Learning to Rematch Mismatched Pairs for Robust Cross-Modal Retrieval

Haochen Han[1*], Qinghua Zheng[1], Guang Dai[2], Minnan Luo[1†], Jingdong Wang[3]
[1] Xi'an Jiaotong University [2] SGIT AI Lab, State Grid Corporation of China [3] Baidu Inc
hhc1997@stu.xjtu.edu.cn, {qhzheng, minnluo}@xjtu.edu.cn
guang.gdai@gmail.com, wangjingdong@baidu.com

Abstract

Collecting well-matched multimedia datasets is crucial for training cross-modal retrieval models. However, in real-world scenarios, massive multimodal data are harvested from the Internet, which inevitably contains Partially Mismatched Pairs (PMPs). Undoubtedly, such semantical irrelevant data will remarkably harm the cross-modal retrieval performance. Previous efforts tend to mitigate this problem by estimating a soft correspondence to down-weight the contribution of PMPs. In this paper, we aim to address this challenge from a new perspective: the potential semantic similarity among unpaired samples makes it possible to excavate useful knowledge from mismatched pairs. To achieve this, we propose L2RM, a general framework based on Optimal Transport (OT) that learns to rematch mismatched pairs. In detail, L2RM aims to generate refined alignments by seeking a minimal-cost transport plan across different modalities. To formalize the rematching idea in OT, first, we propose a self-supervised cost function that automatically learns from explicit similarity-cost mapping relation. Second, we present to model a partial OT problem while restricting the transport among false positives to further boost refined alignments. Extensive experiments on three benchmarks demonstrate our L2RM significantly improves the robustness against PMPs for existing models. The code is available at https://github.com/hhc1997/L2RM.

1. Introduction

The pursuit of general intelligence has advanced the progress of multimodal learning, which aims to understand and integrate multiple sensory modalities like humans. Cross-modal retrieval is one of the most important techniques in multimodal learning due to its flexibility in bridging different modalities [15, 20, 22, 25, 37, 42], which has powered various real-world applications.

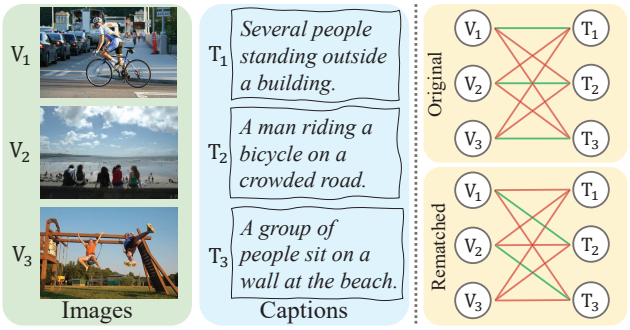


Figure 1. A toy example to illustrate our idea. The potential semantic similarity among unpaired samples makes it possible to excavate useful knowledge from mismatched pairs. Our L2RM aims to rematch PMPs by generating a refined alignment that brings relevant cross-modal samples (green links) together while repelling irrelevant ones (red links) away from each other. We also show some real-world rematched cases for our L2RM in Fig. 5.

Despite the remarkable performance of previous methods, much of their success can be attributed to the voracious appetite for well-matched cross-modal pairs. In practice, collecting such ideal data [24] is notoriously labor-intensive and even impossible. Alternatively, several mainstream cross-modal datasets utilize the co-occurred information to crawl data from the Internet, especially for visual-text samples [11]. Although such a data collection way is free from expensive annotations, it will inevitably introduce partially mismatched pairs. For example, the standard image-caption dataset, Conceptual Captions [36], is estimated to contain about 3% to 20% mismatched pairs. Such semantically irrelevant data will be wrongly treated as the matched pairs for training, which undoubtedly impairs the performance of cross-modal retrieval models. Thus, endowing cross-modal learning with robustness against PMPs is crucial to suit real-world retrieval scenarios.

To alleviate the PMP problem, existing works [19, 23, 45]

*This work was completed during his internship at SGIT AI Lab.

†Corresponding author. This work was supported by the National Key Research and Development Program of China (2022YFB3102600), and National Nature Science Foundation of China (62272374, 62192781, 62202367, 62250009, and 62137002).

typically resort to recasting the estimated soft correspondence into a soft margin to adjust the distance in triplet ranking loss. However, the underuse of mismatched pairs, only limited to down-weighting their contribution, has led to sub-optimal retrieval performance. Hence, it is necessary to address the PMP issue in a data-efficient manner.

A question naturally arises: *Could cross-modal retrieval models even learn useful knowledge from mismatched pairs?* To answer this question, this paper presents L2RM, a general framework that learns to rematch mismatched pairs for robust cross-modal retrieval. As illustrated in Fig. 1, our key idea is to excavate the potential matching relationship among mismatched cross-modal samples. Specifically, we first identify possibly mismatched pairs from training data by modeling the per-sample loss distribution. Then, we formalize the rematching idea as an OT problem to generate a new set of refined alignments for mismatched pairs in every minibatch. Notably, the cost function plays a paramount role when applying OT, which is typically designed as feature-driven distance [5, 10, 18]. However, the over-dependence on representations has led to a cycle of self-reinforcing errors—the existence of PMPs can generate corrupted representations–in turn, preventing the effective transport plan. To handle this problem, we propose a self-supervised learning solution to automatically learn the cost function from explicit similarity-cost mapping relation, which is unexplored in previous OT literature. Moreover, instead of exactly rematching all mismatched samples, we suggest modeling a partial OT problem while restricting the transport among false positives to boost the refined alignment. In practice, we show that our optimization objective could be solved by the Sinkhorn algorithm [9], which only incurs cheap computational overheads.

Our main contributions are summarized as follows: (1) We propose a general OT-based framework to address the widely-existed PMP problem in cross-modal retrieval. The key to our method is learning to rematch mismatched pairs, which goes beyond previous efforts from the data-efficient view. (2) To address the error accumulation faced by the vanilla cost function, we propose a novel self-supervised learner that automatically learns the transport cost from explicit similarity-cost mapping relation. (3) To further boost the refined alignment, we present to model a partial OT problem and restrict the transport among false positives. (4) Extensive experiments on several benchmarks demonstrate our L2RM endows existing cross-modal retrieval methods with strong robustness against PMPs.

2. Related Work

Cross-Modal Retrieval. Approaches for cross-modal retrieval aim to retrieve relevant items across different modalities for the query data. Current dominant methods project different modalities into a shared embedding space to measure the similarity of cross-modal pairs, which generally follow two research lines: 1) Global Alignment focuses on learning the correspondence between whole cross-modal data. Existing studies usually propose a two-stream network to learn comparable global features [13, 30, 47]. 2) Local Alignment. It seeks to align the fine-grained regions for more precise cross-modal matching. For example, [26] employ the cross-attention mechanism to fully excavate the semantic region-word alignments. [43,44] explore the intra-modal relation to facilitate inter-modal alignments.

