Visual Prompting via Partial Optimal Transport

Mengyu Zheng[1,2], Zhiwei Hao[2,3], Yehui Tang[2], and Chang Xu[1⋆]

[1] School of Computer Science, Faculty of Engineering, The University of Sydney
[2] Huawei Noah's Ark Lab
[3] School of information and Electronics, Beijing Institute of Technology
mzhe4259@uni.sydney.edu.au, haozhw@bit.edu.cn, yehui.tang@huawei.com, c.xu@sydney.edu.au

Abstract. Visual prompts represent a lightweight approach that adapts pre-trained models to downstream tasks without modifying the model parameters. They strategically transform the input and output through prompt engineering and label mapping, respectively. Yet, existing methodologies often overlook the synergy between these components, leaving the intricate relationship between them underexplored. To address this, we propose an **O**ptimal **T**ransport-based **L**abel **M**apping strategy (**OTLM**) that effectively reduces distribution migration and lessens the modifications required by the visual prompts. Specifically, we reconceptualize label mapping as a partial optimal transport problem, and introduce a novel transport cost matrix. Through the optimal transport framework, we establish a connection between output-side label mapping and input-side visual prompting. Additionally, we analyze frequency-based label mapping methods within this framework. We also offer an analysis of frequency-based label mapping techniques and demonstrate the superiority of our OTLM method. Our experiments across multiple datasets and various model architectures demonstrate significant performance improvements, which prove the effectiveness of the proposed method.

Keywords: Label mapping · Visual Prompting · Optimal transport

1 Introduction

As the amount of data and computational resources increases, large models trained on extensive datasets are able to effectively capture general features [10, 18], which present impressive generalization ability on adapting to other datasets. Following this, the 'pretraining and finetuning' paradigm—finetuning a pre-trained model on unseen datasets— has achieved notable success and been with widespread application [5]. However, this approach encounters certain limitations in practical applications. As pre-trained models are often typically large (e.g. 632 million parameters in ViT-Huge), there is nearly no chance to deploy several models fine-tuned on different datasets for a variety of scenarios, which would involve unacceptable memory requirement for storing these models with

⋆ Corresponding author

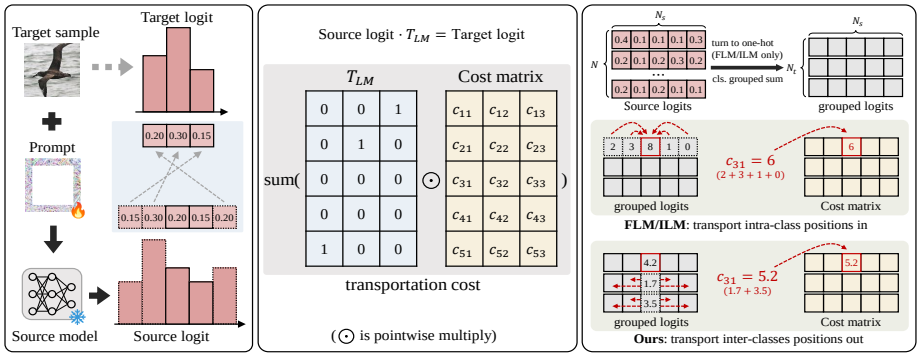
large volume. Moreover, fully finetuning them for each task incurs extensive computational costs, expanding to a broader range of downstream tasks [39].

Drawing inspiration from the outstanding performance of prompting in the NLP domain, Visual Prompting (VP) is proposed to solve this problem [3]. VP reinterprets the problem of adapting a frozen, pre-trained model to an unseen dataset as a data-space adaptation by transforming both the input and the output. Specifically, a small amount of learnable perturbations is introduced into the input. On the output side, label mapping provides a function to map source labels to target labels. This strategy introduces two significant advantages. Firstly, VP considerably reduces the number of parameters requiring tuning. More importantly, it maintains the pre-trained model unchanged, enabling various scenarios to share the same pre-trained model identically. Consequently, this leads to a substantial reduction in both memory requirements and computational expenses.

In visual prompting, only a limited number of learnable parameters are introduced, in the form of prompts that are incorporated with images before the first layer. Therefore, how to efficiently utilize the adaptation capability provided by prompts is a research topic worth exploring. Some approaches aim to enhance the adaptation capability of prompts through input transformation. For instance, DAM-VP [25] offers unique prompts for each subset rather than a universal prompt for all images, aiming to address the challenges posed by highly diverse downstream datasets. Nonetheless, this method necessitates additional training data for clustering and leads to an increase in the number of parameters.

While significant attention has been dedicated to innovations at the input, the transformation applied to the output plays an equally pivotal role within the VP framework. Often referred to as label mapping, this transformation is crucial for aligning the pre-trained model's source domain labels with the target domain's labels. Label mapping serves to correlate the model's source domain classification outcomes with meaningful labels of the target domain. Several methodologies for label mapping have been explored. For example, random-based label mapping is employed in Prom-PATE [32]. Hard-coded label mapping is utilized in the output transformation of VP [3], simply hard-coding the first n outputs of the model layer directly to the task data categories. Apparently, random-based methods do not consider the rationality of the mapping at all. Such random methods completely discard the information of labels. Another kind of label mapping method is frequency-based label mapping [9, 11, 44], which relies on the source model's prediction frequencies for target data points. Furthermore, ILM [9] incorporates the relationships of label mapping as trainable parameters into its training process, aiming to discover interpretable label mapping results. These methods just heuristically rely on the voting of samples to produce mapping results.

