PatchCT: Aligning Patch Set and Label Set with Conditional Transport for Multi-Label Image Classification

Miaoge Li*, Dongsheng Wang*, Xinyang Liu, Zequn Zeng, Ruiying Lu, Bo Chen

National Key Laboratory of Radar Signal Processing

Xidian University, Xi'an, Shanxi 710071, China

`{limiaoge, wds, xinyangatk}@stu.xidian.edu.cn, bchen@mail.xidian.edu.cn`

Mingyuan Zhou

McCombs School of Business

The University of Texas at Austin, Austin, TX 78712, USA

`mingyuan.zhou@mccombs.utexas.edu`

Abstract

Multi-label image classification is a prediction task that aims to identify more than one label from a given image. This paper considers the semantic consistency of the latent space between the visual patch and linguistic label domains and introduces the conditional transport (CT) theory to bridge the acknowledged gap. While recent cross-modal attention-based studies have attempted to align such two representations and achieved impressive performance, they required carefully-designed alignment modules and extra complex operations in the attention computation. We find that by formulating the multi-label classification as a CT problem, we can exploit the interactions between the image and label efficiently by minimizing the bidirectional CT cost. Specifically, after feeding the images and textual labels into the modality-specific encoders, we view each image as a mixture of patch embeddings and a mixture of label embeddings, which capture the local region features and the class prototypes, respectively. CT is then employed to learn and align those two semantic sets by defining the forward and backward navigators. Importantly, the defined navigators in CT distance model the similarities between patches and labels, which provides an interpretable tool to visualize the learned prototypes. Extensive experiments on three public image benchmarks show that the proposed model consistently outperforms the previous methods.

1. Introduction

Multi-label image classification is a fundamental yet challenging task in computer vision, where a set of objects needs
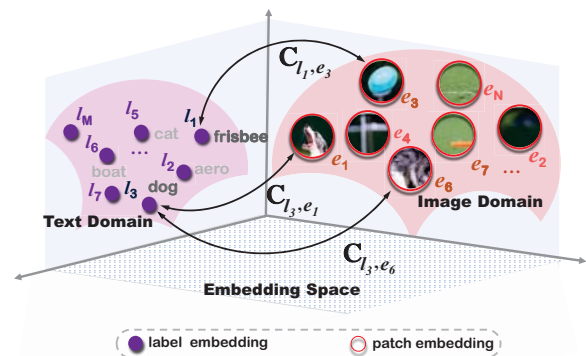

Figure 1: Motivation of the proposed PatchCT. We represent each image as a set of patch embeddings and a set of label embeddings, and then the conditional transport is employed as a semantic regularization to align such two domains.

to be predicted within one image. It has practical applications in wide fields such as image retrieval [45], scene understanding [37], recommendation [48, 24], and biology analysis [21, 1]. In addition to identifying the regions of interest, multi-label image classification also requires special attention on *1)* semantic information of labels, *e.g.*, label correlations, and *2)* interactions between the visual image and textual label domains. Thus, it is often more complex and challenging compared to single-label case.

To address the aforementioned challenges, several significant attempts have been developed from various directions. An increasing research attention is to learn semantic label representation. The core idea is intuitive: labels should be more correlated if they co-occurrence often. For instance, *traffic lights* usually co-appeared with *crosswalk*, and *chair* has a high confidence score to appear if *table* exists in the image. Existing methods adopt pair-wise ranking loss, co-

*Authors contributed equally.

variance matrices, recurrent structures, contrastive learning, graph neural networks (GCNs), and pre-trained language model to this end [55, 2, 43, 50, 25, 14, 1, 9, 8, 51, 29, 62]. Those methods either regular the learning with a pre-defined label graph which is often obtained from the training set or the pre-trained word embedding, *e.g.* Glove [33], or exploring label dependency implicitly (*e.g.*, contrastive learning and BERT embeddings [15]). Meanwhile, some works aim to solve the second issue by adopting cross-modal attention-based framework [23, 29, 26, 56, 11], where a vision transformer (ViT [17]) is often employed to obtain local patch features and the cross-attention between the labels and patches is then applied to align such two modalities. The alignment module acts as the core role in those models and needs to be designed carefully. Moreover, conventional attention mechanisms are guided by task-specific losses, without explicitly training signals to encourage alignment. The learned attention weights are therefore often dense and uninterpretable, leading to less effective relational prediction [5].

We address whether there is a more simple and principled approach to efficient alignments of vision-text domains. To explore this, in this paper, we introduce the conditional transport (CT) theory [59, 39] and reformulate the multi-label classification task as a CT problem, where an image is represented as two discrete distributions over different supports, *e.g.*, the visual patch domain and the textual label domain (as shown in Fig. 1). Our idea is intuitive: those two distributions (or sets) are different modality representations of the same image, and therefore they share semantic consistency. As a result, the learning of multi-label classification can be viewed as the process of aligning the textual label set to be as close to the visual set as possible. Accordingly, it is indeed key to find out how to measure the distance between two empirical distributions with different supports. Fortunately, conditional transport is well-studied in recent researches [59, 38, 40, 42] and provides a powerful tool for measuring the distance from one discrete distribution to another given the cost matrix. It is natural for us to develop a new framework for multi-label classification based on the minimization of CT distance.

Specifically, the visual patch embeddings and textual label embeddings are first extracted by feeding the image and label descriptions into the corresponding encoders. We specify the image encoder as a ViT to capture spatial patch dependencies and the text encoder as a BERT to explore the label semantics. Besides, inspired by the current prompt learning success [31, 60], we here design a simple and efficient template to reduce the domain shift between the language model and the multi-label classification and quickly distill the pre-trained knowledge to the downstream tasks simultaneously. After that, we collect patch features as a discrete probability distribution where each patch has a label-aware probability value that reflects the important score for multi-label predic-

tion, *e.g.* object patches have high probability values while background patches with lower attention. Similarly, we construct the textual label sets as a discrete distribution, whose probability values are obtained by normalizing the ground-truth label of the image. Given such two sets, the cost matrix in CT is then defined according to the similarity between the patches and labels, *e.g.* the cosine distance of the patch and label embeddings. The CT divergence is defined with a bidirectional set-to-set transport, where a forward CT measures the transport cost from the patch set to the label set, and a backward CT that reverses the transport direction. Therefore, by minimizing the bidirectional CT cost, we explicitly minimize the embedding distance between the domains, *i.e.*, optimizing towards better patch-label alignments. Moreover, the learned transport plan models the semantic relations between patches and labels, which provides an interpretable tool to visualize the label concepts.

Our main contributions can be summarized as follows:

- We propose a novel vision-text alignment framework based on conditional transport theory, where the interactions between the patches and labels are explored by minimizing the bidirectional CT distance between those two modalities to produce high semantic consistency for multi-label classification.