Although these prior arts have achieved promising results, their success mainly relies on well-matched data, which is extremely expensive and even impossible to collect. To satisfy a more practical retrieval that is robust against the PMPs, [23, 33, 45] divide the mismatched pairs from training data and estimate a soft correspondence to downweight their training contribution. Recently, [21] resorts to complementary contrastive learning that only utilizes the negative information to avoid overfitting. However, these methods neglect the usage of either the negative information [23, 33, 45] or the positive one [21]. To fully leverage the training data, this paper proposes an OT-based method to rematch those partially mismatched pairs.

Optimal Transport. OT is used to seek a minimal-cost transport plan from one probability measure to another. The original OT model is a linear program that incurs expensive computational cost. [9] proposes the entropy-regularized OT to provide a computationally cheaper solver. Recently, OT has gained increasing attention from different fields in machine learning, including unsupervised learning [4, 29], semi-supervised learning [39], object detection [1, 17], domain adaptation [14, 35], and long-tailed recognition [32, 40]. To the best of our knowledge, we are the first to perform the PMP problem from an OT perspective.

3. Preliminaries

3.1. Background on OT

OT provides a mechanism to infer the correspondence between two measures. We briefly introduce the OT theory to help us better view the PMP problem from an OT perspective. Consider $\boldsymbol{X} = \{x_i\}_{i=1}^m$ and $\boldsymbol{Y} = \{y_j\}_{j=1}^n$ as two discrete variables, and we denote their probability measures as $\boldsymbol{p} = \sum_{i=1}^m p_i \delta(x_i)$ and $\boldsymbol{q} = \sum_{j=1}^n q_j \delta(y_j)$, where δ is the Dirac function, p_i and q_j are the probability mass belonging to the probability simplex. When a meaningful cost function $c(\cdot)$ is defined, we can get the cost matrix $\boldsymbol{C} \in \mathbb{R}^{m \times n}$ between $\boldsymbol{X}$ and $\boldsymbol{Y}$, where $\boldsymbol{C}_{ij} = c(x_i, y_j)$. Based on these, the OT distance can be expressed as:

$$\begin{aligned} \mathrm{OT}(\boldsymbol{p}, \boldsymbol{q}) &\triangleq \min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{p}, \boldsymbol{q})} \langle \boldsymbol{\pi}, \boldsymbol{C} \rangle_F \\ \text{s.t. } \Pi(\boldsymbol{p}, \boldsymbol{q}) &= \{\boldsymbol{\pi} \in \mathbb{R}_+^{m \times n} | \boldsymbol{\pi} \mathbb{1}_n = \boldsymbol{p}, \boldsymbol{\pi}^\top \mathbb{1}_m = \boldsymbol{q}\}, \end{aligned} \tag{1}$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius dot-product and $\mathbb{1}_d$ denotes a d-dimensional all-one vector. $\boldsymbol{\pi}$ is called the optimal transport plan that transport $\boldsymbol{p}$ towards $\boldsymbol{q}$ at the smallest cost.

3.2. Problem Definition

Without losing generality, we take the visual-text retrieval as an example to present the PMP problem in cross-modal retrieval. Consider a training dataset $\mathbb{D} = \{(V_i, T_i, m_i)\}_{i=1}^N$ consisting of N samples, where (V_i, T_i) is the visual-text pair and $m_i \in \{1, 0\}$ indicates whether the bimodal data is semantically matched or not. The key to cross-modal retrieval lies in measuring the similarity across distinct modalities. To achieve this, existing methods usually project the visual and textual modalities into a comparable feature space via the corresponding modal-specific networks f_v and f_t, respectively. Then the similarity of a given visual-text pair is measured through $S_{ij} = g(f_v(V_i), f_t(T_j))$, where g is a nonparametric or parametric mapping function. For convenience, we denote $g(f_v(V_i), f_t(T_j))$ as $g(V_i, T_j)$ in the following.

Ideally, the positive (matched) pairs should have higher similarity while the negative (mismatched) pairs should have lower ones, which can be achieved by minimizing the triplet loss [13] or InfoNCE loss [31]. Consider a batch of N_b pairs $\{(V_i, T_i)\}_{i=1}^{N_b}$, the triplet loss is defined as:

$$\begin{aligned}\mathcal{L}^{\text{triplet}}(V_i, T_i) =& [\alpha - g(V_i, T_i) + g(V_i, \hat{T}_h)]_+ \\ &+ [\alpha - g(V_i, T_i) + g(\hat{V}_h, T_i)]_+,\end{aligned} \tag{2}$$

where α is a margin and $[x]_+ = max(x, 0)$. $\hat{V}_h$ and $\hat{T}_h$ are the most similar negatives in the given batch corresponding to (V_i, T_i). Eq.(2) aims to enforce the negative pairs to be distant from the positives by a certain margin value.

Alternatively, InfoNCE loss is extended to cross-modal scenario [21, 34] that encourages the similarity gap between positives and negatives as large as possible. Formally, the matching probability of j-th textual sample w.r.t. the i-th visual query is defined as $p_{ij}^{v2t} = \frac{\exp(g(V_i, T_j)/\tau)}{\sum_{j'=1}^{N_b} \exp(g(V_i, T_{j'})/\tau)}$, where τ is a temperature parameter. As InfoNCE loss is symmetric, the matching probability p_{ij}^{t2v} is defined similarly. For notation convenience, we denote $\boldsymbol{p}_i^{v2t} = [p_{i1}^{v2t}, \cdots, p_{iN_b}^{v2t}]^\top$ and $\boldsymbol{p}_i^{t2v} = [p_{i1}^{t2v}, \cdots, p_{iN_b}^{t2v}]^\top$ as the probability vectors. To align cross-modal samples, the corresponding one-hot vector $\boldsymbol{y}_i = [y_{i1}, \cdots, y_{iN_b}]^\top$ is used as supervision, where y_{ij} equal to 1 if $i = j$ while other elements are 0. Thus, the cross-modal InfoNCE loss is given by:

$$\mathcal{L}^{\text{InfoNCE}}(V_i, T_i) = \mathcal{H}(\boldsymbol{y}_i, \boldsymbol{p}_i^{v2t}) + \mathcal{H}(\boldsymbol{y}_i, \boldsymbol{p}_i^{t2v}), \tag{3}$$

where $\mathcal{H}$ is the batched cross-entropy function.

The success of both Eq.(2) and Eq.(3) relies on the well-matched pairs. However, in practice, the multimedia datasets are usually web-collected, and thus inevitably contains an unknown portion of irrelevant pairs but are wrongly treated as matched $(m_i = 1)$. Our goal is to combat such PMPs to facilitate robust cross-modal retrieval.

