# Wasserstein Contrastive Representation Distillation

Liqun Chen[1], Zhe Gan[2], Dong Wang[1], Jingjing Liu[2], Ricardo Henao[1], Lawrence Carin[1]
[1]Duke University, [2]Microsoft Dynamics 365 AI Research
{liqun.chen, dong.wang, ricardo.henao, lcarin}@duke.edu, {zhe.gan, jingjl}@microsoft.com

## Abstract

*The primary goal of knowledge distillation (KD) is to encapsulate the information of a model learned from a teacher network into a student network, with the latter being more compact than the former. Existing work, e.g., using Kullback-Leibler divergence for distillation, may fail to capture important structural knowledge in the teacher network and often lacks the ability for feature generalization, particularly in situations when teacher and student are built to address different classification tasks. We propose Wasserstein Contrastive Representation Distillation (WCoRD), which leverages both primal and dual forms of Wasserstein distance for KD. The dual form is used for global knowledge transfer, yielding a contrastive learning objective that maximizes the lower bound of mutual information between the teacher and the student networks. The primal form is used for local contrastive knowledge transfer within a mini-batch, effectively matching the distributions of features between the teacher and the student networks. Experiments demonstrate that the proposed WCoRD method outperforms state-of-the-art approaches on privileged information distillation, model compression and cross-modal transfer.*

## 1. Introduction

The recent success of deep learning methods has brought about myriad efforts to apply them beyond benchmark datasets, but a number of challenges can emerge in real-world scenarios. For one, as the scale of deep learning models continues to grow (*e.g.*, [21, 15]), it has become increasingly difficult to deploy such trained networks on more computationally-restrictive platforms, such as smart phones, remote sensors, and edge devices. Additionally, deep networks require abundant data for training, but large datasets are often private [36], classified [29], or institutional [39], which the custodians may be hesitant to release publicly. Labeled datasets in specialized domains may also be rare or expensive to produce. Finally, despite ample datasets from neighboring modalities, conventional frameworks lack clear ways to leverage cross-modal data.

Knowledge distillation (KD), which has become an increasingly important topic in the deep learning community, offers a potential solution to these challenges. In KD, the goal is to improve a *student* model's performance by supplementing it with additional feedback from a *teacher* model. Often the teacher has larger capacity than the student, or has access to additional data that the student does not. As such, KD can transfer this additional and valuable knowledge from the teacher to the student. In early KD methods [23], this supplemental supervision was imposed by asking the student to minimize the Kullback-Leibler (KL) divergence between its output prediction distribution and the teacher's. Given that the prediction probability distribution contains richer and more informative signals than the one-hot labels, student models have been shown to benefit from this extra supervision. However, the low dimensionality of prediction distribution means that the amount of information encoded (therefore transferred) can be limited. For cross-modal transfer, these predictions may even be irrelevant, making KL divergence unable to transfer meaningful information.

In contrast, intermediate representations present an opportunity for more informative learning signals, as a number of recent works have explored [37, 50, 40, 42, 41]. However, as observed by [42], these methods often compare poorly with the basic KD, potentially due to the challenge of defining a proper distance metric between features of the teacher and student networks. Furthermore, they heavily rely on strategies to copy teacher's behavior, *i.e.*, aligning the student's outputs to those from the teacher. We argue that such practice overlooks a key factor: the teacher's experience may not necessarily generalize well to the student.

Motivated by this, we present **W**asserstein **Co**nstrastive **R**epresentation **D**istillation (WCoRD), a new KD framework that reduces the generalization gap between teacher and student to approach better knowledge transfer. Specifically, our approach constitutes *distillation* and *generalization* blocks, realized by solving the *dual* and *primal* form of the Wasserstein distance (WD), respectively. For better distillation, we leverage the dual form of WD to maximize the mutual information (MI) between student and teacher representation distributions, using an objective inspired by Noise

Contrastive Estimation (NCE) [18]. Unlike previous methods [42], we propose to impose a 1-Lipschitz constraint to the critic via spectral normalization [31]. By shifting the critic to one based on optimal transport, we improve stability and sidestep some of the pitfalls of KL divergence minimization [8, 30]. We term this as *global* contrastive knowledge transfer.

For better generalization, we also use the primal form of WD to indirectly bound generalization error via regularizing the Wasserstein distance between the feature distributions of the student and teacher. This results in a relaxation that allows for coupling student and teacher features across multiple examples within each mini-batch, as opposed to the one-to-one matching in previous methods (*i.e.*, strictly copying the teacher's behavior). In principle, this serves to directly match the feature distributions of the student and teacher networks. We term this *local* contrastive knowledge transfer. With the use of both primal and dual forms, we are able to maximize MI and simultaneously minimize the feature distribution discrepancy.

The main contributions are summarized as follows. $(i)$ We present a novel Wasserstein learning framework for representation distillation, utilizing the dual and primal forms of the Wasserstein distance for *global* constrastive learning and *local* feature distribution matching, respectively. $(ii)$ To demonstrate the superiority of the proposed approach, we first conduct comprehensive experiments on benchmark datasets for model compression and cross-modal transfer. To demonstrate versatility, we further apply our method to a real-world dataset for privileged information distillation.

## 2. Background

### 2.1. Knowledge Distillation

In knowledge distillation, a student network is trained by leveraging additional supervision from a trained teacher network. Given an input sample $(\boldsymbol{x}, y)$, where $\boldsymbol{x}$ is the network input and $y$ is the one-hot label, the distillation objective encourages the output probability distribution over predictions from the student and teacher networks to be similar. Assume $\boldsymbol{z}^T$ and $\boldsymbol{z}^S$ are the logit representations (before the softmax layer) of the teacher and student network, respectively. In standard KD [23], the training of the student network involves two supervised loss terms:

$$\begin{aligned}\mathcal{L} = &\mathcal{L}_{\text{CE}}\left(y, \text{softmax}(\boldsymbol{z}^S)\right) \\ &+ \alpha \cdot \text{KL}\left(\text{softmax}(\boldsymbol{z}^T/\rho) \| \text{softmax}(\boldsymbol{z}^S/\rho)\right) \,, \quad (1)\end{aligned}$$

where $\rho$ is the temperature, and $\alpha$ is the balancing weight. The representation in (1) is optimized with respect to the student network parameters $\boldsymbol{\theta}_S$, while the teacher network (parameterized by $\boldsymbol{\theta}_T$) is pre-trained and fixed. The first term in (1) enforces label supervision, which is conventionally implemented as a cross-entropy loss for classification tasks. The second term encourages the student network to produce distributionally-similar outputs to the teacher network. However, there are inevitable limitations to this approach. Deep neural networks learn structured features through the intermediate hidden layers, capturing spatial or temporal correlations of the input data. These representations are then collapsed to a low-dimensional prediction distribution, losing this complex structure. Furthermore, the KL divergence used here can be unstable numerically due to its asymmetry [8, 33]. For example, it can overly concentrate on small details: a small imperfection can put a sample outside the distribution and explode the KL toward infinity. Despite this, KD objectives based on KL divergence can still be effective and remain popular.

