. Their proofs are provided in the Appendix, where each theory has its corresponding proof.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please check Section 6 and Appendix for experimental details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will release all codes after review.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please check Section 6 and Appendix for details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Please check Section 6 and Appendix for details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide GPU information in Appendix D along with experimental details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This is a theoretical work and fullfills the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: By providing more provable and robust methods for representation learning, this work can have impact in fairness and help to reduce uncertainty in decision making.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This theoretical paper has no risk that we are aware of.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Please check references.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not introduce new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not include crowdsourcing or human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not include human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.