4. Methodology

To tackle the PMP problem, the mainstream pipeline first uses the memorization effect [3] of DNNs, *i.e.*, DNNs learn simpler patterns before memorizing the difficult ones, to partition the dataset into a matched subset $\mathbb{D}_m$, and a mismatched subset $\mathbb{D}_{\widetilde{m}} = \mathbb{D} / \mathbb{D}_m$. After that, $\mathbb{D}_m$ can be used for standard cross-modal training. To mitigate the impact of PMPs, recent advances [19, 23, 45] introduce a soft margin into Eq.(2) to down-weight the samples from $\mathbb{D}_{\widetilde{m}}$. However, due to the underuse of mismatched pairs, the achieved performance by them is argued to be sub-optimal. In this work, we aim to fully leverage PMPs by trying to excavate the potential semantic similarity among mismatched pairs. In the following, we present the details of our method.

4.1. Identifying Mismatched Pairs

Following the mainstream learning style, we first identify possibly mismatched pairs from all training data. The memorization effect of DNNs indicates that mismatched samples tend to have relatively higher loss during the early stage of training. Based on this, we use the difference in loss distribution between the matched and mismatched pairs to divide the training set. Empirically, we observe that the distribution of triplet loss is more distinguishable. Thus, given the retrieval model (f_v, f_t, g), we compute the per-sample loss through Eq.(2):

$$\ell_{(f_v, f_t, g)} = \{\ell_i\}_{i=1}^N = \{\mathcal{L}^{\text{triplet}}(V_i, T_i)\}_{i=1}^N. \tag{4}$$

Then, we fit a two-component beta mixture model [2, 19, 45] to $\ell_{(f_v, f_t, g)}$ using the Expectation-Maximization algorithm. For i-th pair, its probability w_i being mismatched is the posterior probability $p(b|\ell_i)$, where b is the beta component with a higher mean. By setting a threshold on $\{w_i\}_{i=1}^N$, we can divide the training data into the matched subset $\mathbb{D}_m$ and mismatched subset $\mathbb{D}_{\widetilde{m}}$ (we set the threshold to 0.5 in all experiments for brevity).

For initial convergence of the algorithm, we warm up the model for a few epochs by training on all data with Eq.(2) or Eq.(3). However, for extreme mismatching rates, the model would quickly overfit to mismatched pairs and produce unreliable loss. To address this issue, we mitigate the overconfidence of the model by adding a reverse cross entropy [41] term to the InfoNCE loss during warm-up, *i.e.*,

$$\mathcal{L}^{\text{RCE}}(V_i, T_i) = \mathcal{H}(\boldsymbol{p}_i^{v2t}, \boldsymbol{y}_i) + \mathcal{H}(\boldsymbol{p}_i^{t2v}, \boldsymbol{y}_i). \tag{5}$$

In the presence of PMPs, $\boldsymbol{y}_i$ may provide the wrong matching relation. Instead, the estimated probability could reflect

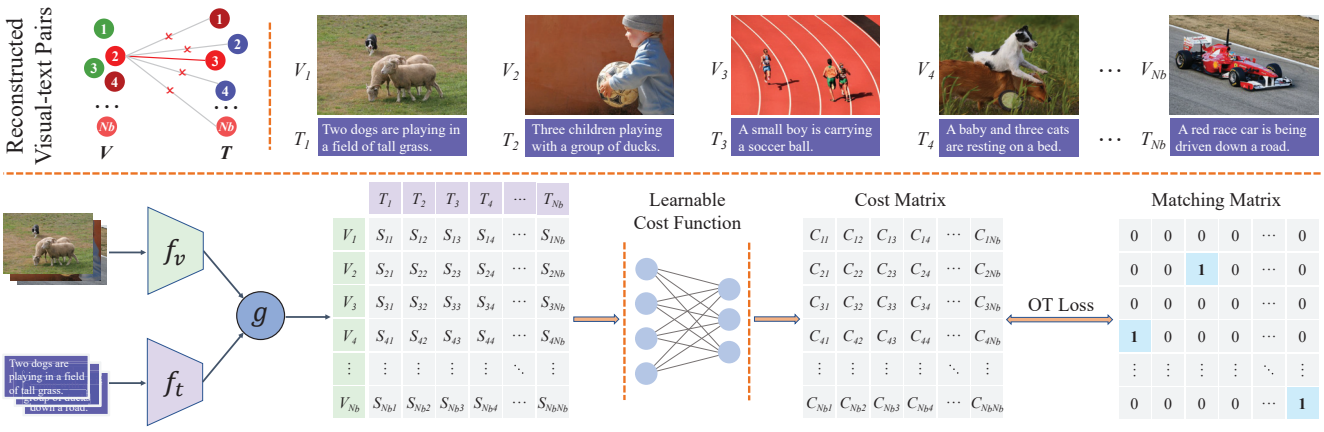


Figure 2. Overview of the learnable cost function with self-supervised learning. The up part illustrates the reconstructed pairs that only (V_4, T_1), (V_2, T_3), and (V_{N_b}, T_{N_b}) are the reserved matching ones. Then, the matching matrix is viewed as supervision to guide the cost function from the explicit similarity-cost mapping relation through an OT loss (the down part).

the truer distribution to a certain extent. Note that we bound the one-hot label into $[\epsilon, 1-\epsilon]$ for computational feasibility. ($\epsilon = 10^{-7}$ in our experiments).

4.2. Rematching Mismatched Pairs

We formalize the rematching idea as an OT problem, generating refined alignments by seeking a minimal-cost transport plan. We will first introduce the novel learnable cost function to suit the PMP scenario, then we show how to boost the refined alignment by a relaxed OT model.

Cost Function with Self-Supervised Learning. Cost function plays a crucial role when learning the transport plan for OT. In general, $\boldsymbol{C}_{ij}$ is set to a distance measure, *e.g.*, L_2-distance [18] or cosine distance [10] to measure the expense of transporting a visual sample i to a textual sample j. However, the existence of PMPs imposes formidable obstacles for these feature-driven distance measures. On the one hand, training with PMPs can wrongly bring irrelevant data together, which undoubtedly prevents effective representation learning. Even worse, different modalities will be embedded into separate regions of the shared space due to the inherent modality gap [27]. On the other hand, the refined alignments produced by those corrupted features would be used to guide subsequent training, leading to the cycle of self-reinforcing errors [8].