- We design sparse and layer-wise CT formulations to reduce the computational cost and enhance the deep interactions across modalities, contributing to robust and accurate alignments between patches and labels.

- Extensive experiments on three widely used visual benchmarks demonstrate the effectiveness of the proposed model by establishing consistent improvements on all data sets.

2. Related Work

2.1. Multi-label Classification

Multi-label classification has attracted increasing interest recently owing to its relevance in real-world applications. A natural idea comes from the single-label case and treats each category independently and then converts the task into a series of binary classification tasks straightforwardly. However, these methods often suffer from limited performance, due to their ignoring of the correlation between labels and objects' spatial dependency which are quite crucial for multi-label image classification [43]. To address this issue, several previous proposed attempts explicitly capture the label dependencies by a CNN-based encoder followed by an RNN [43, 6]. Apart from these sequential methods, some works resort to Graph Convolutional Networks(GCN) to model label relationships [8, 51]. However, it is also arguable that spurious correlations may be learned when the label statistics are insufficient. More recently, motivated by the great success

of ViT in various visual tasks [17, 3, 4], several works aim to align the labels and patches via the attention strategy for improving multi-label prediction. Those align-aware works are closest to our work. For example, Query2Label of [29] adopts the built-in cross-attention in the Transformer decoder as a spatial selector, in which label embeddings are treated as queries to align and pool class-related features from the ViT outputs for subsequent binary classifications. Lanchantin *et al.*[26] utilize a general multi-label framework consisting of a Transformer encoder as well as a ternary encoding scheme during training to exploit the complex dependencies among visual features and labels. Zhao *et al.*[56] propose a multi-label multi-label recognition transformer learning framework with three essential modules for complex alignments and correlations learning among inter- and intra-modalities. Different from those attention-based approaches that usually require carefully designed alignment modules and high computing costs in attention operations, we convert the multi-label classification task to a CT problem and align such two vision-label modalities by minimizing the total transport cost in a bidirectional view.

2.2. Alignment via Transport Distance

Recently, Optimal transport (OT) [41] has been used to solve the distance between two discrete distributions under unsupervised domain adaption [34], label representation [19, 57, 49], and cross-modal semantics [27, 32, 5]. For example, Lee *et al.* introduce a hierarchical OT distance that leverages clustered structure in data to improve alignment between neural data and human movement patterns [27]. Chen *et al.* formulate the cross-domain alignment as a graph matching problem and propose Graph OT (GOT) [5] to various applications, including image caption, machine translation, and text summarization. One of the core ideas behind those models is to align the multi-modalities by minimizing OT cost, they are distinct from ours in terms of task and technique detail. We focus on multi-label image classification cases where the local patch features and a set of class embeddings are considered under the CT framework. CT distance is originally developed to measure the difference between two probability distributions with the mini-batch optimization [59]. Unlike OT which usually needs to optimize the transport plan via an iterative Sinkhorn algorithm [13], CT considers the transport plan in a bidirectional view based on semantic similarity between samples from two distributions. This flexibility of CT potentially facilitates easier integration with deep neural networks with lower complexity and better scalability, showing better results on recent alignment tasks [38, 42, 39, 30]. Those properties motivated us to learn aligned vision-label semantics under the CT framework for multi-label classification.

3. Method

In this section, we begin with the background and relevant notations of multi-label classification. Then we review the preliminaries of conditional transport and introduce the details of our proposed model, showing how to formulate the multi-label classification as a CT problem.

3.1. Background and Notations

Given a dataset with I labeled samples together with a set of M labels: $\mathcal{D} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^I$, where $\boldsymbol{x}_i \in \mathbb{R}^{H \times W \times 3}$ denotes the i-th input RGB image with $H \times W$ size, and $\boldsymbol{y}_i \in \{0, 1\}^M$ denotes the multi-hot binary label vector of $\boldsymbol{x}_i$. $y_{mi} = 1$ means image $\boldsymbol{x}_i$ have label m and vice versa. Through observing $\mathcal{D}$, multi-label classification aims to derive a proper learning model, so that the label $\widetilde{\boldsymbol{y}}$ of a test image $\boldsymbol{x}$ can be predicted accordingly.

3.2. Distance Between Two Set

Let us consider two discrete probability distributions $\mathbf{P}$ and $\mathbf{Q} \in \mathcal{P}(X)$ on space $X \in \mathbb{R}^d$: $\mathbf{P} = \sum_{i=1}^n \theta_i \delta_{\boldsymbol{e}_i}$, and $\mathbf{Q} = \sum_{j=1}^m \beta_j \delta_{\boldsymbol{l}_j}$, where $\boldsymbol{e}_i$ and $\boldsymbol{l}_j$ are two points in the arbitrary same space X. $\boldsymbol{\theta} \in \Sigma^n$ and $\boldsymbol{\beta} \in \Sigma^m$, the simplex of $\mathbb{R}^n$ and $\mathbb{R}^m$, denotes two probability values of the discrete states with the satisfy that $\sum_{i=1}^n \theta_i = 1$ and $\sum_{j=1}^m \beta_j = 1$. $\delta_{\boldsymbol{e}}$ refers to a point mass located at coordinate $\boldsymbol{e} \in \mathbb{R}^d$.

To measure such two discrete distributions, OT distance between $\mathbf{P}$ and $\mathbf{Q}$ is formulated as the optimization problem: $\text{OT}(\mathbf{P}, \mathbf{Q}) = \min_{\mathbf{T} \in \Pi(\boldsymbol{\theta}, \boldsymbol{\beta})} \sum_{i,j} t_{ij} c_{ij}$, with $\mathbf{T} \mathbb{1}^m = \boldsymbol{\theta}$, $\mathbf{T}^T \mathbb{1}^n = \boldsymbol{\beta}$. Where $\mathbb{1}^m$ is the m dimensional vector of ones. $c_{ij} = c(\boldsymbol{e}_i, \boldsymbol{l}_j) \geq 0$ is the transport cost between the two points $\boldsymbol{e}_i$ and $\boldsymbol{l}_j$ defined by an arbitrary cost function $c(\cdot)$. The optimal transport plan $\mathbf{T}$ is often trained by minimizing the OT cost with the iterative Sinkhorn algorithm [13].