The selection of label mapping strategies significantly impacts the effectiveness of visual prompting. We argue that the ideal selection strategy for label mapping should aim to minimize the necessary adjustments a model must make to align its outputs with the target labels, thereby enhancing the efficiency and performance of prompt optimization. A carefully designed label mapping can significantly reduce the alterations required to adjust the output distribution to



(a) Visual prompting (b) POT framework (c) Cost matrix design

Fig. 1: Overview of OTLM: (a) VP pipeline entails transforming input by incorporating prompts with images and transforming output by mapping source labels to target labels; (b) Label mapping within the Partial Optimal Transportation (POT) framework involves computing the transportation cost as the sum of the pointwise multiplication between the mapping matrix and the cost matrix; (C) The design of the cost matrix involves obtaining predictions of the entire target dataset and then aggregating them by class through grouped summation. In FLM/ILM, the cost at each position is defined by transporting the intra-class distribution into the selected position, while OTLM defines it by transporting the inter-class distribution outside the selected position.

match the target labels. Less modification implies a more efficient utilization of prompt optimization capabilities, leading to improved outcomes. Existing label mapping methods have not enhanced the effectiveness from this perspective.

In this paper, we introduce the concept of minimal transfer load for prompt design and define label mapping as a type of partial optimal transport problem, where the transfer load is delineated through distribution migration. To solve this optimal transport problem, we propose a novel cost matrix that defines the mapping cost between labels from the perspective of transfer load. Then, we derive a label mapping strategy **OTLM** (**O**ptimal **T**ransport-based **L**abel **M**apping) that minimizes the transfer load of output distribution (shown in Figure 1). Moreover, we analyze the existing frequency-based label mapping method from the optimal transport perspective, providing insights into the advantages of our approach. The results of the experiment, conducted across multiple datasets and model architectures, showcase the substantial performance improvements brought about by the OTLM. Additionally, we conduct interpretative experiments to understand the effectiveness of the proposed method.

2 Related Work

Natural language prompting and visual prompting. Prompt-based learning in natural language processing is a new paradigm that adapting frozen, pre-trained language models to diverse downstream scenarios [34,36]. Inspired by the

purpose of prompting that adapting a frozen pre-trained model by modifying the data space, Visual Prompting [3] designs a specific prompt for each task and applies it to the input. In addition, Visual-Prompting Tuning [26] are specially proposed for Transformer pre-trained model, adding learnable parameters into Transformer's first or every input layer. The new paradigm "prompting-based learning" in natural language processing and vision enables pre-trained models to be easily applied to various downstream tasks while consuming minimal memory space and computational resources [42, 43].

Label mapping for visual prompting. Label mapping (LM) are applied in various machine learning tasks such as transfer learning [41, 48], domain adaptation [30, 31, 49], and instance segmentation [24]. Unlike the aforementioned tasks where LM is integrated into the model, trainable parameters are only present in the input, and the entire pre-trained model is inaccessible in visual prompting [3]. Consequently, LM has to independently assign source labels to target labels. The existing label mapping methods in visual prompting are not yet mature. Firstly, random-based label mapping strategies are employed in visual prompt methods [3, 25, 32]. In addition, frequency-based mapping [9, 11, 44] maps target labels to source labels by the prediction frequencies of pre-trained model on the target dataset. These methods overlook the potential to alleviate the training burden on input prompts during label mapping. In contrast, our proposed method minimizes the cost when mapping labels.

Optimal transport in machine learning. Optimal Transport (OT) is a mathematical framework of seeking the most cost-efficient way to transport mass from one distribution to another [46]. After the emergence of new methods addressing high computational cost issues [4, 6], OT has found extensive application across various domains in machine learning, such like Generative Adversarial Networks [2, 28, 40], diffusion models [20, 33], and domain adaptation [14–16, 21]. WGAN [2] replaces the KL divergence in traditional GANs with the Wasserstein distance. To the best of our knowledge, Optimal Transport has not yet been employed to tackle the Visual Prompting challenge.

3 Preliminaries

Visual Prompting (VP) Framework. We formally introduce the concepts of Visual Prompting (VP) and Label Mapping (LM), which lay the foundational framework for our proposed methodology.

Visual prompting employs a pre-trained model $\mathcal{M}$, with parameters θ_s, initially trained on a source dataset $\mathcal{D}_S$ with labels $\mathcal{Y}_S \subseteq \mathbb{R}^{n_s}$, where n_s is the number of class of source dataset. VP aims to adapt $\mathcal{M}$ to a target task defined by a new dataset $\mathcal{D}_T$ with labels $\mathcal{Y}_T$, without modifying θ_s. This is achieved through the introduction of the learnable prompt θ_p, which is optimized to manipulate the input data that satisfies the target task performance. Given a target data point $\mathbf{x}_t \in \mathcal{D}_T$, the prompted input $\mathbf{x}'((\theta_p))$ is generated as:

$$\mathbf{x}'(\theta_p) = h(\mathbf{x}_t, \theta_p), \tag{1}$$

where $h(\cdot, \theta_p)$ represents the transformation function incorporating the prompt θ_p into the target input $\mathbf{x}_t$. The objective is to optimize θ_p such that $\mathcal{M}(\mathbf{x}')$ aligns with the target labels $y_t \in \mathcal{Y}_T$, effectively reprogramming $\mathcal{M}$ for the target task.