To address the aforementioned limitations, we propose a novel self-supervised learning approach to automatically learn the cost function. Intuitively, for a given image and caption, the transport cost can be modeled as a function of similarity that higher similarity enjoys a lower cost. Thus, we formulate the cost function as a single-layer feed-forward network with parameters Θ_c, *i.e.*, $f_c(;\Theta_c)$, which takes the similarity matrix of the batched visual-text samples as input and attempts to learn the corresponding cost matrix. To achieve this, we reconstruct the visual-text pairs to guide the cost function from explicit similarity-cost mapping relation. Specifically, for the matched pairs sampled from $\mathbb{D}_m$, we randomly reserve a part of the matching images and substitute the images from $\mathbb{D}_{\widetilde{m}}$ for the remaining ones. With the reserved indexes, we could automatically obtain a matching matrix that indicates the ideal matching probability for each reconstructed pair. For the example illustrated in Fig. 2, (V_4, T_1), (V_2, T_3), and (V_{N_b}, T_{N_b}) are the reserved matching pairs with a matching probability of 1, while the others could be considered as mismatched ones with a matching probability of 0. For convenience, let $\mathbb{D}'$ be the reconstructed data, and $(\boldsymbol{V}, \boldsymbol{T}) \in \mathbb{D}'$ be matrices that contain a batch of images and captions. To relate the similarity-cost mapping with the matching matrix, we optimize the cost function by the following OT loss:

$$\mathcal{L}_{\mathrm{OT}}(\boldsymbol{\pi}^{\mathrm{sup}}, \boldsymbol{V}, \boldsymbol{T}) = \langle \boldsymbol{\pi}^{\mathrm{sup}}, f_c\left(g\left(\boldsymbol{V}, \boldsymbol{T}\right); \Theta_c\right)\rangle_F, \qquad (6)$$

where $\boldsymbol{\pi}^{\mathrm{sup}}$ is the matching matrix, and $g(\boldsymbol{V}, \boldsymbol{T})$ denotes the similarity matrix for the batched visual-text pairs.

Eq.(6) seeks an effective cost function from a reverse perspective of OT, which views the ideal transport plan as the supervision to minimize the transport cost.

Boosting Refined Alignments with Relaxed OT. Given the defined cost function, we could generate the refined alignments for mismatched pairs following the OT objective described in Eq.(1). However, Eq.(1) requires the two distributions to have the same total mass and that all the mass of $\boldsymbol{p}$ should be transported to exactly match the mass of $\boldsymbol{q}$. In practice, due to the limited batch size, one caption may be irrelevant to all images in the batch and vice versa. To this end, we adopt the partial OT model [6, 16] to relax such strict all-to-all mass constraints, which seeks a minimal cost of only transporting $0 \leqslant \rho \leqslant \min\left(\|\boldsymbol{p}\|_1, \|\boldsymbol{q}\|_1\right)$

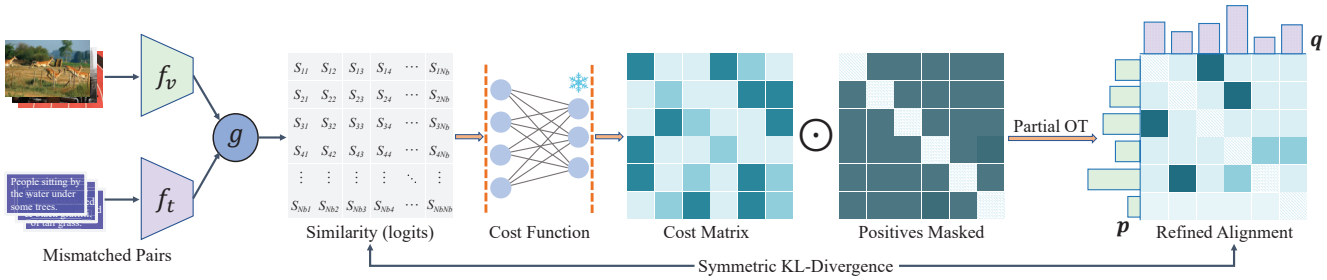


Figure 3. Illustration of the proposed rematching loss. For the mismatched pairs, we formalize a partial OT problem with positive pairs masked to generate the refined alignment in each batch. The refined alignment provides a more reliable matching relation to supervise the mismatched pairs. Then, we compute the symmetric KL-divergence to optimize the retrieval model (f_v, f_t, g).

unit mass between the visual and textual distribution, *i.e.*,

$$\min_{\boldsymbol{\pi}\in\Pi^{\rho}(\boldsymbol{p},\boldsymbol{q})} \langle \boldsymbol{\pi}, \boldsymbol{C}\rangle_F, \text{ s.t. } \Pi^{\rho}(\boldsymbol{p},\boldsymbol{q}) = \{\boldsymbol{\pi}\in\mathbb{R}_{+}^{m\times n} | \boldsymbol{\pi}\mathbb{1}_n \leqslant \boldsymbol{p}, \boldsymbol{\pi}^{\top}\mathbb{1}_m \leqslant \boldsymbol{q}, \mathbb{1}_m^{\top}\pi\mathbb{1}_n = \rho\}. \tag{7}$$

Furthermore, the false positives contained in the mismatched pairs introduce an implicit constraint to our transport plan $\boldsymbol{\pi}$ that the transport mass between the same element in two distributions should be limited. To this end, we propose to impose a mask operation on the transport plan that restricts the transport to only concentrate among the unpaired pairs. Specifically, the mask matrix $\boldsymbol{M}\in\mathbb{R}^{m\times n}$ is defined as:

$$M_{ij} \triangleq \begin{cases} 0, & \text{if } i=j, \\ 1, & \text{otherwise.} \end{cases} \tag{8}$$

Then the masked transport plan is defined as the Hadamard product $\tilde{\boldsymbol{\pi}} = \boldsymbol{M}\odot\boldsymbol{\pi}$ and be optimized through Eq.(7).

4.3. The Training Objective

Given the mismatched pairs $\{(V_i, T_i)\}_{i=1}^{N_b}$ sampled from $\mathbb{D}_{\tilde{m}}$, if we don't have any prior knowledge, we could consider the visual and textual samples follow the uniform distributions, *i.e.*, $\boldsymbol{p} = \sum_{i=1}^{N_b}\frac{1}{N_b}\delta(V_i)$ and $\boldsymbol{q} = \sum_{i=1}^{N_b}\frac{1}{N_b}\delta(T_i)$, respectively. To guarantee the efficiency of our algorithm, we adopt an online strategy to update Θ_c and calculate $\tilde{\boldsymbol{\pi}}$ through a single optimization loop:

$$\begin{aligned} &\min_{\tilde{\boldsymbol{\pi}}\in\Pi^{\rho}(\boldsymbol{p},\boldsymbol{q})} \mathbb{E}_{(\boldsymbol{V},\boldsymbol{T})\in\mathbb{D}_{\tilde{m}}} \left\langle \tilde{\boldsymbol{\pi}}, f_c\left(g\left(\boldsymbol{V},\boldsymbol{T}\right);\Theta_c^*\right)\right\rangle_F - \lambda H(\tilde{\boldsymbol{\pi}}), \\ &\text{s.t. } \Theta_c^* = \arg\min_{\Theta_c} \mathbb{E}_{(\boldsymbol{V},\boldsymbol{T},\boldsymbol{\pi}^{\text{sup}})\in\mathbb{D}'} \mathcal{L}_{OT}(\boldsymbol{\pi}^{\text{sup}}, \boldsymbol{V}, \boldsymbol{T}) \end{aligned} \tag{9}$$

where $\lambda > 0$ is a regularization parameter for the entropic constraint $H(\tilde{\boldsymbol{\pi}}) = -\sum_{ij}\tilde{\boldsymbol{\pi}}_{ij}\log\tilde{\boldsymbol{\pi}}_{ij}$. Note that Eq.(9) introduces an entropy regularization item to the OT model, which enables the transport plan to be solved by the computationally cheaper Sinkhorn-Knopp algorithm [9]. The detailed solution is presented in Appendix A.