More recently, the demand on efficient computation and bidirectional asymmetric transport promotes the development of CT divergence [59], which can be applied to quantify the difference between discrete empirical distributions in various applications [58, 38, 42]. Specifically, given the above source and target distributions $\mathbf{P}$ and $\mathbf{Q}$, the CT cost is defined with a bidirectional distribution-to-distribution transport, where a forward CT measures the transport cost from the source to target, and a backward CT that reverses the transport direction. Therefore, the CT problem can be defined as:

$$\text{CT}_{\mathbf{C}}(\mathbf{P}, \mathbf{Q}) = \min_{\overrightarrow{\mathbf{T}} \overleftarrow{\mathbf{T}}} (\sum_{i,j} \overrightarrow{t}_{ij} c_{ij} + \sum_{j,i} \overleftarrow{t}_{ji} c_{ji}), \quad (1)$$

where $\mathbf{C}$ is the cost matrix, and $\overrightarrow{t}_{ij}$ in $\overrightarrow{\mathbf{T}}$ acts as the transport probability (the navigator) from the source point $\boldsymbol{e}_i$ to the target point $\boldsymbol{l}_j$: $\overrightarrow{t}_{ij} = \theta_i \frac{\beta_j e^{-d_\psi(\boldsymbol{e}_i, \boldsymbol{l}_j)}}{\sum_{j'=1}^m \beta_{j'} e^{-d_\psi(\boldsymbol{e}_i, \boldsymbol{l}_{j'})}}$, hence
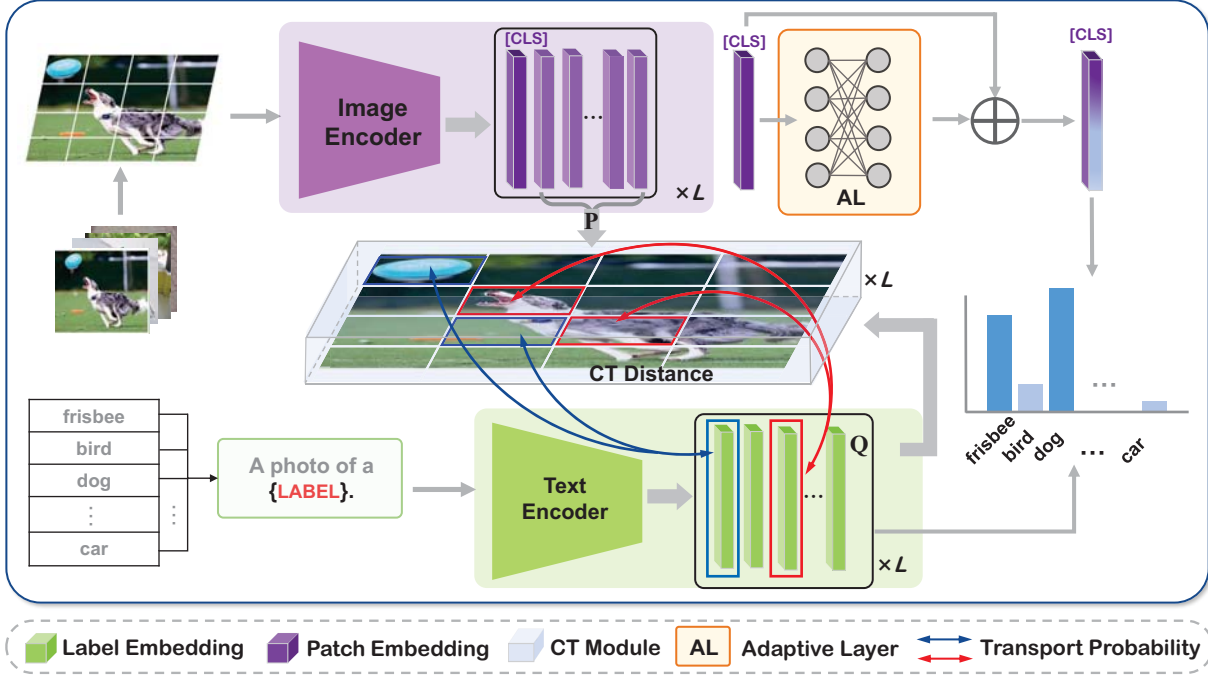
Figure 2: The overall framework of the proposed PatchCT, which loads the pre-trained ViT and BERT to capture visual patch and textual label embeddings. An adaptive module is added to transfer the knowledge implied in ViT to the multi-label classification task. A layer-wise CT distance is then applied to align the vision-text domains.

$\overrightarrow{\mathbf{T}}\mathbb{1}^m = \boldsymbol{\theta}$. Similarly, we have the reversed transport probability: $\overleftarrow{t}_{ji} = \beta_j \frac{\theta_i e^{-d_\psi(l_j, e_i)}}{\sum_{i'=1}^n \theta_{i'} e^{-d_\psi(l_j, e_{i'})}}$, and $\overleftarrow{\mathbf{T}}\mathbb{1}^n = \boldsymbol{\beta}$. The distance function $d_\psi(\boldsymbol{e}_i, \boldsymbol{l}_j)$ parameterized with ψ can be implemented by deep neural networks to measure the semantic similarity between two points, making CT amenable to stochastic gradient descent-based optimization.

3.3. The Proposed Framework

Now we present the details of our proposed PatchCT, which aligns visual **Patch** and textual label domains under **CT** framework for multi-label image classification. As shown in Fig. 2, PatchCT consists of four components, the visual **P** set, the textual label **Q** set, the adaptive layer, and the CT distance between **P** and **Q**.

P set over patch embeddings. For an input image $\boldsymbol{x}_i$, PatchCT first divides it into N patches evenly and then feeds them to the image encoder to obtain the local region features $\{\boldsymbol{e}_n\}_{n=1}^N$ (we omit i for convenience), where $\boldsymbol{e}_n \in \mathbb{R}^d$ denotes the n-th patch embedding, d is the embedding dimension. In addition to the N local features, PatchCT also learns the $[CLS]$ visual token $\boldsymbol{e}_{CLS}$ that acts as the global image representation. One of the main operations in the image encoder is the multi-layer multi-head attention layers that integrate and update the global and local features by considering the spatial and contextual information among patches.

Traditional CT settings often view each point equally, *e.g.*, $\boldsymbol{\theta}_i$ is a uniform distribution over N points. Unfortunately, this is not the case in multi-label tasks where only a few key patches contribute to the final prediction practically. To this end, PatchCT defines a sparse and label-guided $\boldsymbol{\theta}_i$ as below:

$$\boldsymbol{\theta}_i = \text{softmax}(\text{TopK}(\mathbf{E}_i^T \boldsymbol{o}_i, k)), \quad \boldsymbol{o}_i = \mathbf{L}\hat{\boldsymbol{y}}_i \quad (2)$$

where $\mathbf{E}_i \in \mathbb{R}^{d \times N}$ and $\mathbf{L} \in \mathbb{R}^{d \times M}$ are the patch embedding matrix of $\boldsymbol{x}_i$ and label embedding matrix respectively. $\hat{\boldsymbol{y}}_i$ is the normalized label vector $\hat{\boldsymbol{y}}_i = \boldsymbol{y}_i / \sum_{m=1}^M y_{mi}$, thus $\boldsymbol{o}_i$ here denotes the label-aware representation of $\boldsymbol{x}_i$, which is used to select core patches that have close semantics with the ground-truth labels. $\text{TopK}(\cdot, k)$ is a sparsity operation that masks top-k patches with 1 and 0 for others based on the similarity score, and $\text{softmax}(\cdot)$ makes sure the probability simplex of $\boldsymbol{\theta}_i$.