Label mapping is a critical component of VP, facilitating the translation of model predictions from the source label space $\mathcal{Y}_S$ to the target label space $\mathcal{Y}_T$. The LM progress is to find a mapping function $\mathcal{F}_{LM} : \mathcal{Y}_S \to \mathcal{Y}_T$ based on specific principles. With the help of label mapping function $\mathcal{F}_{LM}$, the optimization paradigm can be established. The optimization objective can be expressed as:

$$\min_{\theta_p} \mathbb{E}_{\{\mathbf{x}_t, y_t \in \mathcal{D}_T\}}[\mathcal{L}\left(\mathcal{F}_{LM}(\mathcal{M}(h(\mathbf{x}_t, \theta_p))), y_t\right)], \tag{2}$$

where $\mathcal{L}(\cdot)$ denotes a loss function measuring the discrepancy between the mapped predictions $\mathcal{F}_{LM}(\mathcal{M}(\mathbf{x}'(\theta_p)))$ and the true target labels $y_t \in \mathcal{Y}_T$. This optimization is performed only on the prompt parameters θ_p, leaving the original model parameters $\theta_s \in \mathcal{M}$ untouched.

4 Method

The optimization burden hinges critically on the disparity between the initial and final states of the feature distributions influenced by visual prompts. It is imperative to recognize that the initial state of these features is solely determined by the intrinsic characteristics of the data and the model architecture. Consequently, for a given target data distribution, the magnitude of the optimization burden is directly influenced by the final state to which the system aspires. As delineated in the preceding sections, this final state is fundamentally contingent upon the label mapping function $\mathcal{F}_{LM}$, which governs the transition of prompted output distributions $\hat{f}_i$ towards the desired targets y_i.

Given the pivotal role of $\mathcal{F}_{LM}$ in determining the optimization burden, it is logical to explore methodologies for selecting or designing an appropriate mapping function that can significantly alleviate this burden. To this end, we propose an approach grounded in the principles of optimal transport to conceptualize label mapping.

4.1 Label Mapping from an Optimal Transport Perspective

Our objective is to devise a label mapping strategy, denoted as $\mathcal{F}_{LM}$, aimed at substantially minimizing the necessity for feature transfer. In this framework, a prompt θ_p is meticulously optimized to adjust $f_i = \mathcal{M}(\mathbf{x}'_i(\theta_p)) \in \mathbb{R}^{n_s}$, ensuring f_i closely mirrors the target feature $\hat{f}_i \in \mathbb{R}^{n_t}$, *i.e.*, the one-hot representation of label $y_i \in \mathcal{Y}_T$. The selection of a label mapping strategy introduces variable levels of operational challenge for the prompt, emphasizing that the optimal label mapping is one that requires the least deviation in feature distribution between the source and target. To achieve this, we encapsulate the concept of label mapping within an optimal transport, setting the stage for a strategic

transformation that aligns with our goal of efficiency and precision in feature adaptation.

Considering that label mapping facilitates the transition from the source label distribution $\mathbf{p}$ with n_s classes to the target label distribution $\mathbf{q}$ with n_t classes. $n_s > n_t$ is introduced as a hypothesis. We leverage empirical sampling from input examples $\mathbf{x}'$ to define these distributions as $(\mathbf{p}, \mathbf{q}) \in \Sigma_{n_s} \times \Sigma_{n_t}$:

$$\mathbf{p} = \sum_{j=1}^{n_s} p_j \delta_{f_{,j}}, \quad \mathbf{q} = \sum_{j=1}^{n_t} q_j \delta_{\hat{f}_{,j}}, \tag{3}$$

where $\delta_{f_{,j}}$ is the Dirac function at location $f_{,j} \in \mathbb{R}^N$, with N being the total number of samples in the target dataset $\mathcal{D}_T$ and $f_{i,j}$ signifying the j-th component of feature $f_i \in \mathbb{R}^{n_s}$. Σ_{n_s} and Σ_{n_t} correspond to histograms with n_s and n_t bins, respectively. p_j and q_j are the probability mass associated to the j-th element of feature f_i. In the VP-LM scenario, each target label correlates with a specific source label, disregarding the non-selected source labels. For sample-balance tasks, we adopt $p_j = q_j = \frac{1}{n_s}$, which leads that $\sum_j p_j = 1$ and $\sum_j q_j = \frac{n_t}{n_s}$. As a result, label mapping can be defined as a Partial Optimal Transport (POT), which extends the classical Optimal Transport (OT) concept to accommodate for the transfer of mass between distributions of differing total masses [12], aptly suiting the label mapping context.