The optimal transport plan from Eq.(9) represents a refined alignment that provides a more reliable matching relation for those mismatched visual-text samples. As our refined alignment is generated dynamically, we adopt the KL-divergence to compute the rematching loss instead of the cross entropy. Besides, a reverse term is added to symmetrize the KL-divergence, which makes the training more stable. Formally, let $\tilde{\boldsymbol{\pi}}_i^{v2t}$ and $\tilde{\boldsymbol{\pi}}_i^{t2v}$ be the row-wise and column-wise normalized refined alignment for the i-th sample, respectively. Then, the rematching loss (see Fig. 3) is defined as:

$$\begin{aligned} \mathcal{L}^{\text{re}}(V_i, T_i) = &\frac{1}{2}\left[D_{KL}(\tilde{\boldsymbol{\pi}}_i^{v2t} \parallel \boldsymbol{p}_i^{v2t}) + D_{KL}(\boldsymbol{p}_i^{v2t} \parallel \tilde{\boldsymbol{\pi}}_i^{v2t})\right] \\ &+\frac{1}{2}\left[D_{KL}(\tilde{\boldsymbol{\pi}}_i^{t2v} \parallel \boldsymbol{p}_i^{t2v}) + D_{KL}(\boldsymbol{p}_i^{t2v} \parallel \tilde{\boldsymbol{\pi}}_i^{t2v})\right]. \end{aligned} \tag{10}$$

For the pairs that are divided as matched, we use the triplet ranking loss to directly control the distance gap. Thus, our final objective function is defined as:

$$\mathcal{L}^{\text{Final}} = \sum_{(V_i,T_i)\in\mathbb{D}_m} \mathcal{L}^{\text{triplet}}(V_i, T_i) + \sum_{(V_i,T_i)\in\mathbb{D}_{\tilde{m}}} \mathcal{L}^{\text{re}}(V_i, T_i). \tag{11}$$

The detailed training pseudo-code is shown in Appendix B.

5. Experiment

In this section, we experimentally analyze the effectiveness of L2RM in robust cross-modal retrieval.

5.1. Setup

Datasets. We apply our method to three image-text retrieval datasets varying in scale and scope. Specifically, Flickr30K [46] consists of 31,000 images with five corresponding text annotations for each image from the Flickr website. Following [23], we split 1,000 images for validation, 1,000 images for testing, and the rest for training. MS-COCO [28] is a large-scale cross-modal dataset, which collects 123,287 images with five sentences each. Following [23], we use 5,000 images for validation, 5,000 images for testing, and the rest for training. Conceptual Captions [36] is a web-crawled large-scale dataset containing

MRate	Method	Flickr30K							MS-COCO						
		Image-to-Text			Text-to-Image			rSum	Image-to-Text			Text-to-Image			rSum
		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
0.2	IMRAM	59.1	85.4	91.9	44.5	71.4	79.4	431.7	69.9	93.6	97.4	55.9	84.4	89.6	490.8
	NCR	73.5	93.2	96.6	56.9	82.4	88.5	491.1	76.6	95.6	98.2	60.8	88.8	95.0	515.0
	BiCro	74.7	94.3	96.8	56.6	81.4	88.2	492.0	76.6	95.4	98.2	61.3	88.8	94.8	515.1
	DECL-SGR	74.5	92.9	97.1	53.6	79.5	86.8	484.4	75.6	95.1	98.3	59.9	88.3	94.7	511.9
	DECL-SGRAF	77.5	93.8	97.0	56.1	81.8	88.5	494.7	77.5	95.9	98.4	61.7	89.3	95.4	518.2
	RCL-SGR	74.2	91.8	96.9	55.6	81.2	87.5	487.2	77.0	95.5	98.1	61.3	88.8	94.8	515.5
	RCL-SGRAF	75.9	94.5	97.3	57.9	82.6	88.6	496.8	78.9	96.0	98.4	62.8	89.9	95.4	521.4
	L2RM-SAF	73.7	94.3	97.7	56.8	81.8	88.1	492.4	77.9	96.0	98.3	62.1	89.2	94.9	518.4
	L2RM-SGR	76.5	93.7	97.3	55.5	81.5	88.0	492.5	78.4	95.7	98.3	62.1	89.1	94.9	518.5
	L2RM-SGRAF	**77.9**	**95.2**	**97.8**	**59.8**	**83.6**	**89.5**	**503.8**	**80.2**	**96.3**	**98.5**	**64.2**	**90.1**	**95.4**	**524.7**
0.4	IMRAM	44.9	73.2	82.6	31.6	56.3	65.6	354.2	51.8	82.4	90.9	38.4	70.3	78.9	412.7
	NCR	68.1	89.6	94.8	51.4	78.4	84.8	467.1	74.7	94.6	98.0	59.6	88.1	94.7	509.7
	BiCro	70.7	92.0	95.5	51.9	77.7	85.4	473.2	75.2	95.3	98.1	60.0	87.8	94.3	510.7
	DECL-SGR	69.0	90.2	94.8	50.7	76.3	84.1	465.1	73.6	94.6	97.9	57.8	86.9	93.9	504.7
	DECL-SGRAF	72.7	92.3	95.4	53.4	79.4	86.4	479.6	75.6	95.5	98.3	59.5	88.3	94.8	512.0
	RCL-SGR	71.3	91.1	95.3	51.4	78.0	85.2	472.3	73.9	94.9	97.9	59.0	87.4	93.9	507.0
	RCL-SGRAF	72.7	92.7	96.1	54.8	80.0	87.1	483.4	77.0	95.5	98.3	61.2	88.5	94.8	515.3
	L2RM-SAF	72.1	92.1	96.1	52.7	78.8	85.9	477.7	74.4	94.7	98.3	59.2	87.9	94.4	508.9
	L2RM-SGR	73.1	92.4	96.3	52.3	79.4	86.3	479.8	75.2	94.8	98.1	59.4	87.8	94.1	509.4
	L2RM-SGRAF	**75.8**	**93.2**	**96.9**	**56.3**	**81.0**	**87.3**	**490.5**	**77.5**	**95.8**	**98.4**	**62.0**	**89.1**	**94.9**	**517.7**
0.6	IMRAM	16.4	38.2	50.9	7.5	19.2	25.3	157.5	18.2	51.6	68.0	17.9	43.6	54.6	253.9
	NCR	13.9	37.7	50.5	11.0	30.1	41.4	184.6	0.1	0.3	0.4	0.1	0.5	1.0	2.4
	BiCro	64.1	87.1	92.7	47.2	74.0	82.3	447.4	73.2	93.9	97.6	57.5	86.3	93.4	501.9
	DECL-SGR	64.5	85.8	92.6	44.0	71.6	80.6	439.1	69.7	93.4	97.5	54.5	85.2	92.6	492.9
	DECL-SGRAF	65.2	88.4	94.0	46.8	74.0	82.2	450.6	73.0	94.2	97.9	57.0	86.6	93.8	502.5
	RCL-SGR	62.3	86.3	92.9	45.1	71.3	80.2	438.1	71.4	93.2	97.1	55.4	84.7	92.3	494.1
	RCL-SGRAF	67.7	89.1	93.6	48.0	74.9	83.3	456.6	74.0	94.3	97.5	57.6	86.4	93.5	503.3
	L2RM-SAF	66.1	88.8	93.8	47.8	74.2	82.2	452.9	71.2	93.4	97.5	56.5	85.9	93.0	497.5
	L2RM-SGR	65.1	87.8	93.6	47.0	73.5	81.5	448.5	72.7	93.9	97.5	56.9	86.2	93.3	500.5
	L2RM-SGRAF	**70.0**	**90.8**	**95.4**	**51.3**	**76.4**	**83.7**	**467.6**	**75.4**	**94.7**	**97.9**	**59.2**	**87.4**	**93.8**	**508.4**
0.8	IMRAM	3.1	9.7	5.2	0.3	0.9	1.9	21.1	1.3	5.0	8.3	0.2	0.6	1.3	16.7
	NCR	1.5	6.2	9.9	0.3	1.0	2.1	21.0	0.1	0.3	0.4	0.1	0.5	1.0	2.4
	BiCro	2.3	9.2	17.2	2.6	10.2	16.8	58.3	62.2	88.6	94.6	47.4	79.2	88.5	460.5
	DECL-SGR	44.4	72.6	82.0	33.9	59.5	69.0	361.4	60.0	88.7	94.5	45.9	78.8	88.3	456.2
	DECL-SGRAF	53.4	78.8	86.9	37.6	63.8	73.9	394.4	64.8	90.5	96.0	49.7	81.7	90.3	473.0
	RCL-SGR	47.1	70.5	79.4	30.3	56.1	66.3	349.7	63.2	89.3	95.2	47.6	78.7	88.0	462.0
	RCL-SGRAF	51.7	75.8	84.4	34.5	61.2	70.7	378.3	67.4	90.8	96.0	50.6	81.0	90.1	475.9
	L2RM-SAF	50.8	77.9	85.5	35.6	62.6	72.7	385.1	64.7	90.8	95.8	50.0	80.9	89.4	471.6
	L2RM-SGR	50.5	77.2	83.9	34.2	61.1	71.6	378.5	65.2	90.3	96.1	49.8	81.0	88.2	470.6
	L2RM-SGRAF	**55.7**	**80.8**	**87.8**	**39.4**	**65.4**	**74.9**	**404.0**	**69.0**	**91.9**	**96.4**	**52.6**	**82.4**	**90.3**	**482.6**

Table 1. Image-text retrieval performance under different mismatching rates (MRate) on Flickr30K and MS-COCO.

3.3M one-to-one images and captions. Following [23], we use the subset, *i.e.*, CC152K to conduct experiments, which has 150,000 images for training, 1,000 images for validation, and 1,000 images for testing.

Implementation Details. As a general method, L2RM could be directly applied to almost all cross-modal retrieval methods to improve their robustness. Following [21, 33], we apply L2RM to SGR, SAF, and SGRAF for a comprehensive comparison. We evaluate the retrieval performance with the Recall@K (R@K) metric. Following [23], we save the best performance checkpoint on the validation set w.r.t. the sum of the evaluation scores and report its results on the testing set. We follow the same training setting as [23], our specific parameters setting can be found in Appendix C.1.

Baselines. We compare L2RM with eight state-of-the-art cross-modal retrieval methods, including four general methods (*i.e.*, IMRAM [7], SGR, SAF, and SGRAF [12]) and four robust learning methods against the PMPs (*i.e.*, NCR [23], DECL [33], BiCro [45], and RCL [21]). Note that the original BiCro combines four models, *i.e.*, two co-trained SGR, and two co-trained SAF. For a fair comparison, we report the results of 2 co-trained SGR for BiCro like [23].

5.2. Main Results

In this section, we conduct comparison experiments with different mismatching rates on three datasets to evaluate the performance of our L2RM. As Flickr30K and MS-COCO are well-established datasets, we carry out experiments by

generating the synthesized false positive pairs, *i.e.*, the mismatching rate (MRate) increases from 0.2 to 0.8 in intervals of 0.2. Following [21, 33], we randomly select a specific percentage of images and randomly permute all their corresponding captions, which is more challenging and practical than the setting in [23, 45]. For the web-collected dataset CC152K, which naturally contains about 3% $\sim$ 20 % unknown mismatched pairs [36]. Thus we directly conduct experiments on it to evaluate the performance with real PMPs.

Results on Synthesized PMPs. Tab. 1 shows the experimental results on Flickr30K and MS-COCO. Note that for MS-COCO, the results are computed by averaging over 5 folds of 1K test images like [21, 33]. Due to space limitation, we omit the results of some general methods (SGR, SAF, and SGRAF), and the comparison on original datasets (0 MRate), which could be found in Appendix C.2. From the results, we can find that L2RM achieves the best results on all metrics than the other state-of-the-art methods, which shows the superior robustness of L2RM against PMPs. Moreover, when the mismatching rate is high, *e.g.*, 0.6 and 0.8, the improvement of L2RM is more evident, proving that excavating mismatched pairs could effectively facilitate robust cross-modal retrieval.

Results on Real-World PMPs. We validate our method on the real-world dataset CC152K, which contains an unknown portion of mismatched pairs. As shown in Tab. 2, our method considerably outperforms the best baseline in terms of sum in retrieval by 9.8%. Notably, our L2RM-SGR surpasses all SGR variants by a clear margin, achieving as much as a 16.9% (rSum) absolute improvement over the best variant. It is because the real-world rematched pairs are more likely to involve only local alignments, *e.g.*, Fig. 5(e)-Fig. 5(f), while the SGR model itself is adept at capturing the relationship between local alignments.

Method	Image-to-Text			Text-to-Image			rSum
	R@1	R@5	R@10	R@1	R@5	R@10	
IMRAM	27.8	52.4	60.9	29.2	51.5	61.2	283.0
SAF	32.5	59.5	70.0	32.5	60.7	68.7	323.9
SGR	14.5	35.5	48.9	13.7	36.1	47.9	196.6
NCR	39.5	64.5	73.5	40.3	64.6	73.2	355.6
BiCro	39.7	64.6	72.6	39.2	65.0	74.1	355.2
DECL-SAF	36.6	63.0	73.3	38.5	63.2	73.5	348.1
DECL-SGR	36.2	63.6	73.2	37.1	63.6	73.7	347.4
DECL-SGRAF	39.0	66.1	75.5	40.7	66.3	76.7	364.3
RCL-SAF	37.5	63.0	71.4	37.8	62.4	72.4	344.5
RCL-SGR	38.3	63.0	70.4	39.2	63.2	72.3	346.4
RCL-SGRAF	41.7	66.0	73.6	41.6	66.4	75.1	364.4
L2RM-SAF	37.3	62.7	71.7	38.8	65.7	74.8	351.0
L2RM-SGR	39.5	66.2	**76.0**	41.8	65.9	74.9	364.3
L2RM-SGRAF	**43.0**	**67.5**	75.7	**42.8**	**68.0**	**77.2**	**374.2**

Table 2. Image-text retrieval performance on CC152K.

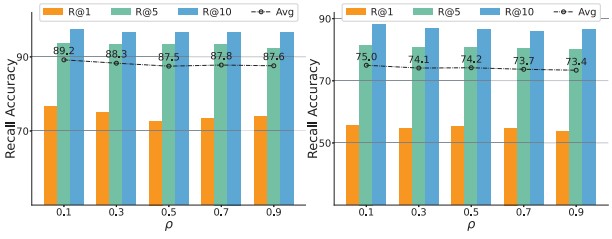
5.3. Ablation Study

Impact of Each Component. To study the influence of specific components in our method, we carry out the ablation study on the Flickr30K with 0.6 MRate. Specifically, we ablate the contributions of three key components of L2RM, *i.e.*, partial OT, positives masked, and the learnable cost function (we use the cosine distance to measure the cost instead). Besides, we compare L2RM with different formulas of rematching loss: KL-divergence and InfoNCE. From Tab. 3, we observe the following conclusions: 1) The full L2RM could achieve the best overall performance, showing that all three components are important to improve the robustness against PMPs. 2) Using the learnable cost function substantially outperforms the variant with the cosine distance cost (*e.g.*,+11.9 in terms of the rSum), which signifies the simple feature-driven cost is sub-optimal to the PMP situation. 3) Formulating different rematching loss could also achieve decent results, which verifies the ability of L2RM to provide effective matching relations.

Ablation	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
L2RM	**65.1**	**87.8**	**93.6**	**47.0**	73.5	**81.5**
L2RM w KL-divergence	64.7	87.6	93.2	46.7	**74.0**	81.5
L2RM w InfoNCE Loss	64.9	87.5	93.5	46.0	72.9	80.9
L2RM w/o Partial OT	62.2	86.4	91.5	44.7	68.6	72.6
L2RM w/o Positives Masked	64.9	87.6	92.7	46.4	73.2	81.1
L2RM w/o Cost Function	61.3	85.6	91.4	44.9	72.4	81.0

Table 3. Ablation studies on Flickr30K with 0.6 MRate.

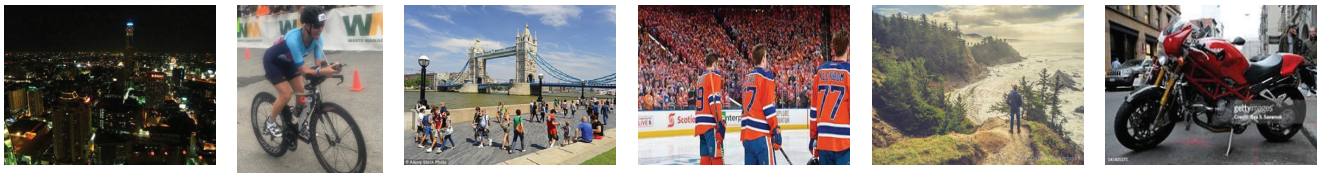
Parameter Analysis. We now investigate the effect of the parameter ρ by plotting the recall scores with incremental ρ on Flickr30K. The figure shows that the overall performance tends to decrease as ρ increases. We further analyze how untransported pairs benefit the model training in Appendix C.3. Experimentally, we find that the refined alignments for untransported pairs can be equivalent to the label smoothing strategy [38].



(a) Image to Text (b) Text to Image

Figure 4. Parameter analysis of L2RM-SGR in terms of recall scores on the testing set of Flickr30K under 0.2 MRate.

Discussions on Warm-up Methods. We use different warm-up methods, *i.e.*, triplet loss [13] and InfoNCE loss [31] for our L2RM-SGR. The experiments are conducted on the Flickr30K with 0.8 MRate and the CC152K with real



Given: view bar out on the lounge chairs

Rematched: the facade can be lit up at night using different coloured led lights (0.988)

(a)

Given: the gift of the present

Rematched: a man on a bicycle (0.996)

(b)

Given: a survey has revealed that a third would be less likely to travel if it was no longer a member

Rematched: suspension bridge over a city (0.963)

(c)

Given: video: fans sing composition after mic issue

Rematched: most popular sports in the world – hockey (0.965)

(d)

Given: day on the place unknown

Rematched: a couple of men stand on a small hill (0.938)

(e)

Given: a general view of atmosphere outside the diesel and launch party

Rematched: a motorcycle winds their way through country (0.626)

(f)

Figure 5. The ability of our L2RM to rematch the mismatched visual-text samples. The figure shows some representative rematched pairs for L2RM-SGR on the training set of CC152K dataset. We highlight the matched words in green and the mismatched words in red.

Method	Flickr30K						CC152K					
	Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Triplet	0.1	0.7	1.3	0.1	0.7	1.1	38.2	64.2	71.4	39.5	64.2	71.9
InfoNCE	45.9	71.4	80.3	29.2	52.7	61.1	39.3	66.8	75.0	40.8	65.2	74.2

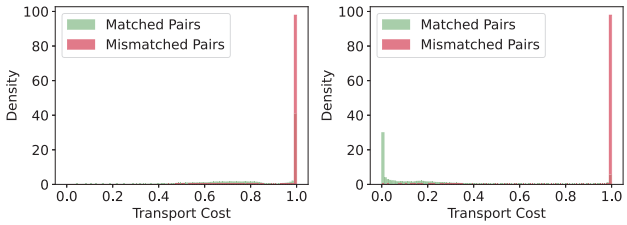
Table 4. Comparison with different warm-up methods on Flickr30K with 0.8 MRate and CC152K.

PMPs. As shown in Tab. 4, one could see that the triplet loss cannot achieve satisfactory performance under the extreme mismatching rate. Compared with the results of the L2RM-SGR in Tab. 1, one could find that it is necessary to limit the overconfidence of the model during the warm-up process. The results on CC152K show that our method is robust to the choice of warm-up methods under a relatively low mismatching rate.

5.4. Visualization and Analysis

Distribution of Transport Cost. To intuitively show the effectiveness of the learnable cost function, we illustrate the transport cost for matched and mismatched training pairs on Flickr30K with 0.8 MRate. From Fig. 6, one could see that our cost function first learns to assign higher transport costs to those mismatched pairs. Although the costs of matched pairs are distributed over a large range in the early stage, they gradually become smaller and tend toward 0 as training proceeds. In conclusion, our cost function could successfully learn to distinguish matched and mismatched data, which lays the foundation for the further OT model.