We aim to provide a general and principled framework for distillation based on *Wasserstein distance*, where both global contrastive learning and local distribution matching are introduced to facilitate knowledge transfer to the student. By using the Wasserstein metric, we also avoid some of the drawbacks of KL-based approaches. Note that our approach utilizes feature representations at the penultimate layer (before logits), denoted as $\boldsymbol{h}^S$ and $\boldsymbol{h}^T$ for the student and teacher networks, respectively.

### 2.2. Wasserstein Distance

One of the more recently-proposed distance measures in knowledge distillation is the contrastive loss [42]. The goal is to move similar samples closer while pushing different ones apart in the feature space (*i.e.*, $\boldsymbol{z}^S$ and $\boldsymbol{z}^T$, or $\boldsymbol{h}^S$ and $\boldsymbol{h}^T$). We further extend and generalize the idea of contrastive loss with Wasserstein Distance (a.k.a. Earth Mover's Distance, or Optimal Transport Distance). In the following, we give a brief introduction to the *primal* and *dual* forms of the general Wasserstein Distance (WD). The primal form [46] is defined as follows.

**Definition 2.1.** Consider two probability distribution: $\boldsymbol{x}_1 \sim \boldsymbol{p}_1$, and $\boldsymbol{x}_2 \sim \boldsymbol{p}_2$. The Wasserstein-1 distance between $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ can be formulated as:

$$\mathcal{W}(\boldsymbol{p}_1, \boldsymbol{p}_2) = \inf_{\pi \in \Pi(\boldsymbol{p}_1, \boldsymbol{p}_2)} \int_{\mathcal{M}\times\mathcal{M}} c(\boldsymbol{x}_1, \boldsymbol{x}_2) d\pi(\boldsymbol{x}_1, \boldsymbol{x}_2) \,,$$

where $c(\cdot)$ is a point-wise cost function evaluating the distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and $\Pi(\boldsymbol{p}_1, \boldsymbol{p}_2)$ is the set of all possible couplings of $\boldsymbol{p}_1(\boldsymbol{x}_1)$ and $\boldsymbol{p}_2(\boldsymbol{x}_2)$; $\mathcal{M}$ is the space of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and $\pi(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is a joint distribution satisfying $\int_{\mathcal{M}} \pi(\boldsymbol{x}_1, \boldsymbol{x}_2) d\boldsymbol{x}_2 = \boldsymbol{p}_1(\boldsymbol{x}_1)$ and $\int_{\mathcal{M}} \pi(\boldsymbol{x}_1, \boldsymbol{x}_2) d\boldsymbol{x}_1 = \boldsymbol{p}_2(\boldsymbol{x}_2)$.

Using the Kantorovich-Rubenstein duality [46], WD can be written in the dual form:

$$\mathcal{W}(\boldsymbol{p}_1, \boldsymbol{p}_2) = \sup_{||g||_L \leq 1} \mathbb{E}_{\boldsymbol{x}_1 \sim \boldsymbol{p}_1}\left[g(\boldsymbol{x}_1)\right] - \mathbb{E}_{\boldsymbol{x}_2 \sim \boldsymbol{p}_2}\left[g(\boldsymbol{x}_2)\right] \,,$$

where $g$ is a function (often parameterized as a neural network) satisfying the 1-Lipschitz constraint (*i.e.*, $||g||_L \leq 1$).

## 3. Method

We present the proposed Wasserstein learning framework for KD, where $(i)$ the dual form is used for global contrastive knowledge transfer (Sec. 3.1), and $(ii)$ the primal form is adopted for local contrastive knowledge transfer (Sec. 3.2). The full algorithm is summarized in Sec. 3.3.

### 3.1. Global Contrastive Knowledge Transfer

For *global* knowledge transfer, we consider maximizing the mutual information (MI) between feature representations $\boldsymbol{h}^S, \boldsymbol{h}^T$ at the penultimate layer (before logits) from the teacher and student networks. That is, we seek to match the joint distribution $\boldsymbol{p}(\boldsymbol{h}^T, \boldsymbol{h}^S)$ with the product of marginal distributions $\boldsymbol{\mu}(\boldsymbol{h}^T)$ and $\boldsymbol{\nu}(\boldsymbol{h}^S)$ via KL divergence:

$$I(\boldsymbol{h}^T; \boldsymbol{h}^S) = \mathrm{KL}(\boldsymbol{p}(\boldsymbol{h}^T, \boldsymbol{h}^S) \| \boldsymbol{\mu}(\boldsymbol{h}^T)\boldsymbol{\nu}(\boldsymbol{h}^S)) \,. \quad (2)$$

Since both the joint and marginal distributions are implicit, (2) cannot be computed directly. To approximate the MI, Noise Contrastive Estimation (NCE) [18] is used. Specifically, we denote a *congruent* pair as one drawn from the joint distribution, and an *incongruent* pair as one drawn independently from the product of marginal distributions. In other words, a congruent pair is one where the same data input is fed to both the teacher and student networks, while an incongruent pair consists of different data inputs. We then define a distribution $\boldsymbol{q}$ conditioned on $\eta$ that captures whether the pair is congruent ($\boldsymbol{q}(\eta = 1)$) or incongruent ($\boldsymbol{q}(\eta = 0)$), with

$$\boldsymbol{q}(\boldsymbol{h}^T, \boldsymbol{h}^S | \eta = 1) = \boldsymbol{p}(\boldsymbol{h}^T, \boldsymbol{h}^S) \,, \quad (3)$$

$$\boldsymbol{q}(\boldsymbol{h}^T, \boldsymbol{h}^S | \eta = 0) = \boldsymbol{\mu}(\boldsymbol{h}^T)\boldsymbol{\nu}(\boldsymbol{h}^S) \,. \quad (4)$$

With one congruent and one incongruent pair, the prior on $\eta$ is

$$\boldsymbol{q}(\eta = 1) = \boldsymbol{q}(\eta = 0) = 1/(1+1) = 1/2 \,. \quad (5)$$

By Bayes' rule, we can obtain the posterior for $\eta = 1$:

$$\boldsymbol{q}(\eta = 1 | \boldsymbol{h}^T, \boldsymbol{h}^S) = \frac{\boldsymbol{p}(\boldsymbol{h}^T, \boldsymbol{h}^S)}{\boldsymbol{p}(\boldsymbol{h}^T, \boldsymbol{h}^S) + \boldsymbol{\mu}(\boldsymbol{h}^T)\boldsymbol{\nu}(\boldsymbol{h}^S)} \,, \quad (6)$$

which can be connected with MI via the following:

$$\log \boldsymbol{q}(\eta = 1 | \boldsymbol{h}^T, \boldsymbol{h}^S) \leq \log \frac{\boldsymbol{p}(\boldsymbol{h}^T, \boldsymbol{h}^S)}{\boldsymbol{\mu}(\boldsymbol{h}^T)\boldsymbol{\nu}(\boldsymbol{h}^S)} \,. \quad (7)$$