The set of partial transportation plans $\Pi(\mathbf{p}, \mathbf{q})$ between two discrete probability measures $\mathbf{p}$ and $\mathbf{q}$ is,

$$\Pi(\mathbf{p}, \mathbf{q}) = \{\mathbf{T} \in \mathbb{R}_+^{n_s \times n_t} | \mathbf{T}\mathbb{1}_{n_t} \leq \mathbf{p}, \mathbf{T}^\top \mathbb{1}_{n_s} = \mathbf{q}\}, \tag{4}$$

where $\mathbb{1}_d$ is a d-dimensional vector of ones, and $\mathbf{T} = (T_{ij})_{i,j}$ is the transportation plan matrix with an entry T_{ij} represents the fraction of mass transported from class $i \in \mathcal{Y}_S$ to class $j \in \mathcal{Y}_T$, which is the label mapping strategy.

The total transportation cost and the optimal transportation plan [8] (Figure 1(b)) can be defined as,

$$\mathrm{POT}(\mathbf{p}, \mathbf{q}) = \min_{\mathbf{T} \in \Pi(\mathbf{p}, \mathbf{q})} \langle \mathbf{C}, \mathbf{T} \rangle_F = \min_{\mathbf{T} \in \Pi(\mathbf{p}, \mathbf{q})} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} C_{ij} T_{ij}, \tag{5}$$

where C_{ij} is the cost of transporting unit mass from class $i \in \mathbf{p}$ to class $j \in \mathbf{q}$, $\mathbf{C} \in \mathbb{R}^{n_s \times n_t}$ is the cost matrix, and $\langle \cdot, \cdot \rangle_F$ is the Frobenius dot product.

4.2 Cost Matrix Definition

As defined by Eq.(5), solving for the optimal transport plan $\mathbf{T}$ necessitates a clear definition of the cost matrix $\mathbf{C}$. As mentioned above, the matrix element C_{ij} quantifies the necessary adjustments to transition a label i to a label j in the source and target domain respectively. An effective mapping strategy aims to minimize the overall feature disparity throughout the data distribution, significantly reducing the prompt's workload. This diminished feature variation

facilitates the prompt's task of tailoring the output distribution to match the target labels more efficiently. Therefore, a suitable transportation cost matrix is of paramount importance. Next, we start with an intuitive approach to define the cost of transport within samples. By analyzing the limitations of this method, we further develop the inter-sample transport cost matrix proposed in this work.

Inter-sample Cost Definition. In traditional classification tasks, aligning a feature vector $f_i \in \mathbb{R}_+^{n_s}$ with its label typically uses Kullback-Leibler (KL) divergence as the loss function to measure disparities. In this study, we instead apply Earth Mover's Distance (EMD) to assess the discrepancy between f_i and the one-hot label vector h_t, denoted by $\mathrm{EMD}(f_i, h_t)$, where $h_t \in \mathbb{R}^{n_s}$represents the n_s-dimensional one-hot vector for label t.

To articulate the dimensions of the feature, we adopt the Earth Mover's Distance, derived from optimal transport theory, to quantify the "transport" needed to adjust f_i to h_t. Therefore, for a normalized feature f_i, we have:

$$\mathrm{EMD}(f_i, h_t) = \sum_{j \neq t} f_{i,j} = 1 - f_{i,t}, \tag{6}$$

where $f_{i,j}$ denotes the j-th component of feature vector f_i, implying that the transport required for feature-to-label alignment equates to the sum necessary to transition all feature components, barring the one associated with label t, to that specific label's position. In our ensuing discussion, we will illustrate that methodologies based on frequency (such as FLM, ILM [9]) potentially offer an approximative resolution to this theoretical framework.

Intra-sample Cost Definition. As mentioned, label mapping essentially constitutes a partial optimal transport problem. This characterizes a scenario where the target domain necessitates fewer features than those generated by the source domain. As a result, the alignment of selected features with their corresponding labels does not require the repositioning of all unassociated feature values to the designated label's position. Instead, it specifically entails the reallocation of chosen feature values to align with the label. Accordingly, we can reformulate the transport cost as:

$$\hat{D}_{\mathbf{T}}(f_i, h_t) = \mathrm{EMD}(f_i\mathbf{T}, h_t\mathbf{T}). \tag{7}$$

This revised definition of transport cost selectively incorporates the n_t features chosen from the source domain labels, while disregarding the rest set of n_s features. An issue emerges from this definition because the transformation $\mathbf{T}$ isolates n_t features out of n_s from the source domain, resulting in the aggregated sum of $f_i\mathbf{T}$ diverging from unity. In practical applications, the alignment of the transformed feature vector $f_i\mathbf{T}$ with the target label t involves reallocating the feature values from non-target positions to the target label's position. This procedure entails setting the feature values at non-target positions to zero while ensuring that the feature value at the target position remains positive, reflecting the activation state of the target label. Consequently, this allows us to reformulate the loss function as,

$$\hat{D}_{\mathbf{T}}(f_i, h_t) = \sum_{j \neq t} (f_i\mathbf{T})_j. \tag{8}$$

The preceding Eq.(8) redefines the cost for the alignment of the feature vector f_i with its corresponding label, in the context of label mapping. It quantifies the cost involved in adjusting the feature values at positions not specified by the label to zero, within the subset of n_t selected features from the source domain. With this framework in mind, we proceed to define C_{ij}, the cost associated with mapping a feature at index i from the source to a corresponding label at index j in the target domain. According to our refined definition, the loss incurred by choosing the source feature at index i for mapping to target label j is indirectly determined by the selection of the other $n_t - 1$ indices, which make a direct calculation of C_{ij} complicated. C_{ij} essentially represents the mapping of feature $f_{,i}$ to class j. For $f_{k,i}$, where $k \in \mathcal{C}_j$ and $\mathcal{C}_j = \{k|y_k = j\}$ denotes the set of sample indices labeled j, the transport cost for $f_{k,i}$ according to our definition is null, while the transport cost for $f_{k',i}$, with $k' \notin \mathcal{C}_j$, accumulates to $\sum_{k'} f_{k',i}$. (see Figure 1(c)) Therefore, C_{ij} can be expressed as,