Visualizing Re-matched Image-Text Pairs. To visually illustrate the rematching ability of our L2RM, we conduct the case study on CC152K to show real-world rematched examples. Specifically, the first two rows of Fig. 5 show the image and its original mismatched caption, respectively. The third row shows the rematched caption provided by our method, and we also show the refined alignment scores in brackets. In particular, we could find that some real-world visual-text pairs are completely uncorrelated (*e.g.*, Fig. 5(a)-Fig. 5(b)) or contain only a few local similarities (*e.g.*, Fig. 5(c)-Fig. 5(e)). Thanks to our L2RM, the



(a) Epoch 5 (b) Epoch 35

Figure 6. Transport cost distribution for matched and mismatched pairs at different training phases of our L2RM. The experiments are conducted on Flickr30K with 0.8 MRate.

potential matching relation among mismatched pairs could be fully excavated to provide refined alignments. For example, one could see that the rematched caption, *i.e.*, *"a man on a bicycle"* nicely expresses the semantic concept in Fig. 5(a). Although some rematched captions could not perfectly share the same semantics with images, they also contain some local similarities to the given images. For example, the image in Fig. 5(f) is correctly described with the words *"a motorcycle"* and our L2RM provides a relatively low refined alignment score as the target. In summary, our proposed rematching strategy could embrace better data efficiency and robustness against PMPs.

6. Conclusion

This work studies the challenge of cross-modal retrieval with partially mismatched pairs (PMPs). To address this problem, we propose L2RM, a generalized OT-based framework that learns to rematch mismatched pairs. Our key idea is to excavate the potential semantic similarity among unpaired samples. To formalize this idea through OT, first, we propose a self-supervised learner to automatically learn effective cost function. Second, we model a partial OT problem and restrict the transport among false positives to further boost refined alignments. Extensive experiments are conducted to verify that our L2RM can endow cross-modal retrieval models with strong robustness against PMPs.

References

[1] Triantafyllos Afouras, Yuki M Asano, Francois Fagan, Andrea Vedaldi, and Florian Metze. Self-supervised object detection from audio-visual correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10575–10586, 2022. 2

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pages 312–321. PMLR, 2019. 3

[3] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017. 3

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 2

[5] Wanxing Chang, Ye Shi, Hoang Tuan, and Jingya Wang. Unified optimal transport framework for universal domain adaptation. *Advances in Neural Information Processing Systems*, 35:29512–29524, 2022. 2

[6] Laetitia Chapel, Mokhtar Z Alaya, and Gilles Gasso. Partial optimal tranport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33:2903–2913, 2020. 4

[7] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12655–12663, 2020. 6

[8] Mingcai Chen, Hao Cheng, Yuntao Du, Ming Xu, Wenyu Jiang, and Chongjun Wang. Two wrongs don't make a right: Combating confirmation bias in learning with label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14765–14773, 2023. 4

[9] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 2, 5

[10] Dan dan Guo, Long Tian, Minghe Zhang, Mingyuan Zhou, and Hongyuan Zha. Learning prototype-oriented set representations for meta-learning. In *International Conference on Learning Representations*, 2021. 2, 4

[11] Karan Desai, Gaurav Kaul, Zubin Trivadi Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 1

[12] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1218–1226, 2021. 6

[13] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 2, 3, 7

[14] Kilian Fatras, Hiroki Naganuma, and Ioannis Mitliagkas. Optimal transport meets noisy label robust loss and mixup regularization for domain adaptation. In *Conference on Lifelong Learning Agents*, pages 966–981. PMLR, 2022. 2

[15] Yanglin Feng, Hongyuan Zhu, Dezhong Peng, Xi Peng, and Peng Hu. Rono: Robust discriminative learning with noisy labels for 2d-3d cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11610–11619, 2023. 1

[16] Alessio Figalli. The optimal partial transport problem. *Archive for rational mechanics and analysis*, 195(2):533–560, 2010. 4

[17] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 303–312, 2021. 2

[18] Xiang Gu, Yucheng Yang, Wei Zeng, Jian Sun, and Zongben Xu. Keypoint-guided optimal transport with applications in heterogeneous domain adaptation. *Advances in Neural Information Processing Systems*, 35:14972–14985, 2022. 2, 4

[19] Haochen Han, Kaiyao Miao, Qinghua Zheng, and Minnan Luo. Noisy correspondence learning with meta similarity correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7517–7526, 2023. 1, 3

[20] Xiaoshuai Hao, Wanqian Zhang, Dayan Wu, Fei Zhu, and Bo Li. Dual alignment unsupervised domain adaptation for video-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18962–18972, 2023. 1

[21] Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2, 3, 6, 7

[22] Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, and Donglin Wang. Vop: Text-video co-operative prompt tuning for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6565–6574, 2023. 1

[23] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419, 2021. 1, 2, 3, 5, 6, 7

[24] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[25] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2787–2797, 2023. 1

[26] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4654–4662, 2019. 2

[27] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 4

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[29] Yijie Lin, Jie Zhang, Zhenyu Huang, Jia Liu, Zujie Wen, and Xi Peng. Multi-granularity correspondence learning from long-term noisy videos. *arXiv preprint arXiv:2401.16702*, 2024. 2

[30] Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15692–15701, 2022. 2

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 7

[32] Hanyu Peng, Mingming Sun, and Ping Li. Optimal transport for long-tailed recognition with learnable cost matrix. In *International Conference on Learning Representations*, 2022. 2

[33] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4948–4956, 2022. 2, 6, 7

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[35] Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 849–858. PMLR, 2019. 2

[36] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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)*, pages 2556–2565, 2018. 1, 5, 7

[37] Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. Human-adversarial visual question answering. *Advances in Neural Information Processing Systems*, 34:20346–20359, 2021. 1

[38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 7

[39] Kai Sheng Tai, Peter D Bailis, and Gregory Valiant. Sinkhorn label allocation: Semi-supervised classification via annealed self-training. In *International Conference on Machine Learning*, pages 10065–10075. PMLR, 2021. 2

[40] Haobo Wang, Mingxuan Xia, Yixuan Li, Yuren Mao, Lei Feng, Gang Chen, and Junbo Zhao. Solar: Sinkhorn label refinery for imbalanced partial-label learning. *Advances in Neural Information Processing Systems*, 35:8104–8117, 2022. 2

[41] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 322–330, 2019. 3

[42] Zheng Wang, Zhenwei Gao, Kangshuai Guo, Yang Yang, Xiaoming Wang, and Heng Tao Shen. Multilateral semantic relations modeling for image text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2830–2839, 2023. 1

[43] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10941–10950, 2020. 2

[44] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing intermodality: Visual parsing with self-attention for vision-and-language pre-training. *Advances in Neural Information Processing Systems*, 34:4514–4528, 2021. 2

[45] Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19883–19892, 2023. 1, 2, 3, 6, 7

[46] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 5

[47] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3536–3545, 2020. 2