By taking the expectation of both sides w.r.t. the joint distribution $\boldsymbol{p}(\boldsymbol{h}^T, \boldsymbol{h}^S)$, we have:

$$I(\boldsymbol{h}^T, \boldsymbol{h}^S) \geq \mathbb{E}_{\boldsymbol{q}(\boldsymbol{h}^T, \boldsymbol{h}^S | \eta = 1)} [\log \boldsymbol{q}(\eta = 1 | \boldsymbol{h}^T, \boldsymbol{h}^S)] \,. \quad (8)$$

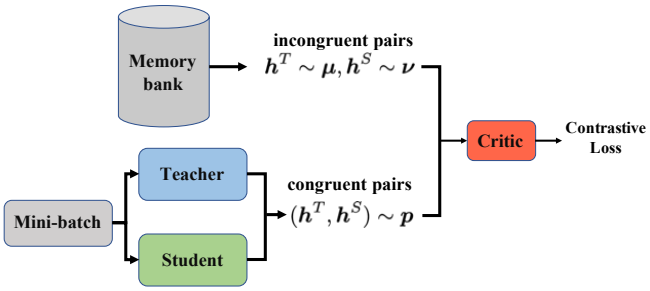


Figure 1: Illustration of Global Contrastive Knowledge Transfer (GCKT) via the use of the dual form for Wasserstein distance.

We can then maximize the right hand side of (8) to increase the lower bound of the MI. Though there is no closed form for $\boldsymbol{q}(\eta = 1 | \boldsymbol{h}^T, \boldsymbol{h}^S)$, a neural network $g$ (called a *critic* with parameters $\boldsymbol{\phi}$) can be used to estimate whether a pair comes from the joint distribution or the marginals. This shares a similar role as the discriminator of a Generative Adversarial Network (GAN) [16]. The critic $g$ can be learned via the following NCE loss:

$$\begin{aligned} \mathcal{L}_{\mathrm{NCE}} = & \mathbb{E}_{\boldsymbol{q}(\boldsymbol{h}^T, \boldsymbol{h}^S | \eta = 1)} [\log g(\boldsymbol{h}^T, \boldsymbol{h}^S)] \\ & + \mathbb{E}_{\boldsymbol{q}(\boldsymbol{h}^T, \boldsymbol{h}^S | \eta = 0)} [\log(1 - g(\boldsymbol{h}^T, \boldsymbol{h}^S))] \,. \end{aligned} \quad (9)$$

The parameters $\boldsymbol{\theta}_S$ and $\boldsymbol{\phi}$ can be optimized jointly by maximizing (9).

In previous work [42], the critic $g$ is a neural network that maps $(\boldsymbol{h}^T, \boldsymbol{h}^S)$ to $[0, 1]$ without other constraints. This can suffer from several drawbacks: $(i)$ $g$ could be sensitive to small numerical changes in the input [42, 33], yielding poor performance, especially when the network architectures or training datasets for the student and teacher networks are different. $(ii)$ It can suffer from mode collapse, as the support for $\boldsymbol{p}(\boldsymbol{h}^T, \boldsymbol{h}^S)$ and $\boldsymbol{\mu}(\boldsymbol{h}^T)\boldsymbol{\nu}(\boldsymbol{h}^S)$ may not overlap [2]. To alleviate these issues, we propose using the dual form of Wasserstein distance, by reformulating (9) as:

$$\begin{aligned} \mathcal{L}_{\mathrm{GCKT}}(\boldsymbol{\theta}_S, \boldsymbol{\phi}) = & \mathbb{E}_{\boldsymbol{q}(\boldsymbol{h}^T, \boldsymbol{h}^S | \eta = 1)} [\hat{g}(\boldsymbol{h}^T, \boldsymbol{h}^S)] \\ & - \mathbb{E}_{\boldsymbol{q}(\boldsymbol{h}^T, \boldsymbol{h}^S | \eta = 0)} [\hat{g}(\boldsymbol{h}^T, \boldsymbol{h}^S)] \,, \end{aligned} \quad (10)$$

where the new critic function $\hat{g}$ has to satisfy the 1-Lipschitz constraint. Equation (10) is otherwise similar to (9), which not only serves as a lower bound for the mutual information between the student and teacher representations, but also provides a robust critic to better match $\boldsymbol{p}(\boldsymbol{h}^T, \boldsymbol{h}^S)$ with $\boldsymbol{\mu}(\boldsymbol{h}^T)\boldsymbol{\nu}(\boldsymbol{h}^S)$.

Instead of enforcing 1-Lipschitz with the gradient penalty as in [17], we apply spectral normalization [31] to the critic $\hat{g}$. Specifically, spectral normalization on an arbitrary matrix $\mathbf{A}$ is defined as $\sigma(\mathbf{A}) = \max_{\|\boldsymbol{\beta}\|_2 \leq 1} \|\mathbf{A}\boldsymbol{\beta}\|_2$, which is equivalent to the largest singular value of $\mathbf{A}$. By applying this regularizer to the weights of each layer in $\hat{g}$, the 1-Lipschitz constraint can be enforced.

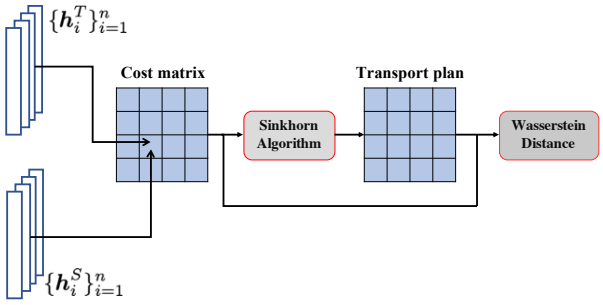


Figure 2: Illustration of Local Contrastive Knowledge Transfer (LCKT) via the use of the primal form for Wasserstein distance.

Note that when multiple incongurent pairs are chosen, the prior distribution on $\eta$ will also change, and (10) will be updated accordingly to:

$$\mathcal{L}_{\mathrm{GCKT}}(\boldsymbol{\theta}_S, \boldsymbol{\phi}) = \mathbb{E}_{\boldsymbol{q}(\boldsymbol{h}^T, \boldsymbol{h}^S|\eta=1)}[\hat{g}(\boldsymbol{h}^T, \boldsymbol{h}^S)] - M\mathbb{E}_{\boldsymbol{q}(\boldsymbol{h}^T, \boldsymbol{h}^S|\eta=0)}[\hat{g}(\boldsymbol{h}^T, \boldsymbol{h}^S)]\,, \tag{11}$$

where $M > 1$, and the lower bound for the mutual information will be tightened with large $M$ [18, 42]. In practice, the incongruent samples are drawn from a memory buffer, that stores pre-computed features of every data sample from previous mini-batches. In this way, we are able to efficiently retrieve a large number of negative samples without recalculating the features. Due to the use of data samples across multiple mini-batches for Wasserstein constrastive learning, we denote this method as *global* contrastive knowledge transfer (GCKT), as illustrated in Figure 1.

### 3.2. Local Contrastive Knowledge Transfer

Contrastive learning can also be applied within a mini-batch to further enhance performance. Specifically, in a mini-batch, the features $\{\boldsymbol{h}_i^T\}_{i=1}^n$ extracted from the teacher network can be viewed as a fixed set when training the student network. Ideally, categorical information is encapsulated in the feature space, so each element $\{\boldsymbol{h}_j^S\}_{j=1}^n$ from the student network should be able to find close neighbors in this set. For instance, nearby samples may share the same class. Therefore, we encourage the model to push $\boldsymbol{h}_j^S$ to several neighbors $\{\boldsymbol{h}_i^T\}_{i=1}^n$ instead of just one from the teacher network for better generalization. As the distribution matching happens in a mini-batch, we denote this as *local* contrastive knowledge transfer (LCKT).

This can be implemented efficiently with the primal form of Wasserstein distance. Specifically, the primal form can be interpreted as a less expensive way to transfer probability mass from $\boldsymbol{\mu}(\boldsymbol{h}^T)$ to $\boldsymbol{\nu}(\boldsymbol{h}^S)$, when only finite training samples are used. That is, we have $\boldsymbol{\mu}(\boldsymbol{h}^T) = \sum_{i=1}^n \mathrm{u}_i \delta_{\boldsymbol{h}_i^T}$, $\boldsymbol{\nu}(\boldsymbol{h}^S) = \sum_{j=1}^n \mathrm{v}_j \delta_{\boldsymbol{h}_j^S}$, where $\delta_{\boldsymbol{x}}$ is the Dirac function centered on $\boldsymbol{x}$. Since $\boldsymbol{\mu}(\boldsymbol{h}^T), \boldsymbol{\nu}(\boldsymbol{h}^S)$ are valid probability distributions, $\boldsymbol{u} = \{\mathrm{u}_i\}_{i=1}^n, \boldsymbol{v} = \{\mathrm{v}_j\}_{j=1}^n$ both lie on a simplex, *i.e.*, $\sum_{i=1}^n \mathrm{u}_i = 1$, and $\sum_{j=1}^n \mathrm{v}_j = 1$. Under this setting, the

**Algorithm 1** Sinkhorn Algorithm.

1: **Input:** $\{\boldsymbol{h}_i^T\}_{i=1}^n$,$\{\boldsymbol{h}_j^S\}_{j=1}^n$, $\epsilon$, probability vectors $\boldsymbol{\mu}$, $\boldsymbol{\nu}$
2: $\boldsymbol{\sigma} = \frac{1}{n}\mathbf{1_n}$, $\boldsymbol{\pi}^{(1)} = \mathbf{1}\mathbf{1}^\top$
3: $\mathbf{C}_{ij} = c(\boldsymbol{h}_i^T, \boldsymbol{h}_j^S)$, $\mathbf{A}_{ij} = \mathrm{e}^{-\frac{\mathbf{C}_{ij}}{\epsilon}}$
4: **for** $t = 1, 2, 3 \ldots$ **do**
5: $\quad \mathbf{Q} = \mathbf{A} \odot \boldsymbol{\pi}^{(t)}$ // $\odot$ is Hadamard product
6: $\quad$ **for** $k = 1, 2, 3, \ldots K$ **do**
7: $\quad\quad \boldsymbol{\delta} = \frac{\boldsymbol{\mu}}{n\mathbf{Q}\boldsymbol{\sigma}}, \boldsymbol{\sigma} = \frac{\boldsymbol{\nu}}{n\mathbf{Q}^\top\boldsymbol{\delta}}$
8: $\quad$ **end for**
9: $\quad \boldsymbol{\pi}^{(t+1)} = \mathrm{diag}(\boldsymbol{\delta})\mathbf{Q}\mathrm{diag}(\boldsymbol{\sigma})$
10: **end for**
11: $\mathcal{W} = \langle \boldsymbol{\pi}, \mathbf{C} \rangle$ // $\langle \cdot, \cdot \rangle$ is the Frobenius dot-product
12: Return $\mathcal{W}$

primal form can be reformulated into:

$$\mathcal{W}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{\pi}} \sum_{i=1}^n \sum_{j=1}^n \boldsymbol{\pi}_{ij} c(\boldsymbol{h}_i^T, \boldsymbol{h}_j^S) = \min_{\boldsymbol{\pi}} \langle \boldsymbol{\pi}, \mathbf{C} \rangle$$
$$\text{s.t.} \quad \sum_{j=1}^n \boldsymbol{\pi}_{ij} = \mathrm{u}_i, \quad \sum_{i=1}^n \boldsymbol{\pi}_{ij} = \mathrm{v}_j\,, \tag{12}$$

where $\boldsymbol{\pi}$ is the discrete joint probability in $\boldsymbol{h}^T$ and $\boldsymbol{h}^S$ (*i.e.*, the transport plan), $\mathbf{C}$ is the cost matrix given by $\mathbf{C}_{ij} = c(\boldsymbol{h}_i^T, \boldsymbol{h}_j^S)$, and $\langle \boldsymbol{\pi}, \mathbf{C} \rangle = \mathrm{Tr}(\boldsymbol{\pi}^\top \mathbf{C})$ represents the Frobenius dot-product. Expression $c(\cdot)$ is a cost function measuring the dissimilarity between the two feature vectors, where cosine distance $c(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{\boldsymbol{x}^\top \boldsymbol{y}}{||\boldsymbol{x}||_2 ||\boldsymbol{y}||_2}$ is a popular choice. Ideally, the global optimum for (12) may be obtained using linear programming [46, 35]. However, this method is not differentiable, making it incompatible with existing deep learning frameworks. As an alternative, the Sinkhorn algorithm [13] is applied to solve (12) by adding a convex regularization term, *i.e.,*

$$\mathcal{L}_{\mathrm{LCKT}}(\boldsymbol{\theta}_S) = \min_{\boldsymbol{\pi}} \sum_{i,j} \boldsymbol{\pi}_{ij} c(\boldsymbol{h}_i^T, \boldsymbol{h}_j^S) + \epsilon \mathcal{H}(\boldsymbol{\pi})\,, \tag{13}$$

where $\mathcal{H}(\boldsymbol{\pi}) = \sum_{i,j} \boldsymbol{\pi}_{ij} \log \boldsymbol{\pi}_{ij}$, and $\epsilon$ is the hyper-parameter controlling the importance of the entropy loss on $\boldsymbol{\pi}$. Detailed procedures for solving this is summarized in Algorithm 1. Although Lines 4-10 in Algorithm 1 constitute an iterative algorithm, its time complexity is small compared to the other feed-forward modules. Also, thanks to the Envelop Theorem [7], we can ignore the gradient flow through $\boldsymbol{\pi}$, meaning that there is no need to back-propagate gradients for Lines 4-10. In practice, we can simply detach/stop-gradient the for-loop module in Pytorch or Tensorflow, while the loss can still help refine the feature representations. Figure 2 illustrates the procedure for calculating the Wasserstein distance.

### 3.3. Unifying Global and Local Knowledge Transfer

Global knowledge transfer is designed for matching the joint distribution $\boldsymbol{p}(\boldsymbol{h}_T, \boldsymbol{h}_S)$ with the product of the marginal

distributions $\boldsymbol{\mu}(\boldsymbol{h}_T)\boldsymbol{\nu}(\boldsymbol{h}_S)$ via contrastive learning under a Wasserstein metric, achieving better distillation. At the same time, local knowledge transfer incentivizes matching the marginal distribution $\boldsymbol{\mu}(\boldsymbol{h}_T)$ with $\boldsymbol{\nu}(\boldsymbol{h}_S)$ via optimal transport, aiming for better generalization. Section 3.1 optimizes the MI by maximizing the lower bound, while Sec. 3.2 minimizes (13) to match the feature distributions.

Although GCKT and LCKT are designed for different objectives, they are complementary to each other. By optimizing LCKT, we aim to minimize the discrepancy between the marginal distributions, which is equivalent to reducing the difference between the two feature spaces, so that LCKT can provide a more constrained feature space for GCKT. On the other hand, by optimizing GCKT, the learned representation can also form a better feature space, which in turn helps LCKT match the marginal distributions.

In summary, the training objective for our method is written as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{WCoRD}}(\boldsymbol{\theta}_S, \boldsymbol{\phi}) = \mathcal{L}_{\mathrm{CE}}(\boldsymbol{\theta}_S) - \lambda_1 \mathcal{L}_{\mathrm{GCKT}}(\boldsymbol{\theta}_S, \boldsymbol{\phi}) \\ + \lambda_2 \mathcal{L}_{\mathrm{LCKT}}(\boldsymbol{\theta}_S)\,, \end{aligned} \tag{14}$$

where besides the parameters $\boldsymbol{\theta}_S$ of the student network, an additional set of parameters $\boldsymbol{\phi}$ for the critic is also learned.

## 4. Related Work

**Knowledge Distillation** Recent interest in knowledge distillation can be traced back to [23], though similar ideas have been proposed before [51, 6]. These methods use the probability distribution of the output over the prediction classes of a large teacher network as additional supervision signals to train a smaller student network. Recent studies have suggested alternative distillation objectives. Later works such as FitNet [37] extend the idea by using the intermediate layers instead of only $\boldsymbol{z}^T$. [50] proposed to use an attention map transfer in KD. SPKD [44] also utilizes intermediate features, but tries to mimic the representation space of the teacher features, rather than preserving pairwise similarities like FitNet. More recently, Contrastive Representation Distillation (CRD) [42] proposed applying NCE [18] to an intermediate layer. Another line of KD research explores alternatives to the teacher-student training paradigm. For example, [53] proposed an on-the-fly ensemble teacher network, in which the teacher is jointly trained with multiple students under a multi-branch network architecture, and the teacher's prediction is a weighted average of predictions from all the branches. Most recenlty, [49] shows that KD can be understood as label smoothing regularization.

**Optimal Transport** Optimal transport distance, *a.k.a.* Wasserstein distance, has a long history in mathematics, with applications ranging from resource allocation to computer vision [38]. Traditionally, optimal transport problems are solved by linear/quadratic programming. Within deep learning, the *dual* form of the Wasserstein distance is used by [2, 17] as an alternative metric for distribution matching in Generative Adversarial Network (GAN) [16], where the dual form is approximated by imposing a 1-Lipschitz constraint on the critic. The *primal* form of Wasserstein distance can be solved by the Sinkhorn algorithm [13], which has been applied to a wide range of deep learning tasks, including document retrieval and classification [28], sequence-to-sequence learning [10], adversarial attacks [47], graph matching [48], and cross-domain alignment [9]. To the authors' knowledge, this work is the first to apply optimal transport to KD, and to utilize both its primal and dual forms to construct a general Wasserstein learning framework.

**Contrastive Learning** Contrastive learning [18, 3] is a popular research area that has been successfully applied to density estimation and representation learning, especially in self-supervised setting [19, 11]. It has been shown that the contrastive objective can be interpreted as maximizing the lower bound of mutual information between different views of the data [24, 32, 4, 22], though it remains unclear whether the success is determined by mutual information or by the specific form of the contrastive loss [43]. Recently, contrastive learning has been extended to Wasserstein dependency measure [33]. Our global contrastive transfer shares similar ideas with it. However, its application to KD has not been studied before. Further, both the primal and dual forms are used to form an integral framework.

## 5. Experiments

We evaluate the proposed WCoRD framework on three knowledge distillation tasks: $(i)$ model compression of a large network, $(ii)$ cross-modal transfer, and $(iii)$ privileged information distillation.

### 5.1. Model Compression

**Experiments on CIFAR-100** CIFAR-100 [27] consists of 50K training images (0.5K images per class) and 10K test images. For fair comparison, we use the public CRD codebase [42] in our experiments. Two scenarios are considered: $(i)$ the student and the teacher share the same network architecture, and $(ii)$ different network architectures are used.

Table 1 and 3 present the top-1 accuracy from different distillation methods. In both tables, models using the original KD is a strong baseline, which only CRD and our WCoRD consistently outperform. The strong performance of the original KD method is manifested because distillation is performed between low-dimensional probability distributions from the teacher and student networks, which makes it relatively easy for the student to learn knowledge from the teacher. However, if knowledge transfer is applied to features from intermediate layers, the numerical scale of features can be different, even when both teacher and student