After giving the patch embedding matrix $\mathbf{E}_i$ and its weights $\boldsymbol{\theta}_i$, PatchCT obtains the discrete distribution $\mathbf{P}_i$ of $\boldsymbol{x}_i$ over the visual set. Note that $\mathbf{P}_i$ collects the detailed visual features of local patches, thus it brings benefits to the downstream multi-label task. For one thing, $\boldsymbol{\theta}_i$ in Eq. 2 guarantees the sparsity of $\mathbf{P}_i$ which is useful for reducing the computing cost and enhancing the interpretability of PatchCT. For another, $\mathbf{P}_i$ concentrates more on the core patches that contain objects via the introduced label-aware selection strategy, leading to more discriminative features.

Q set over textual label embeddings. In addition to the visual set $\mathbf{P}$, PatchCT also represents each image as a set of textual label embeddings. Inspired by the recent success of prompt learning, we perform a simple but efficient prompt: *A photo of {label}* . on each label and acquire M label sentences. One can obtain the label embeddings $\mathbf{L} = \{\boldsymbol{l}_m|_{m=1}^{M}\} \in \mathbb{R}^{d \times M}$. Due to the accessible label vector during training, it is natural to define $\boldsymbol{\beta}$ in $\mathbf{Q}$ by normalizing $\boldsymbol{y}_i$:

$$\boldsymbol{\beta}_i = softmax(\boldsymbol{y}_i).$$

This means that $\mathbf{Q}_i$ collects the ground-truth label features as the textual representation of $\boldsymbol{x}_i$. Besides, with the semantic linguistic knowledge implied in the pre-trained text encoder (*e.g., BERT*) and the designed prompt templates, the learned label embeddings have the ability to capture *1)* the textual semantics for each label, *2)* the correlations among labels, which help to improve the identification of label representations [29].

To summarize briefly, the proposed PatchCT views the multi-label classification as a CT problem and formulates each image $\boldsymbol{x}_i$ with two discrete distributions $\mathbf{P}_i$ and $\mathbf{Q}_i$ over the visual patch and textual label embeddings respectively. Those two representations share semantic consistency but with different supports. One of the core ideas behind PatchCT is to align the vision-text modalities by minimizing the bidirectional CT distance of $\mathbf{P}$ and $\mathbf{Q}$ for multi-label prediction, which will be discussed in Sec. 3.4.

Adaptive layer of visual features. Like previous works [56], we here adopt the pretrained ViT as our image encoder to obtain the local patch embeddings. To further enhance the representation of local patches and distill the pre-trained knowledge of ViT in an efficient way for multi-label classification, PatchCT introduces an adaptive module f_ϕ following the last layer of the image encoder [20], where ϕ is the learnable parameters in f_ϕ. Note that f_ϕ is a lightweight network that consists of two linear layers and aims to adapt ViT features to *new* knowledge that is more suitable for the multi-label task. Moreover, we also apply the residual connection of the adapted features and the original features encoded by the pre-trained ViT for mixing the two pieces of information. Thus the global image feature can be updated as (we still use $\boldsymbol{x}_i$ for convenience):

$$\boldsymbol{x}_i = \boldsymbol{e}_{\text{CLS}} \oplus f_\phi(\boldsymbol{e}_{\text{CLS}}), \tag{3}$$

where $\oplus$ denotes the residual connection. $\boldsymbol{x}_i$ acts as the visual feature and is used to make the multi-label prediction.

3.4. The objective function

The final objective function to minimize is simply the summation of the layer-wise CT distance that aligns the vision-text domains and the asymmetric loss for multi-label classification.

Layer-Wise CT Distance. For image $\boldsymbol{x}_i$, the two discrete distributions $\mathbf{P}_i, \mathbf{Q}_i$ are semantic representations from two different domains. PatchCT bridges the semantic gaps by minimizing the CT distance of $\mathbf{P}_i$ and $\mathbf{Q}_i$, *e.g.*, $\text{CT}(\mathbf{P}_i, \mathbf{Q}_i)$. To make deep interactions between such two modalities and achieve better alignments, we develop a Layer-wise CT distance (LCT) that greedily minimizes the CT distance at each layer. With a L-layer domain-specific encoders, the LCT can be expressed as:

$$\text{LCT}^{(L)} = \sum_{l=1}^{L} \text{CT}(\mathbf{P}_i^{(l)}, \mathbf{Q}_i),$$

where l is the index of layer, $\boldsymbol{\theta}_i^{(l)}$ is calculated according to Eq. 2 by replacing the patch embedding matrix and label-aware embedding at the corresponding layer l. The cost matrix $\mathbf{C}^{(l)}$ in Eq. 1 is computed by the cosine distance:

$$\mathbf{C}_i^{(l)} = 1 - \frac{\mathbf{E}_i^{(l)^T} \mathbf{L}^{(l)}}{||\mathbf{E}_i^{(l)}||||\mathbf{L}^{(l)}||}$$

Asymmetric loss. Given the mixed visual embedding $\boldsymbol{x}_i$ calculated by Eq. 3 and the aligned label embedding matrix at L-th layer $\mathbf{L}^{(L)}$, one can predict the category probabilities of i-th image as $\boldsymbol{p}_i = \sigma(\mathbf{L}^{(L)^T} \boldsymbol{x}_i) \in \mathbb{R}^M$, where $\sigma(\cdot)$ is the sigmoid function. To more effectively address the label imbalance issue, we adopt the asymmetric loss (ASL), which is a variant of focal loss and a valuable choice in multi-label classification tasks [35, 29]:

$$\text{ASL} = \frac{1}{M} \sum_{m=1}^{M} \begin{cases} (1 - p_{mi})^{\gamma+} \log(p_{mi}), & y_{mi} = 1, \\ (p_{mi})^{\gamma-} \log(1 - p_{mi}), & y_{mi} = 0, \end{cases}$$

where $\gamma+, \gamma-$ are two hyper-parameters for positive and negative values.

Let the learnable parameters as $\Omega = \{\text{Enc}, \phi, \psi\}$ that denotes the parameters in image and text encoders, the adaptive module, the defined transport plan in Eq. 1, respectively. Ω is optimized using stochastic gradient descent by minimizing the combined loss:

$$\mathcal{L} = \text{LCT} + \text{ASL}, \tag{4}$$

where the first term ensures the alignments between the vision and text modalities at each layer in two domain-specific encoders and the second term provides supervised information for multi-label classification.

4. Experiments

In this section, we evaluate our PatchCT with known state-of-the-art methods on three widely-used multi-label image benchmarks by reporting a series of metrics. In addition to the numerical results, comprehensive ablation and qualitative studies of the proposed model are also provided. Our code is available at https://github.com/keepgoingjkg/PatchCT.