$$C_{ij} = \sum_{k'} f_{k',i}, \quad \forall k' \notin \mathcal{C}_j. \tag{9}$$

For computational simplicity, we adopt a normalized formulation:

$$C_{ij} = \frac{\sum_{k'} f_{k',i}}{\sum_{n=1}^{N} f_{n,i}}, \quad \forall k' \notin \mathcal{C}_j \rightarrow C_{ij} = 1 - \frac{\sum_{k} f_{k,i}}{\sum_{n=1}^{N} f_{n,i}}, \quad \forall k \in \mathcal{C}_j, \tag{10}$$

where N represents the total number of samples, offering a clear and computationally efficient approach to calculating the mapping cost between source and target label indices.

4.3 Solution of the Optimal Transport Plan

With a well-defined cost matrix $\mathbf{C}$, we can establish the label mapping strategy for our method by resolving the partial optimal transport plan. The objective defined in Eq.(5) is to minimize this cost, leading us to the formulation:

$$\mathbf{T}_0 \leftarrow \underset{\mathbf{T} \in \Pi(\mathbf{p}, \mathbf{q})}{\arg\min} \langle \mathbf{T}, \mathbf{C} \rangle_F. \tag{11}$$

Given the task constraint that $\forall i, j, p_i = q_j$, each target label is uniquely mapped to one source label. In such cases, the optimal transport plan $\mathbf{T}_0$ materializes as a permutation matrix, *i.e.*, a binary matrix with solely 0 and 1, where each row and column contains no more than one entry of 1. The partial optimal transport problem is equivalent to an assignment problem. Concurrently, this specification is also perfectly suited for label mapping in visual prompting contexts.

As a result, the label mapping function can be formulated as $\mathcal{F}_{LM}(f_i) = f_i \mathbf{T}$. In practice, this calculation can be bypassed. We maintain an indices vector $\mathbf{I} \in \mathbb{R}^{n_t}$, where $\mathbf{I}_i$ represents the index of the source feature corresponding to the ith label, determined by $\mathbf{I}_i = \text{Find}(\mathbf{T}_{,i}, 1)$. This approach directly maps each target label to its source index efficiently, leveraging the optimal transport plan without the computational overhead.

This assignment problem can be solved via linear programming algorithms, with a computational complexity of $O((n_s+n_t)n_s n_t \log(n_s+n_t))$ [1]. Considering practical scenarios where n_s and n_t are small enough (less than 10^5), this algorithm's computational demand is trivial compared to the extensive calculations required for deep model's forward and backward progress. Consequently, its impact on the overall training speed is marginal, ensuring the method's feasibility even in computationally intensive scenarios.

4.4 Analysis

Within the Optimal Transport (OT) framework, we examine Iterative Label Mapping (ILM) and Frequency-based Label Mapping (FLM), considering ILM as an extension of FLM. This approach hinges on a frequency-based voting mechanism that selects for each category the most frequently occurring maximum value positions across samples, prioritizing these positions in a descending order of frequency.

We introduce a frequency matrix $\mathbf{A} \in \mathbb{R}_{+}^{n_t \times n_s}$, where $\mathbf{A}_{ij}$ reflects the number of times the maximum value index for samples of target category i is j in the source domain. The transport costs are inversely related to these frequencies, simplified as $\max(\mathbf{A}) - \mathbf{A}$ to derive costs, and finalizing the cost matrix as $\mathbf{C}_f = (\max(\mathbf{A}) - \mathbf{A})^T$. Traditional FLM and ILM methods then use a greedy algorithm to compute the transport plan $\mathbf{T}$.

However, these methods diverge from optimal transport principles, primarily by ignoring the detailed information in soft logits and relying on a greedy solution, limiting their ability to reduce prompts' learning effort and enhance efficiency. Our experiments support this critique, revealing the limitations of frequency-based mapping in prompt learning efficiency.

5 Experiment

5.1 Setup

Frameworks. To perform vision prompt tuning, researchers have proposed various approaches, with two frameworks commonly used: visual prompting (**VP**) [3] and visual prompt tuning (**VPT**) [26].

In our paper, we primarily adhere to the VP framework for conducting our experiments. Additionally, we modify the VPT-deep framework to be compatible with label mapping approaches by excluding the learnable task-specific classification layer. This enables us to compare our proposed method with the baselines, further demonstrating its generalizability.

Baselines. Although VP can adapt pretrained models without changing any model parameters, the hard-coded mapping results in a degradation of performance. To achieve better label mapping, **FLM** and **ILM** [9] adopt label mapping based on prediction frequencies on target data points of the model without

Table 1: Comparison of performance on 10 downstream datasets under the VP framework. The utilized source model is ViT-B pretrained on ImageNet-22K. The highest performance achieved among prompt-based methods is highlighted in **bold**.

Dataset	Fine-tuning		Prompt-based			
	FF	LP	VP	FLM-VP	ILM-VP	OTLM
DTD	64.3	63.2	50.0	52.0	48.3	**56.2**
CUB200	87.3	85.3	32.6	41.4	45.2	**56.6**
NAbirds	82.7	75.9	8.4	19.0	19.6	**33.8**
StanfordDogs	89.4	86.2	57.0	58.5	60.3	**67.9**
Flowers102	98.8	97.9	64.1	64.3	62.6	**86.6**
Food101	84.9	84.4	55.6	55.0	56.1	**61.9**
CIFAR100	68.9	63.4	49.4	54.5	54.0	**62.1**
CIFAR10	97.4	96.3	94.6	94.6	94.7	**95.2**
GTSRB	97.1	68.0	87.7	84.3	85.2	**88.6**
SVHN	87.4	36.6	87.0	88.5	**88.8**	88.7
Average	85.82	75.72	58.60	61.20	61.47	**69.77**

employing a vision prompt. We compare our proposed method with these competitive label mapping approaches, as well as full fine-tuning (**FF**) and linear probing (**LP**), serving as baselines.

Datasets and models. Following previous works, we utilize source models ViT-B [18] pretrained on ImageNet-22K and ResNet50 [22] pretrained on ImageNet-1K under the VP framework. For evaluation, we employ 10 target datasets, including DTD [13], CUB200 [47], NABirds [45], StanfordDogs [27], Flowers102 [38], Food101 [7], CIFAR10/100 [29], GTSRB [23], and SVHN [37]. Under the VPT framework, we leverage Swin-B [35] models pretrained on ImageNet-22K [17] and assess their performance on CUB200 [47], NABirds [45], Flowers102 [38], StanfordDogs [27], and StanfordCars [19].

5.2 Performance Under the VP Framework

Our evaluation begins with experiments conducted under the VP framework [3]. We utilize ten distinct downstream tasks to evaluate the transfer performance. We train the source model for 100 epochs with SGD optimizer.

We first adopt a ViT-B model pretrained on ImageNet-22K as the source model and train it with an initial learning rate of 1e4. The results are summarized in Table 1, with results of fine-tuning methods reported in [25]. Based on the results, it is evident that fine-tuning methods outperform prompt-based approaches in the majority of cases. Among prompt-based methods, the original VP method achieves the worst performance as it only employs a hard-coded mapping of labels. In contrast, the other three approaches achieve better performance. Specifically, our method achieves an average performance of 69.77% across all ten datasets, surpassing FLM and ILM by 8.57% and 8.30%, respectively. Following previous work [9], we employ a convolutional model, *i.e.*,

Table 2: Comparison of performance on 10 downstream datasets under the VP framework. The utilized source model is ResNet50 pretrained on ImageNet-1K. The highest performance achieved among prompt-based methods is highlighted in **bold**.

	Fine-tuning		Prompt-based			
Dataset	FF	LP	VP	FLM-VP	ILM-VP	OTLM
DTD	62.1	64.8	14.2	36.8	38.6	**42.7**
CUB200	76.5	68.1	1.1	9.7	11.0	**12.2**
NAbirds	73.7	58.7	5.9	5.9	7.6	**8.1**
StanfordDogs	75.8	88.5	4.3	51.8	51.9	**56.1**
Flowers102	88.1	81.0	6.9	11.3	17.1	**20.5**
Food101	84.0	71.8	7.8	20.6	**22.2**	21.8
CIFAR100	81.2	71.4	12.2	25.8	30.4	**34.8**
CIFAR10	95.8	89.9	55.9	61.3	64.7	**67.9**
GTSRB	95.2	79.4	33.7	39.4	43.2	**48.2**
SVHN	96.5	45.3	55.2	56.9	61.0	**62.7**
Average	82.89	71.89	19.71	31.94	34.77	**37.48**

ResNet50 pretrained on ImageNet-1K, to further evaluate the proposed method. For this model, we initialize the learning rate to 1e2 and present the results in Table 2. With this source model, similar relevant accuracy is observed across each dataset for each method, where fine-tuning methods consistently outperform prompt-based approaches, and our proposed OTLM method outperforms other prompt-based approaches. Nevertheless, the absolute accuracy values are significantly lower than those achieved by the pretrained ViT-B model, primarily due to the limited model capacity of the ResNet50 model.

5.3 Performance Under the VPT Framework

We further assess our method under the VPT-deep framework [26], employing five different datasets. We adapt VPT by removing its learnable head to accommodate label mapping methods. The number of added prompt tokens for each dataset adheres to the original implementation of VPT. Source model is trained for 100 epochs on each dataset, utilizing a learning rate of 0.1 and a weight decay of 1e−4 with the SGD optimizer. Considering that the original VPT method necessitates training an extra classification layer for each downstream dataset, we exclude it from the comparison of label mapping methods.

We utilize Swin-B models pretrained on ImageNet-22K as source model and train them on downstream datasets. The corresponding results are presented in Table 3. Once again, fully fine-tuning (FF) achieves the best performance. Original VPT enhances performance over LP due to introducing not only trainable tokens at the input layer but also an extra head for each dataset at the output layer. On the contrary, applying label mapping to VPT enables it to map source labels to target labels without an extra classification head, which means less parameters to optimize. Moreover, our method significantly outperforms the other two LM methods and achieves the best results on all five datasets. Notably, av-

Table 3: Comparison of performance on five datasets under the VPT framework using model Swin-B pretrained on ImageNet-22K. The highest performance achieved among label mapping methods without additional heads is highlighted in **bold**.

Dataset	Fine-tuning			Label mapping		
	FF	LP	VPT	FLM	ILM	OTLM
Extra head	✓	✓	✓	✗	✗	✗
CUB-200-2011	88.8	87.7	85.4	80.7	81.8	**84.7**
NABirds	86.3	82.6	81.1	71.3	74.4	**77.2**
Flowers102	98.8	98.0	98.7	95.8	96.1	**98.5**
StanfordDogs	84.3	82.7	84.6	79.1	81.1	**83.1**
StanfordCars	90.4	68.2	80.3	73.4	74.8	**78.0**
Average	89.72	83.8	86.0	80.1	81.6	**84.30**

erage metric of our approach even surpasses that of LP and is comparable to the results of methods with and additional head, except for FF.

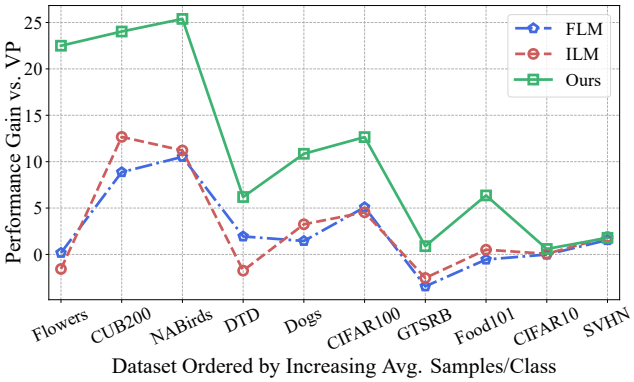
5.4 Data Efficiency in Downstream Task Transfer

In many cases, downstream tasks lack sufficient labeled data, making data efficiency — the ability to achieve proficient performance using limited data — crucial. To assess the data efficiency of various label mapping methods, we analyze the effectiveness of different approaches for integrating label mapping into the VP framework using the prompt-tuned ViT-B model. Specifically, we introduce a metric to measure data scale, defined as the average number of samples in each class. A higher value indicates a larger dataset scale, while a lower value suggests a smaller one. We explore the relationship between data scale and the efficacy of label mapping.

In Figure 2, we present the performance improvement of label mapping methods relative to the original VP method, arranged in ascending order of data scale. The figure illustrates that all three approaches exhibit substantial performance gains with smaller data scales. However, the performance gap compared to VP diminishes as the scale increases. Notably, our OTLM method consistently outperforms both FLM and ILM, particularly on small-scale data. This finding underscores the superior data efficiency of our method, which is significant in practical applications.

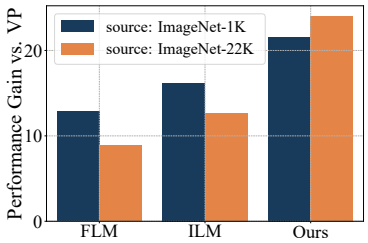
5.5 Scalability

A larger pretrained dataset typically yields a source model with enhanced representation extraction capabilities. However, the commonly concomitant increase in output dimensionality may also complicate the transportation process. To further investigate the performance of label mapping methods across source datasets of varying scales, we utilize Vit-B source models pretrained on both ImageNet-1K and ImageNet-22K, transferring them to the CUB200 downstream dataset.



Dataset	S/C
Flowers	10
CUB200	27
NABirds	39
DTD	40
Dogs	90
CIFAR100	400
GTSRB	496
Food101	600
CIFAR10	4000
SVHN	5860

Fig. 2: Data efficiency of variant label mapping approaches. Compared to FLM and ILM, our method demonstrates significantly higher accuracy gains over VP, particularly evident when the downstream dataset has fewer averaged samples per class. This highlights exceptional data efficiency of our method.



Method	Average cost IN-1K	Average cost IN-22K
FLM	17.96	43.85
ILM	16.28	45.92
Ours	13.53	34.18

Fig. 3: Comparison of different source datasets by employing Vit-B source models pretrained on either ImageNet-1K or ImageNet-22K. The performance improvement over VP and average cost under the CUB200 downstream dataset is reported.

Alongside evaluating performance enhancements over the VP method, we employ the average cost as a metric for comparison. The average cost is computed by averaging the transportation cost of predictions from the source model into the target label space over the whole downstream dataset.

The results are depicted in Figure 3. In the left figure, it is evident that as the dataset scale increases, FLM and ILM show diminished improvement in downstream performance, while our method demonstrates enhanced performance improvement. Upon examining the values of the average cost, we observe that for each method, the cost value increases with the scale of the source dataset. This phenomenon arises because the source model pretrained on ImageNet-22K possesses a larger output dimension, making label mapping to a downstream task more challenging due to the increased number of possible mapping solutions. However, our partial optimal transportation-based method only considers a subset of the output dimension, which matches the output dimension of the downstream dataset.

This approach mitigates the negative impact of the increasing source dataset, as evidenced by the lower cost and higher performance improvement of OTLM compared to FLM and ILM. Furthermore, across each dataset, our OTLM method consistently outperforms FLM and ILM in terms of performance gain over VP and average cost, underscoring its superior performance and scalability.

5.6 Ablation Study

We conducted an ablation study to examine the effectiveness of each module within our proposed OTLM method. The primary enhancements of our method lie in the cost matrix obtained under the optimal transportation framework and the utilization of linear programming for solving the label mapping result. We employ the ImageNet-22K pretrained ViT-B as the source model and adopt the Flowers102 dataset to investigate the impact of these two modules.

Table 4: Ablation study of various solving algorithms and cost matrix designs.

Algorithm	**Cost matrix**	**Accuracy**
Greedy	Frequency	62.56
Linear programming	Frequency	64.29
Greedy	OT	79.92
Linear programming	OT	**86.63**

The ablation results are presented in Table 4. As shown, the combination of the greedy solver and the frequency-based cost matrix achieves the worst result, indicating the limitations of the existing ILM method. Building upon this, integrating either the optimal transportation-based cost matrix or linear programming helps improve performance. Adopting both of these modules, as OTLM does, yields the best performance, achieving a top-1 accuracy of 86.63%. These results confirm the effectiveness of each module within our proposed method.

6 Conclusion

This paper focuses on how appropriate label mapping can help alleviate the burden on input prompts. Specifically, we define a mapping cost matrix and propose OTLM, a novel label mapping strategy based on optimal transport that can minimize the necessary adjustments of the model when training prompts. Additionally, we theoretically analyze why existing frequency-based methods struggle to achieve this goal. Extensive experimental results prove that the proposed method improves accuracy significantly over all other LM methods on VP framework on 10 datasets. Furthermore, OTLM can also be easily applied to the VPT framework and achieves results comparable to VPT with an extra head.

Acknowledgements

This work was supported in part by the Australian Research Council under Projects DP240101848 and FT230100549.

References

1. Ahuja, R.K., Magnanti, T.L., Orlin, J.B.: Network flows: theory, algorithms and applications. Prentice Hall (1995) 9
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. PMLR (2017) 4
3. Bahng, H., Jahanian, A., Sankaranarayanan, S., Isola, P.: Exploring visual prompts for adapting large-scale models. arXiv preprint arXiv:2203.17274 (2022) 2, 4, 9, 10
4. Bai, Y., Schmitzer, B., Thorpe, M., Kolouri, S.: Sliced optimal partial transport. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13681–13690 (2023) 4
5. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021) 1
6. Bonneel, N., Rabin, J., Peyré, G., Pfister, H.: Sliced and radon wasserstein barycenters of measures. Journal of Mathematical Imaging and Vision **51**, 22–45 (2015) 4
7. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13. pp. 446–461. Springer (2014) 10
8. Chapel, L., Alaya, M.Z., Gasso, G.: Partial optimal tranport with applications on positive-unlabeled learning. Advances in Neural Information Processing Systems **33**, 2903–2913 (2020) 6
9. Chen, A., Yao, Y., Chen, P.Y., Zhang, Y., Liu, S.: Understanding and improving visual prompting: A label-mapping perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19133–19143 (2023) 2, 4, 7, 9, 10
10. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12299–12310 (2021) 1
11. Chen, L., Fan, Y., Ye, Y.: Adversarial reprogramming of pretrained neural networks for fraud detection. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. pp. 2935–2939 (2021) 2, 4
12. Chizat, L., Peyré, G., Schmitzer, B., Vialard, F.X.: Unbalanced optimal transport: Dynamic and kantorovich formulations. Journal of Functional Analysis **274**(11), 3090–3123 (2018) 6
13. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3606–3613 (2014) 10
14. Courty, N., Flamary, R., Habrard, A., Rakotomamonjy, A.: Joint distribution optimal transportation for domain adaptation. Advances in neural information processing systems **30** (2017) 4
15. Courty, N., Flamary, R., Tuia, D., Rakotomamonjy, A.: Optimal transport for domain adaptation. IEEE transactions on pattern analysis and machine intelligence **39**(9), 1853–1865 (2016) 4

16. Damodaran, B.B., Kellenberger, B., Flamary, R., Tuia, D., Courty, N.: Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In: Proceedings of the European conference on computer vision (ECCV). pp. 447–463 (2018) 4
17. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition (2009) 10
18. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021) 1, 10
19. Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Fei-Fei, L.: Fine-grained car detection for visual census estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017) 10
20. Gu, X., Yang, L., Sun, J., Xu, Z.: Optimal transport-guided conditional score-based diffusion model. Advances in Neural Information Processing Systems **36**, 36540–36552 (2023) 4
21. Gu, X., Yang, Y., Zeng, W., Sun, J., Xu, Z.: Keypoint-guided optimal transport with applications in heterogeneous domain adaptation. Advances in Neural Information Processing Systems **35**, 14972–14985 (2022) 4
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2016) 10
23. Houben, S., Stallkamp, J., Salmen, J., Schlipsing, M., Igel, C.: Detection of traffic signs in real-world images: The german traffic sign detection benchmark. In: The 2013 international joint conference on neural networks (IJCNN). pp. 1–8. Ieee (2013) 10
24. Hu, J., Chen, C., Cao, L., Zhang, S., Shu, A., