OT-CLIP: Understanding and Generalizing CLIP via Optimal Transport

Liangliang Shi [*1] **Jack Fan** [*2] **Junchi Yan** [1]

Abstract

We propose to understand Contrastive Language-Image Pretraining model (CLIP) from the Optimal Transport (OT) perspective. Specifically, we show that training of CLIP is an embodiment of inverse OT and the adopted two InfoNCE losses in CLIP correspond to a special case of bilevel optimization of a modified entropic OT. We then generalize the original CLIP loss to an OT-based loss family using variants of Regularized OT (e.g. Fused Gromov OT, unbalanced OT, etc.), and show their superior performance on public datasets for downstream tasks in both image and text domain. We also rethink the inference stage of CLIP by using the tool of OT, and propose to adopt the fused Gromov OT for (zero-shot) classification, in which the prediction is based on the graph representation whereby images and texts are nodes for graph matching. By our new technique, we show how to generalize zero-shot classification to other more flexible zero-shot tasks with competitive performance: long-tailed classification and selective classification. The former assumes the known prior distribution of labels, while in the latter case, only a subset of samples are asked to predict, yet with the need of high prediction confidence. The code is available at https://github.com/fan23j/ICML2024-OT-CLIP.

1. Introduction

Using weakly supervised image-text pairs for contrastive pretraining is becoming the preferred method for acquiring a general-purpose computer vision backbone, gradually replacing pretraining on large annotated multiclass datasets. The core idea is to learn aligned representation spaces for both images and text using paired data. Groundbreaking works such as CLIP (Radford et al., 2021b) and ALIGN (Jia et al., 2021) have demonstrated the feasibility of this approach at a large scale. The standard approach to pretraining such models involves using image-text contrastive objectives, which align image and text embeddings to match (positive) image-text pairs while ensuring that unrelated (negative) image-text pairs are dissimilar in the embedding space. This is achieved through the contrastive InfoNCE loss (Oord et al., 2018), which is applied twice to normalize the pairwise similarity scores for all images and then for all texts. However, the batch-level softmax-based InfoNCE loss seems simple and the underlying alignment mechanism seems to not have been explored.

In this paper, the optimal transport (OT) perspective is applied for the contrastive vision-language pertaining models for both learning the representation and zero-shot classification inference. In the training process, the alignment between the image-text pair can be reformulated with inverse optimal transport (IOT) (Stuart & Wolfram, 2020) that learns the cost matrix with the bi-level optimization given the known coupling supervision. Note the cost matrix depends on the distance between image and text features and thus the optimization of IOT is exactly equivalent to learning the representations. In the inference process, zero-shot classification can also be viewed with the optimization of OT: matching the image features to the corresponding labels. Both InfoNCE loss and softmax inference can be a special case under the modification of the constraints of OT.

From this perspective, we can naturally introduce OT technologies for both the learning and inference processes for CLIP. Since softmax can be understood from the perspective of entropic regularized OT with half-constraints as discussed in (Shi et al., 2023), an OT-CLIP loss family is proposed for learning representations by modifying the constraints and utilizing variants of regularized OT. Specifically, we replace the traditional loss function with a series of OT-based losses, such as regularized OT (Dessein et al., 2018) with full constraints, unbalanced OT (Liero et al., 2018), fused Gromov OT (Vayer et al., 2018), and Double-Bounded OT (DBOT) (Shi et al., 2024a), to enhance the training process.

Then, for zero-shot inference, considering the structural information between the representations, we incorporate the

[*]Equal contribution [1]School of Artificial Intelligence & Department of Computer Science and Engineering & MoE Lab of AI, Shanghai Jiao Tong University, Shanghai, China [2]Department of Computer Science, University of North Carolina at Chapel Hill. Correspondence to: Junchi Yan <yanjunchi@sjtu.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

idea of graph matching for classical classification, where both node matching and edge matching are taken into account. We use a modified algorithm for fused Gromov OT (specifically, the half-constraint coupling case) to enhance the prediction. Additionally, with our new technique, we demonstrate how to extend zero-shot classification tasks to other zero-shot tasks with competitive performance, such as long-tailed classification and selective classification. In the former case, we assume a known label distribution for the test data, allowing direct application of OT variants for prediction inference. In the latter case, we employ modified partial OT (Chapel et al., 2020) and unbalanced OT for selective zero-shot classification, leveraging CLIP to avoid incorrect or highly uncertain predictions. In summary, we make the following contributions:

1) We propose to understand CLIP with Optimal Transport in both learning and inference. The learning process can be viewed as Inverse OT, involving a bi-level optimization procedure, while zero-shot inference can be interpreted as optimizing the OT problem with half-constraints of coupling i.e. only the row-wise constraint is enforced.

2) We introduce the OT-CLIP loss family, which replaces the traditional loss in InfoNCE with variants of entropic OT. We show how to enrich the CLIP loss family by employing entropic OT with full constraints, unbalanced OT, fused Gromov OT, Double-Bounded OT, and combinations of these variants. We also provide empirical evidence that fused Gromov OT achieves the best performance in zero-shot classification and learning the text encoder, effectively capturing the relational structure among image/text samples.

3) In CLIP inference, we introduce graph structure and perform graph matching for classification. This is in contrast to the traditional softmax prediction that can be viewed as point matching without considering the edge information i.e. graph structure. We utilize a modified algorithm for fused Gromov OT in the prediction process. Furthermore, we extend our OT methodology to two additional (zero-shot) classification inference settings: long-tailed classification and selective classification, achieving promising performance in both cases.

2. Related Works and Preliminaries

Contrastive Learning and InfoNCE. In recent years, self-supervised contrastive learning has garnered significant attention and has been extensively explored. Specifically, SimCLR (Chen et al., 2020) is one of the influential works, which introduces a simple yet effective framework that maximizes agreement between different views of the same data sample while minimizing agreement between views of different samples. Besides, to reduce the computation space, MoCo (He et al., 2020) introduces a momentum-based update strategy to build a dynamic dictionary of negative samples, while BYOL (Grill et al., 2020) achieves competitive performance and demonstrates the potential of contrastive learning without explicit negative pairs. These methods have achieved impressive performance in downstream tasks such as classification, object detection, and segmentation with fine-tuning on the target datasets. InfoNCE, as the most common loss function in contrastive learning, has been the subject of numerous studies aimed at improving and understanding its properties. Previous works (Oord et al., 2018; Zbontar et al., 2021; Tian et al., 2020a) primarily interpret InfoNCE as maximizing the lower bound of mutual information between different levels of features. However, some works, such as (Tschannen et al., 2019), disagree with this lower-bound interpretation. In a recent study (Shi et al., 2023), InfoNCE was understood from the perspective of optimal transport, where the softmax form can be linked to entropic regularization, and the regularization coefficient corresponds exactly to the temperature. Inspired by this idea, we propose to leverage OT variants to enhance InfoNCE in the context of contrastive vision-language pretraining.

Vision-Language Pretrained Models (VLPMs). There has been significant progress in the field of vision-language understanding, particularly in the areas of vision-language pre-training (VLPMs) (Zhong et al., 2022). Given paired image and text data, these approaches allow neural networks to learn multi-modal representations, enabling the alignment of visual concepts with natural language. One notable approach in VLPMs is the Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021b) framework, which employs both image and text encoders to generate corresponding features that can be aligned in a cross-modality representational space for image-text matching. By training on 400 million image-text pairs, CLIP achieves impressive generalizability and usability, showing competitive performance in aligning image embeddings to broad categories, which is widely applied in downstream tasks both for pre-training and fine-tuning stages (e.g. zero/few-shot classification). To enhance the capabilities of CLIP, SLIP introduces self-supervision to the framework, as detailed by Mu et al. (Mu et al., 2021). Besides, DeCLIP extends the approach by incorporating extensive supervision across image-text pairs, leveraging multi-view data. However, there are still some limitations that need to be addressed. Firstly, the performance of the extracted representations from the image and text encoders falls short of what many unsupervised methods (such as MoCo v3) achieve in downstream tasks like object detection and segmentation. Secondly, training VLPMs requires paired image-text data, and there is limited research on dealing with semi-supervised VLPMs. In this paper, we aim to tackle these two challenges and enhance the performance of our pretrained model in both zero-shot classifications and other downstream tasks that

involve fine-tuning.

Optimal Transport (OT). Known for calculating the distance between (probability) measures, optimal transport (OT) (Kantorovich, 1942), has gained significant attention in various fields due to its powerful mathematical framework for solving transportation (Li et al., 2023a) and matching problems (Wang et al., 2013; 2023). Recently, several related works have focused on exploring and extending the applications of OT such as image registration (Feydy et al., 2017; **?**), barycenter learning (Zhu et al., 2020), style transfer (Kolkin et al., 2019), domain adaptation (Damodaran et al., 2018; Chang et al., 2022) and generative models (Gulrajani et al., 2017; Li et al., 2022b; 2023b). Recently, (Shi et al., 2023) proposed a new Inverse entropic OT-based perspective with a bi-level optimization:

$$\begin{aligned} &\min_{\theta} KL(\tilde{\mathbf{P}}|\mathbf{P}^{\theta}) \\ \text{where} \quad &\mathbf{P}^{\theta} = \arg\min_{\mathbf{P}\in U(\mathbf{a},\mathbf{b})} < \mathbf{C}^{\theta}, \mathbf{P} > -\epsilon H(\mathbf{P}). \end{aligned} \tag{1}$$

Here $\tilde{\mathbf{P}}$ is a known supervision for the transportation, $\mathbf{C}^{\theta}$ is the neural-based cost matrix with parameters θ, $H(\mathbf{P})$ is the entropic regularization with coefficient ϵ. $U(\mathbf{a}, \mathbf{b})$ is the marginal constraint set for coupling $\mathbf{P}$ with marginals $\mathbf{a}$, $\mathbf{b}$:

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

In particular, the work (Shi et al., 2023) shows the equivalence of loss between InfoNCE and the proposed optimization under a simplified constraint set, and the temperature equals the regularization coefficient with this perspective. However, (Shi et al., 2023) focuses on modifying the row and column constraints represented by InfoNCE and the theory can only be applied in learning the representations. Following this study, our work explores different OT optimization objectives (e.g., Unbalanced OT, fused Gromov OT) to obtain more generalized CLIP losses and apply the OT theory in downstream zero-shot classification tasks.

3. OT-CLIP for Representation Learning

3.1. Re-Understanding CLIP via Inverse OT

InfoNCE Losses for CLIP. CLIP is a multi-modal model that jointly learns representations of images and texts. By using the image-text supervision, CLIP trains the image and text encoders via InfoNCE losses. Specifically, in a batch of N image-text pairs $\{(x_i^I, x_i^T)\}_{i=1}^N$, we define x_i^I and x_i^T as the image and text samples of the $i-$th pair. Let $z_i^I = f_\theta(x_i)$ and $z_i^T = f_\phi(x_i^T)$ be the embedding of x_i^I and x_i^T with image encoder $f_\theta(\cdot)$ and text encoder $f_\psi(\cdot)$, then the InfoNCE loss for the image encoder can be denoted as:

$$\mathcal{L}_I = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(z_i^I, z_i^T)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(z_j^I, z_j^T)/\tau)}. \tag{3}$$

where τ is the temperature variable to scale the logits and $\text{sim}(\cdot,\cdot)$ is the cosine similarity between vectors. Symmetrically, we can get the InfoNCE loss $\mathcal{L}_T$ for text encoder similarly and the final loss for CLIP is denoted as

$$\min_{\mathbf{\Theta}} \mathcal{L}_{CLIP} = \frac{1}{2}(\mathcal{L}_I + \mathcal{L}_T). \tag{4}$$

where $\mathbf{\Theta} = (\theta, \psi)$ are the parameters of CLIP. Then we show the loss above equals the bi-level optimization by Inverse OT, which can help us gain a new understanding of the loss of CLIP and facilitate generalization.

Equivalent formulation: OT-based Bilevel Optimization. Typically, the InfoNCE losses for CLIP are understood by aligning the image and text embeddings to match (positive) image-text pairs while ensuring that unrelated (negative) image-text pairs are dissimilar in the embedding space. In this paper, we give a new understanding via the formulation of OT. Specifically, we first define the ground truth $\tilde{\mathbf{P}}$ as supervision (i.e. $\tilde{\mathbf{P}}_{ij} = 1$ for $i = j$ and $\tilde{\mathbf{P}}_{ij} = 0$ for $i \neq j$). Then with the embedding of z_i^I and z_j^T, we can define the cost $\mathbf{C}_{ij}^{\Theta}$ with cosine distance between z_i^I and z_i^T (i.e. $\mathbf{C}_{ij}^{\Theta} = 1 - \text{sim}(z_i^I, z_i^T)$) and the CLIP loss is equal to the minimization of Inverse OT with a bilevel optimization

$$\min_{\mathbf{\Theta}} KL(\tilde{\mathbf{P}}|\mathbf{P}_I^{\theta}) + KL(\tilde{\mathbf{P}}|\mathbf{P}_T^{\theta}), \tag{5}$$

where the inner minimization is

$$\begin{aligned} \mathbf{P}_I^{\mathbf{\Theta}} &= \arg\min_{\mathbf{P}\in\mathcal{C}(\mathbf{a})} < \mathbf{C}^{\Theta}, \mathbf{P} > -\epsilon H(\mathbf{P}), \\ \mathbf{P}_T^{\mathbf{\Theta}} &= \arg\min_{\mathbf{P}\in\mathcal{C}'(\mathbf{b})} < \mathbf{C}^{\Theta}, \mathbf{P} > -\epsilon H(\mathbf{P}). \end{aligned} \tag{6}$$

Note that $\mathcal{C}(\mathbf{a}) = \{\mathbf{P}\mathbf{1} = \mathbf{a}\}$ and $\mathcal{C}'(\mathbf{b}) = \{\mathbf{P}^{\top}\mathbf{1} = \mathbf{b}\}$ are the constraint set of $\mathbf{P}$ with vector $\mathbf{a} = \mathbf{b} = \mathbf{1}$. One can find the equivalence between the optimizations of Eq. 4 and Eq. 5, and the proof can be found in (Shi et al., 2023). Modeling alignment using the mathematical formulation of OT provides us with further insights into the CLIP loss. It reveals that softmax is essentially a transformation based on OT entropy regularization, where the temperature parameter is equivalent to the coefficient of entropic regularization. By using InfoNCE twice, we can interpret it as corresponding to the two constraints in the inner optimizations, namely the row normalization of images and the column normalization of texts. This equivalence inspires us to consider and improve the CLIP InfoNCE loss from the perspective of the mathematical form of OT. In the next subsection, we will introduce a new OT-CLIP loss family by incorporating variants of entropic OT.

3.2. Generalization of CLIP loss via OT variants

Based on the observation that InfoNCE can be seen as a specific instance of IOT, we propose adopting OT variants to generalize the CLIP loss, as shown in Figure 1. We first

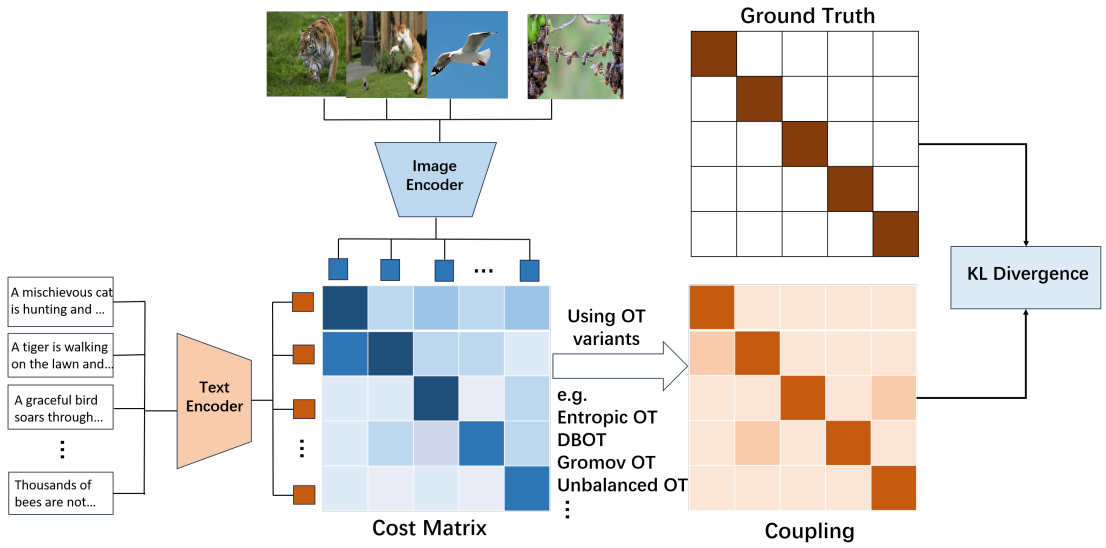

Figure 1. Overview of the OT-CLIP for representation learning. Based on the OT framework, a family of losses for the CLIP model can be derived by adopting different variants of OT. These two worlds to our best knowledge, have not been bridged before.

replace the existing row/column normalization constraints with new ones to obtain the following novel losses.

Improving the CLIP loss by varying the constraints. The previous subsection demonstrates the equivalence between IOT optimization and the CLIP InfoNCE loss, which inspires us to consider and enhance the CLIP loss from the perspective of OT. To improve the representations, we first consider replacing the row/column and normalization constraints in InfoNCE with new constraints:

$$\min_{\Theta} KL(\tilde{\mathbf{P}}|\mathbf{P}^{\Theta}) \text{ s.t. } \mathbf{P}^{\Theta} = \arg\min_{\mathbf{P}\in\mathcal{C}} < \mathbf{C}^{\Theta}, \mathbf{P} > -\epsilon H(\mathbf{P}), \tag{7}$$

where $\mathcal{C}$ is the constraint set for the coupling $\mathbf{P}$. By modifying the constraint $\mathcal{C}$, we can derive a series of CLIP losses using Eq. 7. It is evident that by setting $\mathcal{C} = \mathcal{C}(\mathbf{a})$ and $\mathcal{C}'(\mathbf{b})$ sequentially, we can obtain CLIP losses based on InfoNCE. One direct improvement is to restore the constraints from vanilla Optimal Transport (OT), i.e., setting $\mathcal{C} = U(\mathbf{a}, \mathbf{b})$. Thus, we can use the Sinkhorn algorithm to obtain $\mathbf{P}^{\Theta}$ and subsequently utilize the KL divergence or cross-entropy to derive a new CLIP loss. Another option for $\mathcal{C}$ is to adopt the constraint proposed by DBOT (Shi et al., 2024a):

$$\{\mathbf{P} > \mathbf{0} | \mathbf{P}\mathbf{1} = \mathbf{a}, \mathbf{b} - \delta \leq \mathbf{P}^{\top}\mathbf{1} \leq \mathbf{b} + \delta\}, \tag{8}$$

where the constraints lie between $\mathcal{C}(\mathbf{a})$ and $U(\mathbf{a}, \mathbf{b})$. These constraints do not directly impose restrictions on the columns like $\mathcal{C}(\mathbf{a})$ does or specify column sums like $U(\mathbf{a}, \mathbf{b})$ does. Instead, they confine the column sums im a certain range, providing tolerance for the output coupling.

A more generalized form: Fused Gromov Unbalanced OT. We consider modeling the inner optimization in a more general form. Here, following (Xu & Cheng, 2023), we adopt the formulation of Fused Gromov Unbalanced OT with entropic regularization:

$$\begin{aligned}\min_{\mathbf{P}\geq\mathbf{0}} < \mathbf{C}^{\theta}, \mathbf{P} > -\epsilon H(\mathbf{P}) + \alpha_0 < C(\mathbf{P}, \mathbf{D}_1, \mathbf{D}_2), \mathbf{P} > \\ + \alpha_1 KL(\mathbf{P}\mathbf{1}|\mathbf{a}) + \alpha_2 KL(\mathbf{P}^{\top}\mathbf{1}|\mathbf{b}).\end{aligned} \tag{9}$$

In Eq. 9, we then introduce the regularizers:

i) Marginal Prior Regularization: Instead of imposing hard constraints, we utilize two KL divergence terms in Eq. 9 to penalize the difference between the marginals of $\mathbf{P}$ and the predefined prior distributions. The KL divergence is represented as $KL(\mathbf{a}|\mathbf{b}) = \langle \mathbf{a}, \log \mathbf{a} - \log \mathbf{b} \rangle - \langle \mathbf{a} - \mathbf{b}, \mathbf{1} \rangle$, where $\mathbf{a}$ and $\mathbf{b}$ are vectors, and the two KL regularized terms are controlled by α_1 and α_2, respectively. When $\alpha_1 \to +\infty$ and $\alpha_2 \to +\infty$, the Marginal Prior Regularization degenerates into hard marginal constraints in $U(\mathbf{a}, \mathbf{b})$. The strength of using Marginal Prior Regularization lies in the fact that real data contains noise, and the supervised data for one-to-one matching may not be strict, allowing for one-to-many or many-to-one cases.

ii) Gromov-Wasserstein Discrepancy-Based Structural Regularization: In addition to considering the alignment between image-text pairs, we also take into account the structural matching between the image features and text features within a batch. We define $\mathbf{D}_1$ as the cosine distance matrix for the image features and $\mathbf{D}_2$ as the cosine distance matrix for the text features. Next, we construct a structural cost:

$$C(\mathbf{P}, \mathbf{D}_1, \mathbf{D}_2) = -\mathbf{D}_1\mathbf{P}\mathbf{D}_2^{\top}. \tag{10}$$

The significance of this cost term, controlled by α_0, can be captured by $< C(\mathbf{P}, \mathbf{D}_1, \mathbf{D}_2), \mathbf{P} >= -tr(\mathbf{P}^{\top}\mathbf{D}_1\mathbf{P}\mathbf{D}_2^{\top})$,

aiming to achieve the edge matching between image features and text features. As depicted in Eq. 9, combining the original cost term of OT with the structural regularizer leads to the well-known fused Gromov-Wasserstein (FGW) discrepancy. This discrepancy serves as an optimal transport-based metric for structured data, such as graphs.

Note when $\alpha_0 > 0$ and $\alpha_1 = \alpha_2 \to +\infty$, the optimization in Eq. 9 simplifies to Fused Gromov OT, which we refer to as "**FGromov OT**" in our experiments. When $\alpha_0 = 0$ and $0 < \alpha_1, \alpha_2 < +\infty$, the optimization in Eq. 9 simplifies to (entropic) unbalanced OT, which we refer to as "**Unbal. OT**". Lastly, when $\alpha_0 > 0$ and $\alpha_1, \alpha_2 < +\infty$, the optimization problem represents the general case, which we refer to as "**U-FG OT**" in our experiments.

More Variants for More OT-CLIP loss. It can be observed that each type of regularized OT can be used to derive a new OT-CLIP loss. Therefore, this paper effectively proposes a family of OT-CLIP losses through IOT bilevel optimization. Other OT variants (Shi et al., 2024b), such as those utilizing L2 (Essid & Solomon, 2018) or Tsallis entropic (Muzellec et al., 2017) regularization instead of entropic regularization and adopting the semi-relaxed optimal transport instead of hard constraints and unbalanced case, can also be used to derive new loss functions. Besides, adopting other divergence e.g. JS instead of KL can also be derived for a series of new CLIP losses. Due to space limitations, this paper only derives the aforementioned OT variants as specific examples.

3.3. Experiments on CLIP Training

Datasets. Our models are pretrained on the popular Conceptual Captions 3M (CC3M) (Sharma et al., 2018) image-text pairs and primarily evaluated on ImageNet1K (Deng et al., 2009) zero-shot classification.

Training and Architecture. For each experiment, we train for 30 epochs with a batch size of 256 across 4 32GB V100 GPUs for an effective batch size of 1024. We use learning rate *lr = 5e-4* with the AdamW optimizer and weight decay of 0.1 in all our experiments. Besides, following CLIP (Radford et al., 2021b), we utilize a modified ResNet-50 (He et al., 2016) backbone as the image encoder. These enhancements include modifications to the input stem proposed by (He et al., 2018), implementing anti-aliasing rect-2 blur pooling from (Zhang et al., 2019), and replacing the final average pooling layer with a multi-head QKV attention layer. For our text encoder, we also use the modified Transformer (Vaswani et al., 2017) utilized in CLIP, which contains a 63M-parameter 12-layer 512-wide model with 8 attention heads.

Additional Benchmarks. We also evaluate zero-shot performance on popular downstream datasets: SUN397 (Xiao et al., 2010), Food-101 (Bossard et al., 2014), CIFAR10 (Krizhevsky, 2009), Caltech-101 (Li et al., 2022a), Oxford Pets (Parkhi et al., 2012), STL10 (Coates et al., 2011), KITTI Distance (Geiger et al., 2012), UCF101 (Soomro et al., 2012), Country211 (Radford et al., 2021a), Patch Camelyon (Veeling et al., 2018), Imagenet-A (Hendrycks et al., 2021b), Imagenet-R (Hendrycks et al., 2021a) and Imagenet-Sketch (Wang et al., 2019). Additionally, we isolate the image and text encoders after joint pretraining and evaluate them independently.

We evaluate the image encoder's transfer learning capabilities on the Common Objects in Context (COCO) dataset (Lin et al., 2014). We attach a simple prediction module using instance-aware mask heads from CondInst (Tian et al., 2020b) to our pretrained ResNet-50 models to train on instance segmentation. We train for 100 epochs and evaluate performance on epoch 100. We also evaluate the image encoder on CIFAR100 classication. We attach a single layer linear classifier to the final embedding layer of the ResNet backbone and evaluate one two scenarios: first, by fully fine-tuning the ResNet backbone along with a linear classifier for 10 epochs, and second, by freezing the backbone and only fine-tuning the linear classifier for 10 epochs. With the text encoder, we conduct zero-shot evaluation tasks 2012-2016 from Semantic Textual Similarity (STS12-STS-16) (Agirre et al., 2012; 2013; 2014; 2015; 2016), STS Benchmark (STS-B) (Wang et al., 2018), and SICK-Relatedness (SICK-R) (Marelli et al., 2014).

Evaluation of the CLIP framework. We first assess the performance of the complete CLIP model with zero-shot classification. As depicted in Table 1, we compare OT-CLIP losses to previously proposed contrastive losses/models e.g. InfoNCE, HNS, Triplet, and CyCLIP. We discover that our loss outperforms traditional InfoNCE and other losses on most datasets. On average, our OT-based loss with Gromov inner optimization exhibits the best performance.

Evaluation of the Image Encoder. We further evaluate the image encoder to assess the effectiveness of different losses based on various downstream tasks. Table 2 presents the results for classification, segmentation, and object detection tasks. For detection and segmentation, we report mean average precision (mAP) and mean average recall (mAR) over *IoU=0.5:0.95* in Table 2, where we observe that vanilla Entropic OT (with Sinkhorn-based loss) achieves the best segmentation results and performs competitively in other evaluations. Regarding the classification task, we find inner optimization with Unbalanced OT performs competitively.

Evaluation of the Text Encoder. We compare our OT-based loss to others by evaluating sentence representations on STS tasks. Table 3 presents the evaluation results for 7 STS tasks. It can be observed that the Gromov OT-based loss performs competitively for the STS12 task and achieves

Table 1. Top-1 classification accuracy (%) by Zero-shot on CLIP models pretrained on CC3M by different losses. Best results are in **bold**.

	Food-101	CIFAR-10	CIFAR-100	SUN397	Pets	Flowers	Caltech-101	STL-10	EuroSAT	GTSRB	KITTI Dist	UCF-101	Country211	PCAM	ImgNt-A	ImgNt-R	ImgNt-Skt	ImgNtHK	Average
InfoNCE	9.3	23.6	9.7	29.0	11.2	1.7	38.7	61.7	21.7	3.4	30.8	16.8	0.6	60.82	1.44	10.59	6.83	16.3	19.2
HardContrastive	10.0	19.7	7.64	28.3	10.6	0.78	39.5	62.44	23.26	4.7	**43.45**	16.63	0.5	49.91	1.41	10.74	6.5	15.9	19.1
Triplet	9.38	17.79	9.06	28.8	11.16	0.92	40.6	65.86	20.64	6.59	14.1	16.86	0.61	51.84	**1.65**	11.06	6.7	16.8	18.1
CyCLIP	10.7	16.26	9.39	33.2	8.72	1.02	38.9	**73.0**	16.7	6.3	35.6	19.8	0.62	47.8	1.48	10.4	5.9	16.7	19.2
OT-Based Loss Family with different OT variants for inner optimization																			
Entropic OT	10.3	**25.2**	8.6	28.7	**13.2**	1.8	39.7	63.6	18.2	3.5	27.3	16.9	0.6	50.02	1.39	11.55	6.78	16.7	18.7
DBOT	**10.8**	24.6	8.8	30.7	12.1	1.3	**41.1**	64.3	21.9	4.1	33.6	19.2	**0.7**	49.96	1.28	10.81	**7.93**	17.0	19.6
FGromov OT	10.6	24.9	9.1	**31.2**	12.5	1.8	39.7	63.9	20.3	3.0	28.5	**19.4**	0.5	**64.98**	1.49	11.27	7.79	16.6	**20.6**
Unbal. OT	10.1	26.62	**11.71**	30.55	10.53	1.27	38.7	63.4	11.2	**8.75**	31.4	17.9	0.59	50.31	1.37	**11.65**	6.57	**17.1**	19.0
U-FG OT	10.1	24.8	10.55	30.46	12.9	**2.05**	39.5	64.95	**26.18**	5.07	33.5	17.0	0.63	54.9	1.48	11.35	7.27	16.8	20.1

Table 2. Top-1 classification accuracy of image encoder after CC3M pretrain on CIFAR100. Mean average precision and recall are reported for detection and segmentation on COCO at *IoU=0.5:0.95* with 100 max detections. Best are in **bold**.

	Classification		**Detection**		**Segmentation**	
	Linear	Full	mAP	mAR	mAP	mAR
InfoNCE	45.4	68.0	0.206	0.295	0.155	0.221
HNS	**46.8**	69.2	0.204	0.294	0.122	0.185
Triplet	45.5	65.6	**0.207**	0.295	0.152	0.217
CyCLIP	31.3	**70.1**	**0.207**	**0.297**	0.162	0.230
OT-Based Loss Family with different OT variants for inner optimization						
Entropic OT	46.4	67.9	0.205	0.295	**0.166**	**0.237**
DBOT	44.7	68.1	**0.207**	0.296	0.161	0.228
FGromov OT	45.9	67.6	0.204	0.295	0.151	0.216
Unbal. OT	46.0	68.4	0.196	0.288	0.116	0.179
U-FG OT	45.4	68.1	**0.207**	0.295	0.157	0.222

Table 3. Semantic evaluation on text encoder after CC3M pretrain. Best results highlighted in **bold**. 'STSB' and 'SICKR' denotes 'STSBenchmark' and 'SICKRelatedness', respectively.

	STS12	STS13	STS14	STS15	STS16	STSB	SICKR	Average
InfoNCE	44.26	56.51	52.56	66.95	56.62	61.27	62.87	57.29
HNS	45.3	52.28	51.63	67.52	55.90	62.36	62.25	56.75
Triplet	44.15	52.15	52.10	67.70	55.47	**64.66**	63.74	57.14
CyCLIP	26.68	39.93	37.27	50.84	36.98	45.89	57.67	42.47
OT-Based Loss Family with different OT variants for inner optimization								
Entropic OT	44.25	56.28	52.24	67.89	55.91	61.60	62.76	57.28
DBOT	**44.90**	49.48	50.58	64.25	53.77	60.77	62.83	55.23
FGromov OT	44.89	**56.67**	**53.70**	**67.94**	**57.04**	63.75	63.42	**58.20**
Unbal. OT	42.81	53.21	51.53	67.46	56.78	63.25	63.26	56.90
U-FG OT	41.86	54.15	51.05	67.48	57.51	62.83	**64.21**	57.01

the best performance for the remaining STS tasks. This indicates that learning the structural information of batch data can help improve the text encoder.

4. OT-CLIP for Inference Process

In addition to viewing CLIP training as an optimization problem with OT, we can also adopt an OT perspective for the zero-shot classification task of CLIP during inference. The main difference is that the training process is more focused on learning representations for one-to-one matching, while the inference process does not aim to get one-to-one matching results. Instead, the inference can involve many-to-one conditions where the label measure is not one vector. In the next subsection, we first propose an approach from the graph matching perspective for zero-shot classification in CLIP with fused Gromov-Wasserstein distance. This involves leveraging the similarity or dissimilarity between images/labels to assist in the classification process.

4.1. Graph Matching View for Zero-shot Classification

From the perspective of OT, zero-shot classification in CLIP can be seen as a special case of matching between image feature nodes and text feature nodes, where the distribution of labels is not constrained. Building upon fused Gromov OT, which is commonly used for graph matching, we propose a many-to-one graph matching approach for zero-shot classification, where the label distribution is not constrained. This can be formulated as follows:

$$\min_{\mathbf{P}\in\mathcal{C}(\mathbf{a})} \mathcal{L} =< \mathbf{C}, \mathbf{P} > -\alpha_0 < \mathbf{D}_1, \mathbf{P}\mathbf{D_2}\mathbf{P}^\top > -\epsilon H(\mathbf{P}), \tag{11}$$

where $\mathcal{C}(\mathbf{a}) = \{\mathbf{P}|\mathbf{P1} = \mathbf{1}\}$. Note the optimziation is non-convex and one method is to follow the algorithm in Fused Gromov OT (Xu & Cheng, 2023). The solution can be simplified by an iterative algorithm by

$$\mathbf{P}^{(l+1)} = \min_{\mathbf{P}\in\mathcal{C}(\mathbf{a})} < \mathbf{C}^{(l)}, \mathbf{P} > -\epsilon H(\mathbf{P}), \tag{12}$$

where $\mathbf{C}^{(l)} = \mathbf{C} - \alpha_0\mathbf{D}_1\mathbf{P}^{(l)}\mathbf{D}_2^\top$. Note the optimization in Eq. 12 can be solved with a closed form via softmax given the logit matrix $-\mathbf{C}^{(l)}$. For simplicity, we refer to the method as **FGromov OT-ISoftmax**, which iteratively uses the softmax for prediction. Another method uses the projected gradient descent algorithm (Peyré et al., 2016) where the projection is computed by KL divergence. The

iterations can be given by

$$\begin{aligned} &\mathbf{P} \leftarrow \text{Proj}^{\text{KL}}_{\mathcal{C}(\mathbf{a})}(\mathbf{P} \cdot e^{-\tau(\nabla \mathcal{L}(\mathbf{P}))}) \\ \text{where} \quad &\text{Proj}^{\text{KL}}_{\mathcal{C}(\mathbf{a})}(\mathbf{K}) = \arg \min_{\mathbf{P}' \in \mathcal{C}(\mathbf{a})} KL(\mathbf{P}'|\mathbf{K}). \end{aligned} \tag{13}$$

We refer to this approach as **FGromov OT-PGD**, where the projected gradient descent method is employed for prediction. Compared to traditional algorithms for fused Gromov OT, we replace $U(\mathbf{a}, \mathbf{b})$ with $\mathcal{C}(\mathbf{a})$, assuming that the distribution of labels is unknown. In the next subsection, we consider a setting based on the long-tail problem, where we assume that the distribution of labels is known, allowing us to evaluate the effectiveness of the Sinkhorn-based method.

4.2. Long-tailed Inference

Next, we study the long-tailed classification problem, in which it is typically assumed that the training dataset follows a long-tailed distribution for labels, while the test data may follow a uniform, long-tailed, or reverse long-tailed distribution. From the perspective of OT, when considering this particular classification problem in the inference process, the constraint set $\mathcal{C}(\mathbf{a})$ can be restored to the original $U(\mathbf{a}, \mathbf{b})$, where $\mathbf{b} = N \cdot \mathbf{r}$ is given by the label distribution $\mathbf{r}$ and the number of samples N. Similar to Section 4.1, we adopt a graph matching approach to solve the long-tail classification problem. Adopting the fused Gromov-Wasserstein optimization for long-tailed inference, the zero-shot classification can be formulated as

$$\min_{\mathbf{P} \in U(\mathbf{a}, \mathbf{b})} < \mathbf{C}, \mathbf{P} > -\alpha_0 < \mathbf{D}_1, \mathbf{P}\mathbf{D}_2\mathbf{P}^\top > -\epsilon H(\mathbf{P}). \tag{14}$$

We follow the approach of (Xu & Cheng, 2023; Peyre & Cuturi, 2019) and use the iterative Sinkhorn algorithm to solve the optimization problem. The resulting coupling solution $\mathbf{P}$ actually represents the prediction confidence, and we select the label with the highest confidence for each sample. In addition to zero-shot long-tailed inference, where $\mathbf{C}$ is directly calculated by OpenAI CLIP, we also do fine-tuning based on long-tailed data and adopt different inference method for classification. Comparing to other inference methods as shown in Table 5, our approach performs competitively.

4.3. Selective Zero-shot Classification

In this subsection, we show the OT-based inference is helpful for other zero-shot classification task, i.e. Selective Zero-shot classification (Song et al., 2018). Due to current technological limitations, zero-shot classification struggles to achieve the same level of accuracy and reliability as supervised learning. An alternative approach is Selective Zero-shot Classification, which reduces the risk of misclassifications by rejecting examples that fall below a confidence threshold. Previous works may focus on designing a confidence function, which can involve multiple hyperparameters that need to be controlled through cross-validation.

In this paper, we attempt to simplify Selective Zero-shot Classification by introducing a selective rate (sample percentage), in which the users directly specify the number or percentage of samples that need to be classified, and then the model selects samples based on their confidence levels for classification. One straightforward method is to use the softmax probabilities as confidence scores for comparing samples and selecting high-confidence samples for classification while rejecting low-confidence ones. Building upon the OT-based inference perspective, we propose two approaches to improve performance:

Unbalanced OT. In the matching perspective, there are cases where points from the source set and from the target set cannot be matched one-to-one, i.e. certain points in both sets do not have any corresponding matches. The idea is to use unbalanced OT by replacing the hard constraints with a KL penalty. We introduce this concept into selective classification, which allows us to classify only a subset of images, rather than all of them. The optimization for selective classification can be specified as follows:

$$\min_{\mathbf{P} \geq \mathbf{0}} < \mathbf{C}, \mathbf{P} > -\epsilon H(\mathbf{P}) - \tau_1 KL(\mathbf{P}\mathbf{1}|\mathbf{a}), \tag{15}$$

where τ_1 is a hyperparameter, and the optimization can be solved by the algorithm proposed in (Chizat et al., 2018). Note that when $\tau_1 \to +\infty$, the solution of Eq. 15 equals the Softmax of $-\mathbf{C}$. The advantage of adopting unbalanced OT is that if the model considers that a sample's features differ significantly from all label features, the probability sum of that sample for all labels will decrease, which is helpful in rejecting the sample.

Partial OT. Another alternative selection is to adopt Partial OT, which replaces the strict equality constraint with an inequality constraint. For selective classification in CLIP, we modify the constraints accordingly, and the optimization can be specified as follows:

$$\min_{\mathbf{P} \geq \mathbf{0}} < \mathbf{C}, \mathbf{P} > -\epsilon H(\mathbf{P}) \text{ s.t. } \mathbf{P}\mathbf{1} \leq \mathbf{1},\ \mathbf{1}^\top \mathbf{P}\mathbf{1} = s, \tag{16}$$

where s is a prior constant representing the number of accepted samples, indicating that CLIP selects to trust s predicted results. The optimization of Eq. 16 can be solved by iterative Bregman projections as proposed in (Benamou et al., 2015) and Table 6 shows the results for Partial OT.

4.4. Experiments on CLIP Inference

Datasets. We primarily evaluate OT-CLIP in the inference setting on the same downstream datasets used in Section 3. We generate long-tailed distributions of the ImageNet, CIFAR100, and Places365 (Zhou et al., 2017) validation sets using an imbalance ratio of 10. For long-tailed training, we use the training distributions by (Kang et al., 2019).

Table 4. Top-1 accuracy (%) comparison of Zero-shot inference methods using OpenAI pretrained CLIP. Vanilla softmax is compared with our FGromov OT-based methods (FGromov OT-ISoftmax and FGromov OT-PGD), which consider the classification by graph matching.

Methods	Food-101	CIFAR-10	CIFAR-100	SUN397	DTD	Pets	Flowers	Caltech-101	STL-10	EuroSAT	GTSRB	KITTI Dist	UCF-101	Country211	PCAM	ImgNt-A	ImgNt-R	ImgNt-Skt	ImgN1K	Average
Vanilla Softmax	75.1	65.74	33.23	57.99	39.8	82.18	7.15	67.67	92.15	31.98	29.52	17.7	58.34	13.83	65.4	10.0	28.22	**6.21**	59.8	44.3
FGromov OT-ISoftmax (ours)	**81.92**	**69.6**	**33.62**	**67.6**	39.6	**83.5**	**7.7**	67.6	89.8	**45.6**	**52.8**	11.36	58.6	**36.9**	51.2	**20.2**	**58.5**	6.2	**65.4**	**49.9**
FGromov OT-PGD (ours)	75.5	67.5	33.5	58.3	39.7	82.3	7.1	**67.8**	**92.3**	32.9	29.9	**17.9**	**58.4**	13.8	**67.1**	10.4	29.1	6.2	59.9	44.7

Table 5. Comparison of top-1 accuracy (%) across different testing inference methods under non-finetuned (Zero-shot) and finetuned conditions, for datasets (CIFAR100, ImageNet, Places385) and data distributions (longtailed - LT, uniform - U, reverse longtailed - RLT), based on OpenAI pretrained CLIP (Radford et al., 2021b).

	CIFAR100						**ImageNet**						**Places385**					
	Zero-shot			Finetuned			Zero-shot			Finetuned			Zero-shot			Finetuned		
Testing Inference	LT	U	RLT	LT	U	RLT	LT	U	RLT	LT	U	RLT	LT	U	RLT	LT	U	RLT
Classifier Normalize	–	–	–	52.7	**43.3**	34.0	–	–	–	50.3	44.4	42.7	–	–	–	27.7	22.5	21.9
Class-Aware Bias	–	–	–	41.5	38.8	36.5	–	–	–	50.9	47.9	43.3	–	–	–	24.1	23.7	25.7
Vanilla Softmax	31.8	32.8	32.7	45.1	36.6	28.2	58.6	57.6	60.9	50.3	44.4	42.7	37.9	36.8	39.7	23.4	22.5	25.0
FGromov OT (ours)	**43.5**	**33.1**	**37.4**	**66.4**	37.3	**42.6**	**63**	**63.8**	**65.1**	**52.2**	**59.9**	**43.4**	**56.7**	**63.6**	**56.7**	**43.8**	**45.5**	**40.7**

Table 6. Top-1 accuracy (%) comparison of selective zero-shot classification methods using OpenAI pretrained CLIP. Vanilla softmax is compared with our Unbalanced OT and Partial OT based methods, which modifying hard constraints in Softmax to soft ones.

	Food-101	CIFAR-10	CIFAR-100	SUN397	DTD	Pets	Flowers	Caltech-101	STL-10	EuroSAT	GTSRB	KITTI Dist	UCF-101	Country211	PCAM	ImgNt-A	ImgNt-R	ImgNt-Skt	ImgN1K	Average
Softmax	92.1	85.4	50.6	71.5	51.5	93.6	7.0	92.1	99.5	39.3	**43.7**	15.5	78.6	22.8	58.6	12.9	48.0	11.2	68.8	54.9
Unbal. OT	**93.0**	85.1	50.3	**73.0**	57.7	**95.0**	7.3	92.0	**99.6**	**41.4**	43.5	**18.0**	79.0	**23.0**	75.1	**13.0**	**48.3**	**11.7**	**76.0**	**56.9**
Partial OT	92.4	**85.6**	**50.8**	71.8	**57.7**	94.1	**7.4**	**92.2**	**99.6**	39.6	43.6	17.9	**79.1**	22.9	**77.0**	**13.0**	48.1	11.3	72.2	56.6

Additional Benchmarks. Similar to Sec. 3.3, we evaluate zero-shot performance on popular downstream datasets: SUN397 (Xiao et al., 2010), Food-101 (Bossard et al., 2014), CIFAR10 (Krizhevsky, 2009), Caltech-101 (Li et al., 2022a), Oxford Pets (Parkhi et al., 2012), STL10 (Coates et al., 2011), KITTI Distance (Geiger et al., 2012), UCF101 (Soomro et al., 2012), Country211 (Radford et al., 2021a), Patch Camelyon (Veeling et al., 2018), Imagenet-A (Hendrycks et al., 2021b), Imagenet-R (Hendrycks et al., 2021a) and Imagenet-Sketch (Wang et al., 2019).

Experiments on (vanilla) zero-shot classification. Given a pretrained CLIP model, we enhance zero-shot classification by employing Fused GW methods, specifically FGW-ISoftmax and FGW-PGrad. The results are presented in Table 4. It can be observed that Fused GW-based algorithms perform competitively across most datasets.

Experiments on Long-tailed Inference. We tested the long-tailed inference task on CIFAR-100, ImageNet-1K, and Places385 datasets. Alongside traditional softmax, we also include classifier normalization (Kang et al., 2019) and class-aware bias (Menon et al., 2020) methods for comparison in the finetune setting. We finetune CLIP on long-tailed data for 10 epochs with a standard linear classification head using cross-entropy loss. For inference, we predict without the pretrained/trained prediction head and calculate logits using both the image and text encoders.The exception is the classifier normalization method, which uses a classification head with cosine normalization (Kang et al., 2019) and retains its prediction head for inference. Firstly, for the zero-shot case, we found that FGWISoftmax outperforms the direct use of Softmax by a significant margin across all datasets. Secondly, for the fine-tuning case, we found that FGWISoftmax is superior to all other inference comparisons, including classifier normalization which retained its classification head from training as presented in Table 5.

Experiments on Selective Zero-shot Classification. Table 6 presents the results of selective zero-shot classification. We set the selective rate to 50% and compare the results of Softmax, Unbalanced OT, and Partial OT in Table 6. It can be observed that both Unbalanced OT and Partial OT improve performance compared to directly using Softmax. Furthermore, by comparing the results in Table 4 and Table 6, we can observe that Selective Zero-shot Classification can help users reject misclassified samples.

5. Conclusion

We have proposed an OT perspective to CLIP for both its training and inference. Our mathematical understanding and

derived techniques help establish a new family of loss functions for CLIP training, whose efficacy is verified on public benchmarks. We also devise graph-based representations for images and texts, and the inference could be regarded as a matching procedure. It leads to new techniques not only for standard zero-shot classification but also the practical settings in terms of long-tailed recognition as well as selective classification where the model is required to perform well on a subset of samples with high prediction scores.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning, specifically the multi-modal CLIP model. CLIP itself has broad applications and impact thus our proposed technique shall be considered when such models are used in any case.

Acknowledgments

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

References

Agirre, E., Diab, M., Cer, D., and Gonzalez-Agirre, A. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, 2012.

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., and Guo, W. * sem 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 1, pp. 32–43, 2013.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., et al. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 81–91, 2014.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., and Wiebe, J. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 252–263, 2015.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., and Wiebe, J. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 497–511, 2016.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101 – mining discriminative components with random forests. In *Computer Vision – ECCV 2014*, pp. 446–461. Springer International Publishing, 2014. ISBN 978-3-319-10599-4.

Chang, W., Shi, Y., Tuan, H., and Wang, J. Unified optimal transport framework for universal domain adaptation. *Advances in Neural Information Processing Systems*, 35: 29512–29524, 2022.

Chapel, L., Alaya, M. Z., and Gasso, G. Partial optimal tranport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33: 2903–2913, 2020.

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. PMLR, 2020.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

Coates, A., Ng, A. Y., and Lee, H. An analysis of single-layer networks in unsupervised feature learning. *Journal of Machine Learning Research*, 15:215–223, 2011.

Damodaran, B. B., Kellenberger, B., Flamary, R., Tuia, D., and Courty, N. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 447–463, 2018.

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.

Dessein, A., Papadakis, N., and Rouas, J.-L. Regularized optimal transport and the rot mover's distance, 2018.

Essid, M. and Solomon, J. Quadratically regularized optimal transport on graphs. *SIAM Journal on Scientific Computing*, 40(4):A1961–A1986, 2018.

Feydy, J., Charlier, B., Vialard, F.-X., and Peyré, G. Optimal transport for diffeomorphic registration. In *Medical Image Computing and Computer Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pp. 291–299. Springer, 2017.

Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361, 2012.

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.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. Improved training of wasserstein gans. In *NIPS*, 2017.

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.

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., and Li, M. Bag of tricks for image classification with convolutional neural networks, 2018.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. *CVPR*, 2021b.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., and Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

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

Kolkin, N., Salavon, J., and Shakhnarovich, G. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10051–10060, 2019.

Krizhevsky, A. Learning multiple layers of features from tiny images. 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

Li, F.-F., Andreeto, M., Ranzato, M., and Perona, P. Caltech 101, 2022a.

Li, Y., Mo, Y., Shi, L., and Yan, J. Improving generative adversarial networks via adversarial learning in latent space. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 8868–8881. Curran Associates, Inc., 2022b.

Li, Y., Guo, J., Wang, R., and Yan, J. T2t: From distribution learning in training to gradient search in testing for combinatorial optimization. In *Advances in Neural Information Processing Systems*, 2023a.

Li, Y., Shi, L., and Yan, J. Iid-gan: an iid sampling perspective for regularizing mode collapse. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, 2023b.

Liero, M., Mielke, A., and Savaré, G. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. *European conference on computer vision*, pp. 740–755, 2014.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. A sick cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 2014.

Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A., and Kumar, S. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Mu, N., Kirillov, A., Wagner, D. A., and Xie, S. SLIP: self-supervision meets language-image pre-training. *CoRR*, abs/2112.12750, 2021. URL https://arxiv.org/abs/2112.12750.

Muzellec, B., Nock, R., Patrini, G., and Nielsen, F. Tsallis regularized optimal transport and ecological inference. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

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

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Peyre, G. and Cuturi, M. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6): 355–607, 2019.

Peyré, G., Cuturi, M., and Solomon, J. Gromov-wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pp. 2664–2672. PMLR, 2016.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021a. URL https://arxiv.org/abs/2103.00020.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021b.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Shi, L., Zhang, G., Zhen, H., Fan, J., and Yan, J. Understanding and generalizing contrastive learning from the inverse optimal transport perspective. In *International Conference on Machine Learning*, 2023.

Shi, L., Shen, Z., and Yan, J. Double-bounded optimal transport for advanced clustering and classification. *AAAI*, 2024a.

Shi, L., Zhen, H., Zhang, G., and Yan, J. Relative entropic optimal transport: a (prior-aware) matching perspective to (unbalanced) classification. *Advances in Neural Information Processing Systems*, 36, 2024b.

Song, J., Shen, C., Lei, J., Zeng, A.-X., Ou, K., Tao, D., and Song, M. Selective zero-shot classification with augmented attributes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 468–483, 2018.

Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

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, 2020a.

Tian, Z., Shen, C., and Chen, H. Conditional convolutions for instance segmentation. *CoRR*, abs/2003.05664, 2020b. URL https://arxiv.org/abs/2003.05664.

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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. Fused gromov-wasserstein distance for structured objects: theoretical foundations and mathematical properties, 2018.

Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., and Welling, M. Rotation equivariant cnns for digital pathology. *CoRR*, abs/1806.03962, 2018. URL http://arxiv.org/abs/1806.03962.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pp. 10506–10518, 2019.

Wang, R., Guo, Z., Jiang, S., Yang, X., and Yan, J. Deep learning of partial graph matching via differentiable top-k. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6272–6281, 2023.

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.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey

to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. doi: 10.1109/CVPR.2010.5539970.

Xu, H. and Cheng, M. Regularized optimal transport layers for generalized global pooling operations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

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. PMLR, 2021.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. Making convolutional networks shift-invariant again. *arXiv preprint arXiv:1904.11486*, 2019.

Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L. H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16793–16803, 2022.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 40, pp. 1452–1464. IEEE, 2017.

Zhu, J., Shi, L., Yan, J., and Zha, H. Automix: Mixup networks for sample interpolation via cooperative barycenter learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pp. 633–649. Springer, 2020.

A. OT Losses for CLIP Training

A.1. Entropic OT

Algorithm 1 PyTorch-style pseudocode for Entropic OT

```
# M:Metric cost matrix
# a:Samples weights in the source domain
# b:Samples weights in the target domain
# reg:Regularization term > 0
# N:Max number of iterations

# Initialize u and v as uniform distributions
u = torch.ones(a.shape[0]) / a.shape[0]
v = torch.ones(b.shape[0]) / b.shape[0]

# Compute the kernel K using the negative cost matrix scaled by the
regularization term
K = torch.exp(M / -reg)
# Precompute the row normalization factor
Kp = (1/a).reshape(-1, 1) * K

# Iteratively update u and v using Sinkhorn algorithm
for i in range(N):
    # Compute the matrix-vector product of K transposed and u
    KtransposeU = K.t() @ u
    # Update v based on the current u
    v = b / KtransposeU
    # Update u using the new v and precomputed Kp
    u = 1.  / (Kp @ v)

# Return the optimal transport plan P computed from u and v
return u.reshape((-1, 1)) * K * v.reshape((1, -1))
```

For training, we used $reg = 0.01$ and $N = 5$. a was initialized as a tensor of ones of length equivalent to the training batch size. b was also a tensor of length training batch size, but with values 1 / batch size.

A.2. DBOT

Algorithm 2 PyTorch-style pseudocode for DBOT

```
# M:Metric cost matrix between samples in the source and target domains
# reg:Regularization term > 0
# N:Max number of iterations

# Initialize the number of samples based on the cost matrix size
n = M.shape[0]
# Initialize source distribution as uniform
a = torch.ones((n,))
# Define lower bound for target distribution weights
b_d = torch.ones((n,), 0.1 * n)
# Define upper bound for target distribution weights
b_u = torch.full((n,), 0.9 * n)

# Initialize the transport plan P with regularized cost matrix
P = torch.exp(-M/reg)
# Iteratively adjust P to satisfy constraints
for i in range(N):
    # Normalize P row-wise to match source distribution
    sum_P = P.sum(dim=1)
    P = torch.diag(a / sum_P) @ P

    # Adjust P not to exceed upper bound of target distribution
    sum_P_t = P.t().sum(dim=1)
    P = P @ torch.diag(torch.max(b_d / sum_P_t, torch.ones(P.shape[1])))

    # Adjust P to meet at least lower bound of target distribution
    sum_P_t = P.t().sum(dim=1)
    P = P @ torch.diag(torch.min(b_u / sum_P_t, torch.ones(P.shape[1])))
# Return the adjusted optimal transport plan P
return P
```

For training, we used $N = 5$ and $reg = 1.0$.

A.3. Fused Gromov OT

Algorithm 3 PyTorch-style pseudocode for Fused Gromov OT

```
# image_f:Image features
# text_f:Text features
# a1:Regularization term for scaling the KL divergence term 1
# a2:Regularization term for scaling the KL divergence term 2
# reg:Regularization term > 0
# N:Max number of iterations for adjusting the structural cost matrix

# KL term 1:  Compute the self-similarity matrix for image features
sigma_1 = image_f @ image_f.t()
# KL term 2:  Compute the self-similarity matrix for text features
sigma_2 = text_f @ text_f.t()

# Construct initial cost matrix based on the negative dot product (similarity)
between image and text features
C = 1.0 - image_f @ text_f.t()
# Initialize the transport plan P
P = softmax(-C / reg)

# Iteratively refine the structural cost matrix
for i in range(N):
    # Adjust the cost matrix C
    C = C - a1 * sigma_1 @ P @ sigma_2 * a2
    # Update the transport plan P
    P = sinkhorn(1.0 - C, numIters=5)
# Return transport plan P
return P
```

For training, we used $N = 5$, $reg = 0.01$, $a1 = 0.01$, and $a2 = 0.01$.

A.4. Unbalanced-Fused Gromov OT / Unbalanced OT

Algorithm 4 PyTorch-style pseudocode for Unbalanced-Fused Gromov OT

```
# x:Cosine sim matrix of image and text features
# c1:Cost matrix of image features
# c2:Cost matrix of text features
# p0:Marginal prior of dimensions
# q0:Marginal prior of samples
# a0:Weight of GW term
# a1:Weight of entropic term
# a2:Weight of KL term for p0
# a3:Weight of KL term for q0
# tau:Regularization coefficient
# num:Number of outer iterations
# inner:Number of inner Sinkhorn iterations
# eps:Epsilon to avoid numerical instability

# Initial setup:
t = q0 * p0 # Initialize transport matrix t
log_p0 = torch.log(p0 + eps)
log_q0 = torch.log(q0 + eps)

# Iterative optimization:
for m in range(num):
    n = min(m, a1.shape[0] - 1)
    a11 = a1[n] + tau
    tmp1 = torch.matmul(c2, t)
    tmp2 = torch.matmul(tmp1, c1)
    cost = -x - a0[n] * tmp2 - tau * torch.log(t + eps)
    a = torch.zeros_like(p0)
    b = torch.zeros_like(q0)
    y = -cost / a11
    # Sinkhorn normalization:
    for k in range(inner):
        log_p = torch.logsumexp(y - log_p0, dim=2, keepdim=True)
        log_q = torch.logsumexp(y - log_q0, dim=1, keepdim=True)
        a = a2[n] / (a2[n] + a11) * (log_p0 - log_p)
        b = a3[n] / (a3[n] + a11) * (log_q0 - log_q)
        y = -cost / a11 + a + b
     t = torch.exp(y)
return t
```

For training with U-FG OT, we use $a0 = a1 = a2 = a3 = 0.01$, $inner = 5$, and $num = 4$. $q0$ and $p0$ are initialized as a tensor of 0.001 of length batch size. For training with Unbal. OT, we use the same parameters as U-FG OT except with $num = 1$ and $a0 = 0$.

Special thanks to authors (Xu & Cheng, 2023) for their implementation of regularized optimal transport layers for pooling (ROTP) that provided much of the structural implementation for this algorithm.

B. OT Losses for CLIP Inference

B.1. FGWISoftmax

Algorithm 5 PyTorch-style pseudocode for FGWISoftmax

```
# image_f:Image features
# text_f:Text features
# a1:Regularization term for scaling the KL divergence term 1
# a2:Regularization term for scaling the KL divergence term 2
# reg:Regularization term > 0
# N:Max number of iterations for adjusting the structural cost matrix

# KL term 1:  Compute the self-similarity matrix for image features
sigma_1 = image_f @ image_f.t()
# KL term 2:  Compute the self-similarity matrix for text features
sigma_2 = text_f @ text_f.t()

# Construct initial cost matrix based on the negative dot product (similarity)
between image and text features
C = 1.0 - image_f @ text_f.t()
# Initialize the transport plan P
P = softmax(-C / reg)

# Iteratively refine the structural cost matrix
for i in range(N):
    # Adjust the cost matrix C
    C = C - a1 * sigma_1 @ P @ sigma_2 * a2
    # Update the transport plan P
    P = softmax(-C / reg)
# Return transport plan P
return P
```

B.2. FGW-PGrad

Algorithm 6 PyTorch-style pseudocode for FGW-PGrad

```
# image_f:Image features
# text_f:Text features
# a1:Regularization term for scaling the KL divergence term 1
# a2:Regularization term for scaling the KL divergence term 2
# reg:Regularization term > 0
# N:Max number of iterations for adjusting the structural cost matrix

# KL term 1:  Compute the self-similarity matrix for image features
sigma_1 = image_f @ image_f.t()
# KL term 2:  Compute the self-similarity matrix for text features
sigma_2 = text_f @ text_f.t()

# Construct initial cost matrix based on the negative dot product (similarity)
between image and text features
C = 1.0 - image_f @ text_f.t()
# Initialize the transport plan P
P = softmax(-C)

# Iteratively refine the structural cost matrix
for i in range(N):
    # Adjust the cost matrix C
    grad = C - a1 * sigma_1 @ P @ sigma_2 * a2

    K = P * torch.exp(-grad/reg)
    P = K / K1  # Row normalization
# Return transport plan P
return P
```

B.3. Partial OT

Algorithm 7 PyTorch-style pseudocode for Partial OT

```
# image_f:Image features
# text_f:Text features
# reg:Regularization term > 0
# N:Max number of iterations for adjusting the structural cost matrix

P = torch.exp(image_f @ text_f.t() / reg
m = torch.full((num_class,), 1/num_class)

for i in range(N):
    row_P = P.sum(dim=1)
    scaling = torch.max(row_P, 1)
    P = div(P, scaling)
    P *= m/P.sum(dim=1)  # Row normalization
# Return transport plan P
return P
```

C. Inference Hyperparameters and Datasets

Table 7. Hyperparameter settings for inference results. Reference Appendix Section A for parameter details.

	FGWISoftmax			**FGromov OT**						**FGW-PGrad**				**Partial OT**		**Unbalanced OT**								
	Graph Zero-shot			LT Zero-shot			LT Finetune			Graph Zero-shot				Selective Zero-shot		Selective Zero-shot								
Dataset	reg	$a1$	N	reg	$a1$	N	reg	$a1$	N	$a2$	reg	$a1$	N	reg	N	tau	num	$inner$	$q0$	$p0$	$a0$	$a1$	$a2$	$a3$
Food-101	0.01	0.01	3	–	–	–	–	–	–	0.01	0.01	0.01	5	0.1	100	0.01	5	1	0.01	0.01	0	0.01	0.01	0.01
CIFAR-10	1.0	0.01	3	–	–	–	–	–	–	0.01	0.01	0.01	5	0.1	100	0.01	5	1	0.01	0.01	0	0.01	0.01	0.01
CIFAR-100	1.0	0.01	3	0.01	0.01	3	0.01	0.01	3	0.01	0.01	0.01	5	0.1	100	0.01	5	1	0.01	0.01	0	0.01	0.01	0.01
SUN397	0.01	0.01	3	–	–	–	–	–	–	0.01	0.01	0.01	5	0.1	100	0.01	5	1	0.01	0.01	0	0.01	0.01	0.01
DTD	1.0	0.01	5	–	–	–	–	–	–	0.01	0.01	0.01	5	0.1	100	0.01	5	1	0.01	0.01	0	0.01	0.01	0.01
Oxford Pets	1.0	0.01	5	–	–	–	–	–	–	0.01	0.01	0.01	5	0.1	100	0.01	5	1	0.01	0.01	0	0.01	0.01	0.01
Oxford Flowers	0.01	0.01	3	–	–	–	–	–	–	0.01	0.01	0.01	5	0.1	100	0.01	5	1	0.01	0.01	0	0.01	0.01	0.01
Caltech-101	1.0	0.01	3	–	–	–	–	–	–	0.01	0.01	0.01	5	0.1	100	0.01	1	1	0.01	0.01	0	0.1	0.1	0.1
STL-10	2.0	0.01	5	–	–	–	–	–	–	0.01	0.01	0.01	5	0.1	100	0.01	5	1	0.01	0.01	0	0.01	0.01	0.01
EuroSat	0.01	0.01	3	–	–	–	–	–	–	0.01	0.01	0.01	5	0.1	100	0.01	5	1	0.01	0.01	0	0.01	0.01	0.01
GTSRB	0.01	0.01	3	–	–	–	–	–	–	0.01	0.01	0.01	5	0.1	100	0.01	1	1	0.01	0.01	0	0.1	0.1	0.1
KITTI Distance	1.0	0.01	3	–	–	–	–	–	–	0.01	0.01	0.01	5	0.1	100	0.01	1	1	0.01	0.01	0	0.1	0.1	0.1
UCF101 Frames	1.0	0.01	5	–	–	–	–	–	–	0.01	0.01	0.01	5	0.1	100	0.01	5	1	0.01	0.01	0	0.01	0.01	0.01
Country211	0.01	0.01	3	–	–	–	–	–	–	0.01	0.01	0.01	5	0.1	100	0.01	1	1	0.01	0.01	0	0.1	0.1	0.1
Patch Camelyon	1.0	0.01	3	–	–	–	–	–	–	0.01	0.01	0.01	5	0.1	100	0.01	1	1	0.01	0.01	0	0.1	0.1	0.1
ImageNet-A	0.01	0.01	3	–	–	–	–	–	–	0.01	0.01	0.01	5	0.1	100	0.01	5	1	0.01	0.01	0	0.01	0.01	0.01
ImageNet-R	0.01	0.01	3	–	–	–	–	–	–	0.01	0.01	0.01	5	0.1	100	0.01	5	1	0.01	0.01	0	0.01	0.01	0.01
ImageNet-Sketch	1.0	0.01	5	–	–	–	–	–	–	0.01	0.01	0.01	5	0.1	100	0.01	5	1	0.01	0.01	0	0.01	0.01	0.01
ImageNet1K	0.01	0.01	3	0.1	0.1	5	0.1	0.1	5	0.01	0.01	0.01	5	0.1	100	0.01	5	1	0.01	0.01	0	0.01	0.01	0.01
ImageNet-LT/RLT	–	–	–	0.1	0.1	5	0.1	0.1	5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
CIFAR100-LT/RLT	–	–	–	0.01	0.01	3	0.01	0.01	3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Places365	–	–	–	0.01	0.1	3	0.01	0.1	3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Places365-LT/RLT	–	–	–	0.01	0.1	3	0.01	0.1	3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

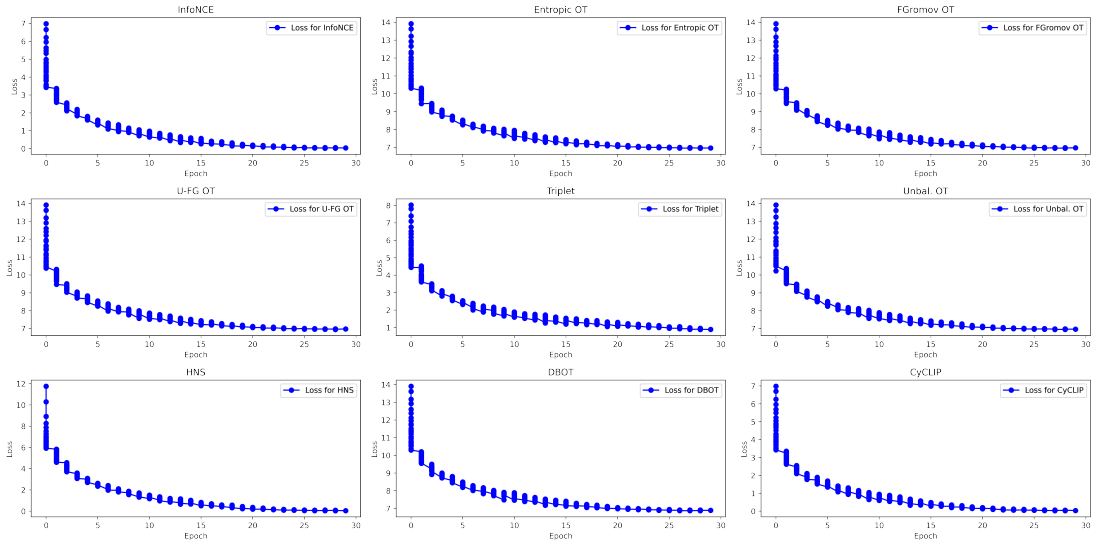

Figure 2. Convergence graphs of proposed OT methods and baselines.

D. Additional Figures

D.1. Computational Cost and Convergence

We compare the computational cost of OT-based Loss and InfoNCE loss for 30 epochs of training below. Since we solely apply OT to the loss function, our optimizations are only applied after classification pooling. Therefore, we do not assume much cost during the training process. Admittedly, the computational cost was only loosely enforced through effective batch size, and our OT methods are more expensive than baseline InfoNCE despite some of our OT methods having slightly lower training times. Most importantly, we want to highlight the comparable training time to that of baseline InfoNCE. Training time is reported on CC3M using effective batch size 1024. Although training losses are not utilized during validation, we provide the inference time on Imagenet1K-val for a more controlled comparison of computational cost. Inference time is reported using batch size 1000, averaged over 10 runs. Inference hyperparameters are set to the same used during training.

Method	Training Time (hours)	Inference Time (seconds)
InfoNCE	17.46	89.34
HNS	17.01	–
Triplet	30.01	–
CyCLIP	17.73	–
Entropic OT	17.01	90.83
DBOT	17.02	90.42
FGromov OT	17.45	90.36
Unbal. OT	17.33	91.2
U-FG OT	17.51	90.37

Table 8. Training and Inference Time on ImageNet-1K

D.2. Influence of Batch Size on Graph-Matching Perspective

OT-based inference will degrade to traditional Softmax prediction if the model needs to predict one sample or a small number of samples at a time. However, in practical applications, it is rare to set the batch size as 1 for inference due to the increased computational cost. Instead, higher batch sizes are often used to enable parallel prediction. Below we show our OT inference results in the graph matching perspective for ImageNet-1K using varying batch sizes. Generally, we can observe an upward trend in Top-1 Accuracy as we increase the batch size. With the exception for FGromov-OT-ISoftmax, where the best results aligned with the batch size with closest to the number of samples per class during non-shuffled evaluation. For

comparison, baseline Softmax achieves **59.82%** across all batch sizes.

Table 9. ImageNet-1K Classes in Sequential Order — Graph Matching Perspective using Increasing Batch Size

Method	BS 32	BS 64	BS 128	BS 256	BS 512	BS 1024	BS 2048	BS 4096	Whole
FGromov-OT-PGD	59.82	59.82	59.82	59.84	59.86	59.87	59.92	59.91	59.95
FGromov-OT-ISoftmax	59.9	73.6	66.3	65.4	63.0	61.0	61.1	61.1	61.1

Table 10. ImageNet-1K Random Shuffle — Graph Matching Perspective using Increasing Batch Size

Method	BS 32	BS 64	BS 128	BS 256	BS 512	BS 1024	BS 2048	BS 4096	Whole
FGromov-OT-PGD	59.84	59.84	59.86	59.91	59.92	59.95	59.92	59.96	59.95
FGromov-OT-ISoftmax	59.87	59.93	60.0	60.0	60.1	60.3	60.5	60.9	60.7

Table 11. Hyperparameters for Table 9

	FGWISoftmax				**FGW-PGrad**			
Batch Size	reg	$a1$	$a2$	N	reg	$a1$	$a2$	N
16	0.01	0.1	0	3	0.01	0.01	0.01	5
32	0.01	0.1	0	3	0.01	0.01	0.01	5
64	0.01	0.1	0	3	0.01	0.01	0.01	5
128	0.1	0.1	0	3	0.01	0.01	0.01	5
256	0.1	0.1	0	3	0.01	0.01	0.01	5
512	0.1	0.1	0	3	0.01	0.01	0.01	5
1024	0.01	0.01	0.01	3	0.01	0.01	0.01	5
2048	0.01	0.01	0.01	10	0.01	0.01	0.01	10
4096	0.01	0.01	0.01	10	0.01	0.01	0.01	15
Whole	0.01	0.01	0.01	13	0.01	0.01	0.01	20