Methods	ALL						Top-3						mAP
	CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1	
CNN-RNN [43]	-	-	-	-	-	-	66.0	55.6	60.4	69.2	66.4	67.8	61.2
ResNet-101 [22]	80.2	66.7	72.8	83.9	70.8	76.8	84.1	59.4	69.7	89.1	62.8	73.6	77.3
ML-GCN [10]	85.1	72.0	78.0	85.8	75.4	80.3	89.2	64.1	74.6	90.5	66.5	76.7	83.0
SSGRL† [8]	**89.5**	68.3	76.9	**91.2**	70.7	79.3	91.9	62.1	73.0	**93.6**	64.2	76.0	83.6
CMA [53]	82.1	73.1	77.3	83.7	76.3	79.9	87.2	64.6	74.2	89.1	66.7	76.3	83.4
TSGCN [47]	81.5	72.3	76.7	84.9	75.3	79.8	84.1	67.1	74.6	89.5	69.3	69.3	83.5
MulCon [16]	-	-	78.6	-	-	81.0	-	-	-	-	-	-	84.0
C-Tran† [26]	86.3	74.3	79.9	87.7	76.5	81.7	90.1	65.7	76.0	92.1	**71.4**	77.6	85.1
ADD-GCN [52]	84.7	75.9	80.1	84.9	79.4	82.0	88.8	66.2	75.8	90.3	68.5	77.9	85.2
ASL [36]	87.2	76.4	81.4	88.2	79.2	81.8	91.8	63.4	75.1	92.9	66.4	77.4	86.6
CSRA [61]	89.1	74.2	81.0	89.6	77.1	82.9	**92.5**	65.8	76.9	93.4	68.1	78.8	86.9
Q2L [29]	87.6	76.5	81.6	88.4	78.5	83.1	91.9	66.2	77.0	93.5	67.6	78.5	87.3
M3TR [56]	88.4	77.2	82.5	88.3	79.8	83.8	91.9	68.1	78.2	92.6	69.6	79.4	87.5
PatchCT	83.3	**82.3**	**82.6**	84.2	**83.7**	**83.8**	90.7	**69.7**	**78.8**	90.3	70.8	**79.8**	**88.3**

Table 1: Comparison of PatchCT and known SOTA models on MS-COCO dataset under the settings of all and top-3 labels. All metrics all in %. The symbol † means using a larger input image resolution (576 × 576).

Datasets	# Train	# Test	# Class	# Object
MS-COCO	82,081	40,504	80	2.9
VOC 2007	5,011	4952	20	2.5
NUS-WIDE	125,449	83898	81	2.4

Table 2: Statistics of the used datasets. # Object denotes the average object labels per image.

4.1. Datasets and Evaluation Metrics

Datasets. We conduct extensive experiments on three popular image datasets, including **MS-COCO** [28], **PASCAL VOC 2007** [18], and **NUS-WIDE** [12]. MS-COCO is a commonly-used benchmark to evaluate the multi-label image classification task. It contains 82,081 images as the training set and 40,504 images as the validation set. The objects are categorized into 80 classes with about 2.9 object labels per image. Pascal VOC 2007 contains 20 categories and total 9,963 images, in which 5,011 images form the train-val set, and the remaining 4,952 images are taken as the test set for evaluation. For fair comparisons, the current competitors and our model are all trained on the train-val set and evaluated on the test set. NUS-WIDE is a real-world web image dataset with 269,648 images and 5018 labels from Flickr. These images are further manually annotated with 81 visual concepts. After removing the no-annotated images, the training set and test set contain 125,449 and 83,898 images respectively. We summarize the statistics of datasets at Table. 2.

Evaluation Metrics. To comprehensively evaluate the performance, we follow previous works [43, 9, 29] and report the mean average precision(mAP), the average per-class precision (CP), recall (CR), F1 (CF1), and the average overall precision(OP), recall(OR), F1(OF1). Besides, we also report the results of top-3 labels. For all metrics, higher values indicate better performance, and in general, mAP is the most important metric. For each image, labels are identified as positive if their predicted probabilities are greater than 0.5. All metrics are reported as the mean results of five runs with different random seeds.

4.2. Implementation Details

Like previous works, PatchCT employs the ViT-B16 that pretrained on ImageNet21k as the image encoder and loads the 12-layer BERT trained on Wikipedia data as the text encoder, and the hidden dimension in both encoders is $d = 512$. For a fair comparison with competitors, we resize all images to $H \times W = 448 \times 448$ as input resolution in both training and testing phases throughout all experiments. The number of top-k patches is set as $k = 200$. We set $\gamma+ = 0$ and $\gamma- = 2$ in the asymmetric loss. The optimization of PatchCT is done by AdamW with a learning rate of $1e - 5$, True-weight-decay $1e - 2$, and batch size 12 for maximally 40 epochs. All experiments are performed on one Nvidia RTX-3090Ti GPU and our model is implemented in PyTorch. Please refer to the attached code for more details.

4.3. Comparison with State-of-the-art Methods

In order to demonstrate the effectiveness of the proposed framework, we compare PatchCT with a number of state-of-the-art methods from the literature, including conventional CNN-based methods [43, 22, 54], graph-based methods [10, 8, 47], and attention-based methods [61, 11, 26, 29, 56]. The numbers of the compared methods are taken from the best-reported results in their papers (We report results of Q2L with TResNet-L to keep close compute and numbers of parameters fairly with the image encoder of

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
CNN-RNN [43]	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0	92.4	91.7	84.2	93.7	59.8	93.2	75.3	99.7	78.6	84.0
RLSD [54]	96.4	92.7	93.8	94.1	71.2	92.5	94.2	95.7	74.3	90.0	74.2	95.4	96.2	92.1	97.9	66.9	93.5	73.7	97.5	87.6	88.5
HCP [46]	98.6	97.1	98.0	95.6	75.3	94.7	95.8	97.3	73.1	90.2	80.0	97.3	96.1	94.9	96.3	78.3	94.7	76.2	97.9	91.5	90.9
RDAR [44]	98.6	97.4	96.3	96.2	75.2	92.4	96.5	97.1	76.5	92.0	87.7	96.8	97.5	93.8	98.5	81.6	93.7	82.8	98.6	89.3	91.9
RARL [7]	98.6	97.1	97.1	95.5	75.6	92.8	96.8	97.3	78.3	92.2	87.6	96.9	96.5	93.6	98.5	81.6	93.1	83.2	98.5	89.3	92.0
SSGRL† [8]	99.5	97.1	97.6	97.8	82.6	94.8	96.7	98.1	78.0	97.0	85.6	97.8	98.3	96.4	98.8	84.9	96.5	79.8	98.4	92.8	93.4
ML-GCN [10]	99.5	98.5	98.6	98.1	80.8	94.6	97.2	98.2	82.3	95.7	86.4	98.2	98.4	96.7	99.0	84.7	96.7	84.3	98.9	93.7	94.0
TSGCN [47]	98.9	98.5	96.8	97.3	87.5	94.2	97.4	97.7	84.1	92.6	89.3	98.4	98.0	96.1	98.7	84.9	96.6	87.2	98.4	93.7	94.3
ASL [36]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	94.6
CSRA [61]	99.9	98.4	98.1	98.9	82.2	95.3	97.8	97.9	84.6	94.8	90.8	98.1	97.6	96.2	99.1	86.4	95.9	88.3	98.9	94.4	94.7
MlTr-1 [11]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	95.8
Q2L [29]	99.9	98.9	99.0	98.4	**87.7**	98.6	98.8	99.1	84.5	98.3	89.2	99.2	99.2	**99.2**	**99.3**	90.2	98.8	88.3	99.5	95.5	96.1
M3TR [56]	99.9	99.3	**99.1**	99.1	84.0	97.6	98.0	99.0	85.9	**99.4**	93.9	**99.5**	99.4	98.5	99.2	90.3	**99.7**	91.6	**99.8**	96.0	96.5
PatchCT	**100.0**	**99.4**	98.8	**99.3**	87.2	**98.6**	**98.8**	**99.2**	**87.2**	99.0	**95.5**	99.4	**99.7**	98.9	99.1	**91.8**	99.5	**94.5**	99.5	**96.3**	**97.1**