| **Teacher**<br>**Student** | **WRN-40-2**<br>**WRN-16-2** | **WRN-40-2**<br>**WRN-40-1** | **resnet56**<br>**resnet20** | **resnet110**<br>**resnet20** | **resnet110**<br>**resnet32** | **resnet32x4**<br>**resnet8x4** | **vgg13**<br>**vgg8** |
|---|---|---|---|---|---|---|---|
| Teacher | 75.61 | 75.61 | 72.34 | 74.31 | 74.31 | 79.42 | 74.64 |
| Student | 73.26 | 71.98 | 69.06 | 69.06 | 71.14 | 72.50 | 70.36 |
| KD | 74.92 | 73.54 | 70.66 | 70.67 | 73.08 | 73.33 | 72.98 |
| FitNet | 73.58 (↓) | 72.24 (↓) | 69.21 (↓) | 68.99 (↓) | 71.06 (↓) | 73.50 (↑) | 71.02 (↓) |
| AT | 74.08 (↓) | 72.77 (↓) | 70.55 (↓) | 70.22 (↓) | 72.31 (↓) | 73.44 (↑) | 71.43 (↓) |
| SP | 73.83 (↓) | 72.43 (↓) | 69.67 (↓) | 70.04 (↓) | 72.69 (↓) | 72.94 (↓) | 72.68 (↓) |
| CC | 73.56 (↓) | 72.21 (↓) | 69.63 (↓) | 69.48 (↓) | 71.48 (↓) | 72.97 (↓) | 70.71 (↓) |
| VID | 74.11 (↓) | 73.30 (↓) | 70.38 (↓) | 70.16 (↓) | 72.61 (↓) | 73.09 (↓) | 71.23 (↓) |
| RKD | 73.35 (↓) | 72.22 (↓) | 69.61 (↓) | 69.25 (↓) | 71.82 (↓) | 71.90 (↓) | 71.48 (↓) |
| PKT | 74.54 (↓) | 73.45 (↓) | 70.34 (↓) | 70.25 (↓) | 72.61 (↓) | 73.64 (↑) | 72.88 (↓) |
| AB | 72.50 (↓) | 72.38 (↓) | 69.47 (↓) | 69.53 (↓) | 70.98 (↓) | 73.17 (↓) | 70.94 (↓) |
| FT | 73.25 (↓) | 71.59 (↓) | 69.84 (↓) | 70.22 (↓) | 72.37 (↓) | 72.86 (↓) | 70.58 (↓) |
| FSP | 72.91 (↓) | n/a | 69.95 (↓) | 70.11 (↓) | 71.89 (↓) | 72.62 (↓) | 70.23 (↓) |
| NST | 73.68 (↓) | 72.24 (↓) | 69.60 (↓) | 69.53 (↓) | 71.96 (↓) | 73.30 (↓) | 71.53 (↓) |
| CRD | 75.48 (↑) | 74.14 (↑) | 71.16 (↑) | 71.46 (↑) | 73.48 (↑) | 75.51 (↑) | 73.94 (↑) |
| CRD+KD | 75.64 (↑) | 74.38 (↑) | 71.63 (↑) | 71.56 (↑) | 73.75 (↑) | 75.46 (↑) | 74.29 (↑) |
| LCKT | 75.22 (↑) | 74.11 (↑) | 71.14 (↑) | 71.23 (↑) | 72.32 (↑) | 74.65 (↑) | 73.50 (↑) |
| GCKT | 75.47 (↑) | 74.23 (↑) | 71.21 (↑) | 71.43 (↑) | 73.41 (↑) | 75.45 (↑) | 74.10 (↑) |
| WCoRD | 75.88 (↑) | **74.73** (↑) | 71.56 (↑) | 71.57 (↑) | 73.81 (↑) | 75.95 (↑) | 74.55 (↑) |
| WCoRD+KD | **76.11** (↑) | 74.72 (↑) | **71.92** (↑) | **71.88** (↑) | **74.20** (↑) | **76.15** (↑) | **74.72** (↑) |

Table 1: CIFAR-100 test *accuracy* (%) of student networks trained with a number of distillation methods, when sharing the same architecture type as the teacher. See Appendix for citations of the compared methods. ↑ denotes outperformance over KD, and ↓ denotes underperformance. For all other methods, we used author-provided or author-verified code from the CRD repository. Our reported results are averaged over 5 runs. Note that $\lambda_1 = 0.8$ is the same as the weight on CRD, and $\lambda_2 = 0.05$.

| | Teacher | Student | AT | KD | SP | CC | CRD | CRD+KD | LCKT | GCKT | WCoRD | WCoRD+KD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Top-1 | 26.69 | 30.25 | 29.30 | 29.34 | 29.38 | 30.04 | 28.83 | 28.62 | 29.10 | 28.78 | 28.51 | **28.44** |
| Top-5 | 8.58 | 10.93 | 10.00 | 10.12 | 10.20 | 10.83 | 9.87 | 9.51 | 10.05 | 9.92 | 9.84 | **9.45** |

Table 2: Top-1 and Top-5 error rates (%) of student network ResNet-18 on ImageNet validation set.

share the same network architecture. As shown in Table 3, directly applying similarity matching to align teacher and student features even hurts performance.

WCoRD is a unified framework bridging GCKT and LCKT, which improves the performance of CRD, a current state-of-the-art model. When the same network architecture is used for both the teacher and student networks, an average relative improvement[1] of $48.63\%$ is achieved (derived from Table 1). This performance lift is $43.27\%$ when different network architectures are used (derived from Table 3).

We can also add the basic KD loss to WCoRD, obtaining an ensemble distillation loss (denoted as WCoRD+KD), similar to [42]. In most cases, this ensemble loss can further improve the performance. However, in ResNet50 → MobileNetV2 and ResNet50 → VGG8, WCoRD still works better than WCoRD+KD.

[1]The relative improvement is defined as $\frac{\text{WCoRD}-\text{CRD}}{\text{CRD}-\text{KD}}$, where the name of each method represents the corresponding accuracy of the student model.

**Ablation Study** We report results using only global or local contrastive knowledge transfer in Table 1 and 3. LCKT performs better than KD but slightly worse than CRD and GCKT, as both CRD and GCKT are NCE-based algorithms, where negative samples are used to improve performance. Additionally, GCKT enforces an 1-Lipschitz constraint on the critic function, which includes an extra hyper-parameter. Results show that CRD and GCKT have comparable results, and in some cases, GCKT performs slightly better (*e.g.*, from VGG13 to VGG8).

We perform an additional ablation study on the weight of the $\mathcal{L}_{\text{LCKT}}$ loss term (*i.e.*, $\lambda_2$ in (14)). We adjust $\lambda_2$ from 0 to 0.2, and set $\lambda_1 = 0.8$, which is the same as in CRD for fair comparison. Results are summarized in Table 4. The standard deviation (Std) is reported based on 5 runs. We observe that: $(i)$ when $\lambda_2 = 0.05$, ResNet-8x4 performs the best; and $(ii)$ WCoRD can consistently outperform GCKT and CRD methods, when $\lambda_2 \in (0, 0.2]$.

| **Teacher**<br>**Student** | **vgg13**<br>**MobileNetV2** | **ResNet50**<br>**MobileNetV2** | **ResNet50**<br>**vgg8** | **resnet32x4**<br>**ShuffleNetV1** | **resnet32x4**<br>**ShuffleNetV2** | **WRN-40-2**<br>**ShuffleNetV1** |
|---|---|---|---|---|---|---|
| Teacher | 74.64 | 79.34 | 79.34 | 79.42 | 79.42 | 75.61 |
| Student | 64.6 | 64.6 | 70.36 | 70.5 | 71.82 | 70.5 |
| KD | 67.37 | 67.35 | 73.81 | 74.07 | 74.45 | 74.83 |
| FitNet | 64.14 (↓) | 63.16 (↓) | 70.69 (↓) | 73.59 (↓) | 73.54 (↓) | 73.73 (↓) |
| AT | 59.40 (↓) | 58.58 (↓) | 71.84 (↓) | 71.73 (↓) | 72.73 (↓) | 73.32 (↓) |
| SP | 66.30 (↓) | 68.08 (↑) | 73.34 (↓) | 73.48 (↓) | 74.56 (↑) | 74.52 (↓) |
| CC | 64.86 (↓) | 65.43 (↓) | 70.25 (↓) | 71.14 (↓) | 71.29 (↓) | 71.38 (↓) |
| VID | 65.56 (↓) | 67.57 (↑) | 70.30 (↓) | 73.38 (↓) | 73.40 (↓) | 73.61 (↓) |
| RKD | 64.52 (↓) | 64.43 (↓) | 71.50 (↓) | 72.28 (↓) | 73.21 (↓) | 72.21 (↓) |
| PKT | 67.13 (↓) | 66.52 (↓) | 73.01 (↓) | 74.10 (↑) | 74.69 (↑) | 73.89 (↓) |
| AB | 66.06 (↓) | 67.20 (↓) | 70.65 (↓) | 73.55 (↓) | 74.31 (↓) | 73.34 (↓) |
| FT | 61.78 (↓) | 60.99 (↓) | 70.29 (↓) | 71.75 (↓) | 72.50 (↓) | 72.03 (↓) |
| NST | 58.16 (↓) | 64.96 (↓) | 71.28 (↓) | 74.12 (↑) | 74.68 (↑) | 74.89 (↑) |
| CRD | 69.73 (↑) | 69.11 (↑) | 74.30 (↑) | 75.11 (↑) | 75.65 (↑) | 76.05 (↑) |
| CRD+KD | 69.94 (↑) | 69.54 (↑) | 74.58 (↑) | 75.12 (↑) | 76.05 (↑) | 76.27 (↑) |
| LCKT | 68.21 (↑) | 68.81 (↑) | 73.21 (↑) | 74.62 (↑) | 74.70 (↑) | 75.08 (↑) |
| GCKT | 68.78 (↑) | 69.20 (↑) | 74.29 (↑) | 75.18 (↑) | 75.78 (↑) | 76.13 (↑) |
| WCoRD | 69.47 (↑) | **70.45** (↑) | **74.86** (↑) | 75.40 (↑) | 75.96 (↑) | 76.32 (↑) |
| WCoRD+KD | **70.02** (↑) | 70.12 (↑) | 74.68 (↑) | **75.77** (↑) | **76.48** (↑) | **76.68** (↑) |

Table 3: CIFAR-100 test *accuracy* (%) of a student network trained with a number of distillation methods, when the teacher network architecture is significantly different. We use the same codebase from the CRD repository. Our reported results are averaged over 5 runs. Note that $\lambda_1 = 0.8$ is the same as the weight on CRD, and $\lambda_2 = 0.05$.

| $\lambda_2$ | 0 | 0.01 | 0.03 | 0.05 | 0.08 | 0.1 | 0.2 |
|---|---|---|---|---|---|---|---|
| Mean | 75.45 | 75.66 | 75.75 | **75.95** | 75.83 | 75.66 | 75.62 |
| Std | 0.31 | 0.46 | 0.44 | 0.40 | 0.34 | 0.29 | 0.47 |

Table 4: CIFAR-100 test accuracy (%) of student network ResNet-8x4 with different weights on the local knowledge transfer term. The teacher network is ResNet-32x4.

| Layer | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| CRD | **55.0** | 63.64 | 73.76 | 74.75 |
| WCoRD | 54.6 | **63.70** | **74.23** | **75.43** |

Table 5: Top-1 Accuracy (%) on chrominance view of STL-10 testing set with ResNet-18. The modal is distilled on the network trained with the luminance view of Tiny-ImageNet.

**Experiments on ImageNet** We also evaluate the proposed method on a larger dataset, ImageNet [14], which contains 1.2M images for training and 50K for validation. In this experiment, we use ResNet-34 [20] as the teacher and ResNet-18 as the student, and use the same training setup as in CRD [42] for fair comparison. Top-1 and Top-5 error rates (lower is better) are reported in Table 2, showing that the WCoRD+KD method achieves the best student performance on the ImageNet dataset. The relative improvement of WCoRD over CRD on Top-1 error is 44.4%, and the relative improvement on Top-5 error is 23.08%, which further demonstrates the scalability of our method.

### 5.2. Cross-Modal Transfer

We consider a setting where one modality $\mathcal{X}$ contains a large amount of labeled data, while the other modality $\mathcal{Y}$ has only a small labeled dataset. Transferring knowledge from $\mathcal{X}$ to $\mathcal{Y}$ should help improve the performance of the model trained on $\mathcal{Y}$. We use linear probing to evaluate the quality of the learned representation, a common practice proposed in [1, 52]. Following [42], the same architecture is applied to both teacher and student networks. We first map images in the RGB space to the Lab color space (L: Luminance, ab: Chrominance), then train a ResNet18 (teacher) on the Luminance dimension of Labeled Tiny-ImageNet [14], which we call L-Net. The accuracy of L-Net on Tiny-ImageNet is 47.76%. The student network, denoted as ab-Net, is trained on the Chrominance dimension of unlabeled STL-10 dataset [12].

In experiments, we distill general knowledge from L-Net to ab-Net with different objective functions such as CRD and WCoRD. Linear probing is performed by fixing the ab-Net. We then train a linear classification module on top of features extracted from different layers in the ResNet18 for 10-category classification. Results are summarized in Table 5. For reference, training student model from scratch with ResNet18 on STL-10 can reach an accuracy of 64.7%. From Table 5, WCoRD outperforms CRD when using fea-

| | Parameter Size | Teacher | Student | AT | SP | CC | CRD | LCKT | WCoRD |
|---|---|---|---|---|---|---|---|---|---|
| ResNet-8x4 | 4.61 Mb | - | 82.16 | 82.43 | 82.45 | 80.97 | 83.43 | 81.30 | **84.50** |
| Inception-v3 | 84.61 Mb | - | 79.12 | 80.75 | 80.81 | 79.01 | 79.68 | 80.74 | **80.85** |

Table 6: AUC (%) of the ResNet-8x4 and Inception-v3 student networks on the OCT-GA dataset.

tures extracted from the 2nd-4th residual blocks, indicating features extracted from these layers via WCoRD is more informative than those from CRD.

### 5.3. Privileged Information Distillation

In many real-world scenarios, large datasets required to train deep network models cannot be released publicly, due to privacy or security concerns. One solution is privileged information distillation [45]: instead of releasing the original data, a model trained on the data is made public, which can serve as a teacher model. Other researchers can train a student model on a smaller public dataset, leveraging this teacher model as additional supervision to enhance the performance of the student model.

For example, Geographic Atrophy (GA) is an advanced, vision-threatening form of age-related macular degeneration (AMD), currently affecting a significantly large number of individuals. Optical coherence tomography (OCT) imaging is a popular method for diagnosing and treating many eye diseases, including GA [5]. To automatically detect GA in OCT images, a binary classifier can be trained with labeled data such as OCT-GA, an institutional dataset consisting of 44,520 optical coherence tomography (OCT) images of the retina of 1088 patients; 9640 of these images exhibit GA, and each image contains 512 by 1000 pixels.

The resolution of images in OCT-GA is relatively low, and the small size of the dataset puts limitations on the learned model. One way to improve model performance is by leveraging additional larger high-resolution datasets [39]. Two challenges prevent us from doing this in real-life scenarios: $(i)$ additional datasets may be private, and only a pre-trained model is publicly available; and $(ii)$ the disease of interest may not be among those labeled categories in the additional datasets (*e.g.*, GA may not be the focus of interest in other imaging datasets).

One example is the OCT dataset introduced by [25], consisting of 108,312 images from 4,686 patients for 4-way classification: choroidal neovascularization (CNV), diabetic macular edema (DME), Drusen, and Normal. To test our proposed framework in privileged distillation setting, we treat this larger dataset as inaccessible and only use a pre-trained model as the teacher, as is the case in many real-life scenarios. Then we train a model on the smaller OCT-GA dataset as the student network, and use privileged distillation to transfer knowledge from the teacher to the student.

We test both Inception-v3 and ResNet8x4 models as student networks for GA disease identification. KL divergence cannot be used here, as both the learning goal and the training datasets for teacher and student networks are different. This is designed to test the knowledge generalization ability of a model. As shown in Table 6, WCoRD achieves an improvement of $2.34\%$ and $0.96\%$ compared to the basic student and CRD methods, respectively. The relative improvement with ResNet8x4 is $\frac{\text{WCoRD}-\text{CRD}}{\text{CRD}-\text{AT}} = 107.0\%$. Since the goal of the teacher and student models are different, features from the teacher are biased. When the student uses the same architecture as the teacher (Inception-v3), CRD performs worse than both AT and SPKD in Table 6, which can be interpreted as low generalizability. With the help from LCKT, WCoRD is still able to get a comparable accuracy. These results serve as strong evidence that WCoRD possesses better knowledge generalization ability than CRD.

| $\lambda_2$ | 0 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 | 0.1 | 0.2 |
|---|---|---|---|---|---|---|---|---|
| Mean | 83.43 | 83.80 | 84.15 | **84.50** | 83.83 | 83.70 | 83.65 | 82.28 |
| Std | 0.48 | 0.71 | 0.69 | 0.91 | 0.91 | 0.66 | 0.49 | 0.50 |

Table 7: AUC (%) of student network ResNet-8x4 with different weights on the local knowledge transfer term.

**Ablation Study** We investigate the importance of the local knowledge transfer term $\mathcal{L}_{\text{LCKT}}(\cdot)$ in WCoRD. As shown in Tables 1 and 3, without it, WCoRD cannot consistently outperform CRD in different student-teacher architecture settings. By fixing $\lambda_1$ for the $\mathcal{L}_{\text{GCKT}}(\cdot)$ loss with $\lambda_1 = 1$, we adjust the hyper-parameter $\lambda_2$ from $0.01$ to $0.2$. Table 7 reports the results, where it is evident that with $\lambda_2 = 0.07$ WCoRD performs the best. Also, we observe that when choosing $\lambda_2$ from $(0, 0.1]$, it is consistently better than the model variant with $\lambda_2 = 0$. This indicates that our model is relatively robust given different hyper-parameter choices.

## 6. Conclusions

We present Wasserstein Contrastive Representation Distillation (WCoRD), a new framework for knowledge distillation. WCoRD generalizes the concept of contrastive learning via the use of Wasserstein metric, and introduces an additional feature distribution matching term to further enhance the performance. Experiments on a variety of tasks show that our new framework consistently improves the student model performance. For future work, we plan to further extend our framework to other applications, such as federated learning [26] and adversarial robustness [34].

## References


[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *ICLR*, 2017. 7

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 3, 5

[3] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019. 5

[4] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *NeurIPS*, 2019. 5

[5] David S Boyer, Ursula Schmidt-Erfurth, Menno van Lookeren Campagne, Erin C Henry, and Christopher Brittain. The pathophysiology of geographic atrophy secondary to age-related macular degeneration and the complement pathway as a therapeutic target. *Retina (Philadelphia, Pa.)*, 37(5):819, 2017. 8

[6] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *SIGKDD*, 2006. 5

[7] Michael Carter. *Foundations of mathematical economics*. MIT Press, 2001. 4

[8] Liqun Chen, Shuyang Dai, Yunchen Pu, Chunyuan Li, Qinliang Su, and Lawrence Carin. Symmetric variational autoencoder and connections to adversarial learning. In *AISTATS*, 2018. 2

[9] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. *arXiv preprint arXiv:2006.14744*, 2020. 5

[10] Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, and Lawrence Carin. Improving sequence-to-sequence learning via optimal transport. *ICLR*, 2019. 5

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 5

[12] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. *AISTATS*, 2011. 7

[13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013. 4, 5

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 7

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2018. 1

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 3, 5

[17] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *NeurIPS*, 2017. 3, 5

[18] Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *AISTATS*, 2010. 2, 3, 4, 5

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019. 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016. 1

[22] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019. 5

[23] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *NeurIPS Deep Learning and Representation Learning Workshop*, 2015. 1, 2, 5

[24] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *ICLR*, 2019. 5

[25] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 2018. 8

[26] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016. 8

[27] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[28] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *ICML*, 2015. 5

[29] Kevin J Liang, Geert Heilmann, Christopher Gregory, Souleymane O Diallo, David Carlson, Gregory P Spell, John B Sigman, Kris Roe, and Lawrence Carin. Automatic threat recognition of prohibited items at aviation checkpoint with x-ray imaging: a deep learning approach. *SPIE Anomaly Detection and Imaging with X-Rays (ADIX) III*, 2018. 1

[30] David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. *AISTATS*, 2020. 2

[31] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018. 2, 3

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[33] Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron van den Oord, Sergey Levine, and Pierre Sermanet. Wasserstein dependency measure for representation learning. *NeurIPS*, 2019. 2, 3, 5

[34] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy (SP)*, 2016. 8

[35] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. Technical report, 2017. 4

[36] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. Detecting and classifying lesions in mammograms with deep learning. *Scientific reports*, 2018. 1

[37] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *ICLR*, 2014. 1, 5

[38] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. A metric for distributions with applications to image databases. In *ICCV*, 1998. 5

[39] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. *ICCV*, 2017. 1, 8

[40] Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*, 2019. 1

[41] Siqi Sun, Zhe Gan, Yu Cheng, Yuwei Fang, Shuohang Wang, and Jingjing Liu. Contrastive distillation on intermediate representations for language model compression. *arXiv preprint arXiv:2009.14167*, 2020. 1

[42] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *ICLR*, 2020. 1, 2, 3, 4, 5, 6, 7

[43] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019. 5

[44] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *ICCV*, 2019. 5

[45] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 2009. 8

[46] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008. 2, 4

[47] Eric Wong, Frank R Schmidt, and J Zico Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. *ICML*, 2019. 5

[48] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin. Gromov-wasserstein learning for graph matching and node embedding. In *NeurIPS*, 2019. 5

[49] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *CVPR*, 2020. 5

[50] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 1, 5

[51] Xinchuan Zeng and Tony R. Martinez. Using a neural network to approximate an ensemble of classifiers. *Neural Processing Letters*, 2000. 5

[52] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *CVPR*, 2017. 7

[53] Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-the-fly native ensemble. In *NeurIPS*, 2018. 5