Table 3: Results on Pascal VOC 2007 dataset in terms of class-wise precision (AP) and mean average precision (mAP) in %.

Methods	ALL		Top-3		mAP
	CF1	OF1	CF1	OF1	
CNN-RNN [43]	-	-	34.7	55.2	56.1
ResNet-101 [22]	51.9	69.5	56.8	69.1	59.8
CMA [53]	60.5	73.7	55.5	70.0	61.4
MlTr-1 [11]	65.0	**75.8**	-	-	66.3
MulCon [16]	59.0	73.8	-	-	62.5
ASL [36]	63.6	75.0	-	-	65.2
Q2L [29]	64.0	75.0	-	-	66.3
PatchCT	**65.5**	74.7	**61.2**	**71.0**	**68.1**

Table 4: Results on the NUS-WIDE dataset under the setting of all and top-3 labels. All metrics in %.

PatchCT). Table.(1-4) report the comparisons of PatchCT with those SOTA methods on the three multi-label image datasets. We have the following remarks about the numerical results: *1)* Overall, our proposed PatchCT achieves the best mAP scores on all datasets. Despite scoring high on some metrics, other approaches often either fail to balance accuracy and recall or fail to achieve stable performance on all categories (*e.g.*, chair, table, horse). *2)* Compared to models that leverage additional information of label dependency [9, 10, 47], PatchCT accomplishes significant improvements in most metrics. This suggests the efficiency of the employed pre-trained language model that contains rich semantic knowledge and shows great potential to capture label correlations. Besides, the AP scores of our PatchCT on all 20 categories in Pascal VOC 2007 dataset are above 87.2, and PatchCT exceeds 97.1 mAP on the test set. This demonstrates that the text encoder can also provide discriminative label representations. *3)* Furthermore, developed from the similar motivation that aligns the vision-text modalities for the multi-label task, our PatchCT shows a relatively stable and superior performance compared to recent attention-based models [29, 56, 26, 11] that adopt the cross-attention mechanism to explore the interactions between those two domains. We attribute this success to the sparsity layer-wise CT distance. Further, LCT provides an efficient option to align the semantics of patches and labels progressively by minimizing the transport cost from a bidirectional view. Moreover, we also visualize the learned transport probabilities at Sec. 4.5 which indicates the interpretability of PatchCT.

4.4. Further Analyses

To fully understand the introduced CT module in PatchCT, we perform a series of further analyses in this section, Specifically, we are interested in the following variants.

Effects of the CT loss. We treat the LCT distance and the asymmetric loss equally in previous experiments for convenience. Here we adopt a weighted combined loss: $\mathcal{L}_\alpha = \alpha\text{LCT} + \text{ASL}$, where α controls the weight of the LCT. We report the mAP of PatchCT under the setting of $\alpha = [0 : 0.2 : 1]$ on MS-COCO and VOC 2007 datasets at the first two subfigures of Fig. 3. Note that, $\alpha = 0$ means we remove LCT from the combined loss, which results in our base model without CT alignment. The results of this variant are also reported at Table 5 for a clear comparison. We find that The introduced LCT plays a pivotal role in PatchCT and improves the performance by a significant margin, which demonstrates the efficiency of the vision-text alignment by minimizing the bidirectional transport cost.

Effects of the starting layer in LCT. Due to its irreplaceable role in PatchCT, we further explore the effects of the starting layer l_s in LCT and report the mAP scores on MS-COCO and VOC 2007 with the setting $l_s = [1 : 2 : 11]$ at the last two subfigures of Fig. 3. We have the following interesting findings. Overall, aligning the vision-text
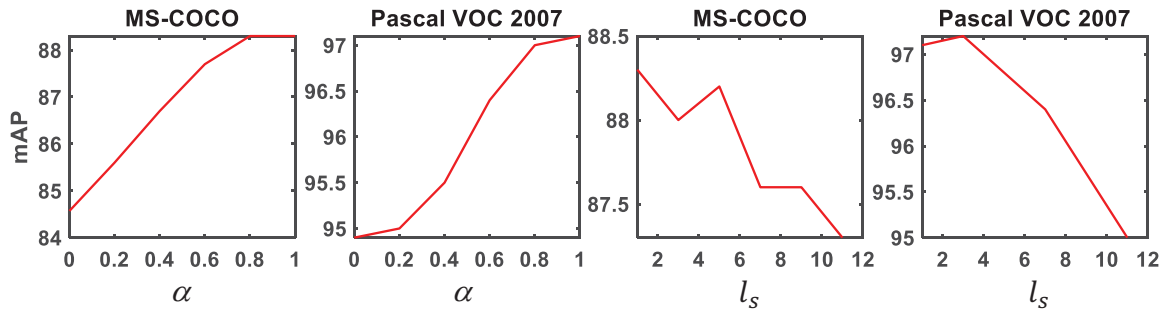
Figure 3: Parameter sensitivity of PatchCT in terms of α that controls the weights of LCT in the combined loss (the first two subfigures) and the starting layer of the CT alignments l_s (the last two subfigures) on MS-COCO and Pascal VOC 2007.
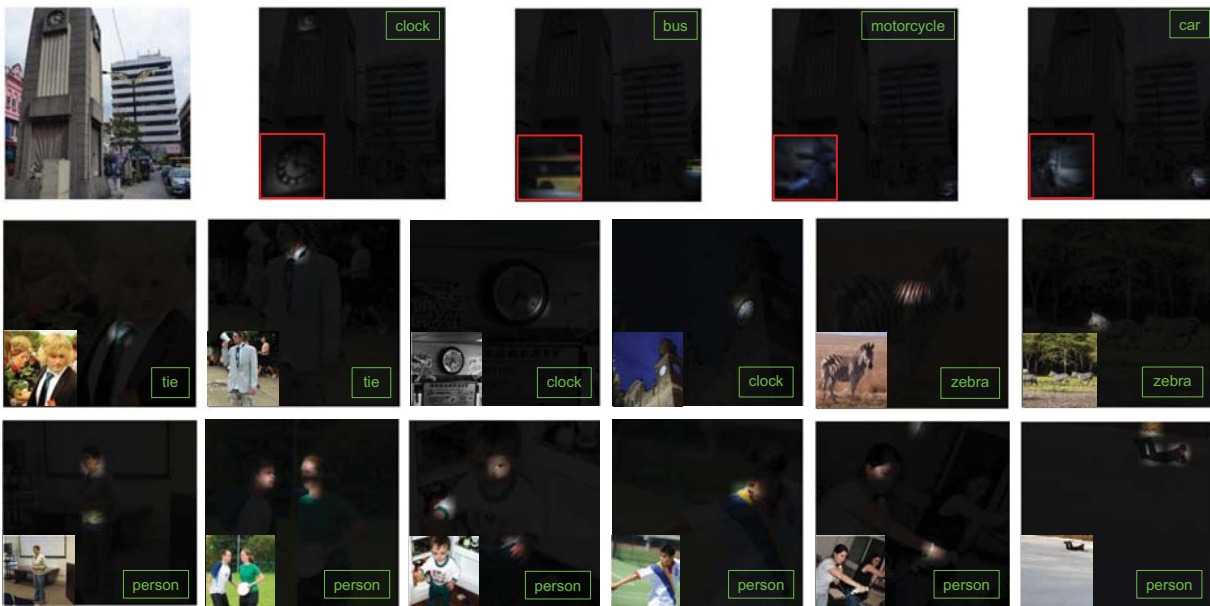

Figure 4: Visualization of the learned backward transport probabilities. The top row includes the input image and the different label queries. The activated areas are zoomed in for clear display (red boxes). The bottom two rows show activated patches in different images for labels "tie", "clock", "zebra", and "person". The raw images are included in the bottom left corner.

modalities at the early stages (*e.g.* $l_s < 6$) often produces higher performance, which again verifies our motivation that effective alignment is crucial for multi-label classification tasks. Besides, aligning from $l_s = 1$ may not give the best results as expected. This may be due to the fact that vision and text require specific modality information at the very early layers.

Effects of TopK$(\cdot, k)$ Length. In our previous experiments, we fix the hyperparameter $k = 200$ in Eq. 2 for all images and find it works well in most cases. To further explore whether our model is sensitive to k, we report the ablation results with various k on Pascal VOC 2007 at Table. 6. We observe that PatchCT is robust to k, and one can obtain even better results after finetuning k.

Datasets	MS-COCO	Pascal VOC 2007	NUS-WIDE
PatchCT(label)	88.0	96.7	68.0
PatchCT(w/o CT)	84.7	94.9	65.8
PatchCT	88.3	97.1	68.1

Table 5: Ablation results (mAP) of prompt learning and CT alignment. PatchCT(label) denotes the variant that extracts label embedding only using label name, and Patchct (w/o CT) denotes the variant that learns parameters without the CT regularization.

Effects of the Prompt Learning. In general, prompt learning can help bridge the distribution gap between training and inference. Empirically, CLIP finds that the simple prompt "*A photo of {label}* ." often improves performance over the baseline of using only the label text (improves accuracy on

k	50	100	150	200	250
mAP	96.8	97.2	96.9	97.1	97.0

Table 6: Ablation on k length in $\text{TopK}(\cdot, k)$ on Pascal VOC 2007 dataset.

ImageNet by 1.3%). To evaluate the efficiency of prompt learning for multi-label classification, we here report the results of the variant that extracts the textual features only using the label name at Table. 5. Compared to utilizing contextless class names, prompt learning boost multi-label classification performance on three datasets.

4.5. Qualitative Analysis

Another main property of PatchCT is the interpretability brought by the learned transport probabilities, which provide a visualization tool to understand the interactions between labels and patches. Recall that m-th column of backward transport plan $\overleftarrow{\mathbf{T}}$ in Eq. 1 means the transport probabilities from the m-th label to a set of patches. For a backward transport plan of interest $\overleftarrow{\mathbf{T}}_m \in \mathbb{R}^N$, we first normalize it at the range of [0,1], and then reshape it to the grid matrix $\overleftarrow{\mathbf{T}}_m \in \mathbb{R}^{g \times g}$, where $g = \sqrt{N}$ is the number of grids. To visualize the transport plan, we resize the grid matrix to the same size as the raw image via the bicubic interpolation and highlight the most related patches according to the transport probabilities. The results are shown at Fig. 4 in the following cases: the same image with different labels (the top row), and the same query label with different images (the bottom two rows). From the top row, we find that for a given image, the label queries successfully retrieve the patches that cover the corresponding objects and the transport plan can precisely highlight the main regions of that object. We are also interested in the transport plans in different images. From the middle row, we can see that the transport plans have the ability to adjust the most related regions according to contextual images dynamically. For example, for a given label *zebra*, it tends to focus on the stripes in a close-up image and more on the face in the large background image. We also visualize the top-3 related patches for the *person* query at the bottom row and find that the patches tend to capture the multiple regions of "person" in the image. This improves the flexibility and robustness of PatchCT and widens its applications.

5. Conclusion

In this paper, we reformulate the multi-label image classification as a conditional transport problem and introduce PatchCT for aligning the vision and text modalities under the CT framework, where each image is represented as two discrete distributions over the patches and labels respectively. PatchCT is optimized by minimizing the combined loss that consists of the widely used asymmetric loss and the layer-wise CT distance end-to-end. Extensive experiments on three datasets consistently demonstrate the superiority of the proposed PatchCT. Two ablation studies and visualizations confirm our motivation and the core role of the introduced CT. Since its natural flexibility and simplicity, we hope PatchCT provides innovative ideas for follow-up studies.

References

[1] Junwen Bai, Shufeng Kong, and Carla P Gomes. Gaussian mixture variational autoencoder with contrastive learning for multi-label classification. In *International Conference on Machine Learning*, pages 1383–1398. PMLR, 2022. 1, 2

[2] Wei Bi and James Kwok. Multilabel classification with label correlations and missing labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[4] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 3

[5] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pages 1542–1553, 2020. 2, 3

[6] Shang-Fu Chen, Yi-Chen Chen, Chih-Kuan Yeh, and Yu-Chiang Wang. Order-free rnn with visual attention for multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[7] Tianshui Chen, Zhouxia Wang, Guanbin Li, and Liang Lin. Recurrent attentional reinforcement learning for multi-label image recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 7

[8] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 522–531, 2019. 2, 6, 7

[9] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5177–5186, 2019. 2, 6, 7

[10] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5177–5186, 2019. 6, 7

[11] Xing Cheng, Hezheng Lin, Xiangyu Wu, Dong Shen, Fan Yang, Honglin Liu, and Nian Shi. Mltr: Multi-label classification with transformer. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022. 2, 6, 7

[12] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009. 6

[13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 3

[14] Son D Dao, Ethan Zhao, Dinh Phung, and Jianfei Cai. Multi-label image classification with contrastive learning. *arXiv preprint arXiv:2107.11626*, 2021. 2

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2

[16] Son D Dao Ethan Zhao Dinh and Phung Jianfei Cai. Contrast learning visual attention for multi label classification. 6, 7

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[18] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 6

[19] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. *Advances in neural information processing systems*, 28, 2015. 3

[20] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *CoRR*, abs/2110.04544. 5

[21] Zongyuan Ge, Dwarikanath Mahapatra, Suman Sedai, Rahil Garnavi, and Rajib Chakravorty. Chest x-rays classification: A multi-label and fine-grained problem. *arXiv preprint arXiv:1807.07247*, 2018. 1

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 6, 7

[23] Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8776–8786, 2020. 2

[24] Himanshu Jain, Yashoteja Prabhu, and Manik Varma. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 935–944, 2016. 1

[25] Jack Lanchantin, Arshdeep Sekhon, and Yanjun Qi. Neural message passing for multi-label classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 138–163. Springer, 2019. 2

[26] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16478–16488, 2021. 2, 3, 6, 7

[27] John Lee, Max Dabagia, Eva Dyer, and Christopher Rozell. Hierarchical optimal transport for multimodal distribution alignment. *Advances in neural information processing systems*, 32, 2019. 3

[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. 6

[29] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021. 2, 3, 5, 6, 7

[30] Xinyang Liu, Dongsheng Wang, Miaoge Li, Zhibin Duan, Yishi Xu, Bo Chen, and Mingyuan Zhou. Patch-token aligned bayesian prompt learning for vision-language models. *arXiv preprint arXiv:2303.09100*, 2023. 3

[31] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *arXiv preprint arXiv:2103.10385*, 2021. 2

[32] Shweta Mahajan, Teresa Botschen, Iryna Gurevych, and Stefan Roth. Joint wasserstein autoencoders for aligning multimodal embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[33] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2

[34] 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. 3

[35] Tal Ridnik, Emanuel Ben Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. 5

[36] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor.

Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91, 2021. 6, 7

[37] Jing Shao, Kai Kang, Chen Change Loy, and Xiaogang Wang. Deeply learned attributes for crowded scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4657–4666, 2015. 1

[38] Korawat Tanwisuth, Xinjie Fan, Huangjie Zheng, Shujian Zhang, Hao Zhang, Bo Chen, and Mingyuan Zhou. A prototype-oriented framework for unsupervised domain adaptation. In *NeurIPS 2021: Neural Information Processing Systems*, Dec. 2021. 2, 3

[39] Korawat Tanwisuth, Shujian Zhang, Huangjie Zheng, Pengcheng He, and Mingyuan Zhou. POUF: Prompt-oriented unsupervised fine-tuning for large pre-trained models. In *ICML 2023: International Conference on Machine Learning*, July 2013. 2, 3

[40] Long Tian, Jingyi Feng, Wenchao Chen, Xiaoqiang Chai, Liming Wang, Xiyang Liu, and Bo Chen. Prototypes-oriented transductive few-shot learning with conditional transport, 2023. 2

[41] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009. 3

[42] Dongsheng Wang, Dandan Guo, He Zhao, Huangjie Zheng, Korawat Tanwisuth, Bo Chen, and Mingyuan Zhou. Representing mixtures of word embeddings with mixtures of topic embeddings. In *ICLR 2022: International Conference on Learning Representations*, 2022. 2, 3

[43] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2016. 2, 6, 7

[44] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *Proceedings of the IEEE international conference on computer vision*, pages 464–472, 2017. 7

[45] Shikui Wei, Lixin Liao, Jia Li, Qinjie Zheng, Fei Yang, and Yao Zhao. Saliency inside: learning attentive cnns for content-based image retrieval. *IEEE Transactions on image processing*, 28(9):4580–4593, 2019. 1

[46] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Hcp: A flexible cnn framework for multi-label image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1901–1907, 2015. 7

[47] Jiahao Xu, Hongda Tian, Zhiyong Wang, Yang Wang, Wenxiong Kang, and Fang Chen. Joint input and output space learning for multi-label image classification. *IEEE Transactions on Multimedia*, 23:1696–1707, 2020. 6, 7

[48] Xitong Yang, Yuncheng Li, and Jiebo Luo. Pinterest board recommendation for twitter users. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 963–966, 2015. 1

[49] Yang Yang, Yi-Feng Wu, De-Chuan Zhan, Zhi-Bin Liu, and Yuan Jiang. Complex object classification: A multi-modal

multi-instance multi-label deep network with optimal transport. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2594–2603, 2018. 3

[50] Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartlomiej Twardowski, and Joost van de Weijer. Orderless recurrent models for multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13440–13449, 2020. 2

[51] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXI*. 2

[52] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *European conference on computer vision*, pages 649–665. Springer, 2020. 6

[53] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. Cross-modality attention with semantic graph embedding for multi-label classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12709–12716, 2020. 6, 7

[54] Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, and Jianfeng Lu. Multilabel image classification with regional latent semantic dependencies. *IEEE Transactions on Multimedia*, 20(10):2801–2813, 2018. 6, 7

[55] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013. 2

[56] Jiawei Zhao, Yifan Zhao, and Jia Li. M3tr: Multi-modal multi-label recognition with transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 469–477, 2021. 2, 3, 5, 6, 7

[57] Peng Zhao and Zhi-Hua Zhou. Label distribution learning by optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 3

[58] Huangjie Zheng, Xu Chen, Jiangchao Yao, Hongxia Yang, Chunyuan Li, Ya Zhang, Hao Zhang, Ivor Tsang, Jingren Zhou, and Mingyuan Zhou. Contrastive conditional transport for representation learning. *arXiv preprint arXiv:2105.03746*, 2021. 3

[59] Huangjie Zheng and Mingyuan Zhou. Exploiting chain rule and bayes' theorem to compare probability distributions. *Advances in Neural Information Processing Systems*, 34:14993–15006, 2021. 2, 3

[60] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022. 2

[61] Ke Zhu and Jianxin Wu. Residual attention: A simple but effective method for multi-label recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 184–193, 2021. 6, 7

[62] Xuelin Zhu, Jiuxin Cao, Jiawei Ge, Weijia Liu, and Bo Liu. Two-stream transformer for multi-label image classification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3598–3607, 2022. 2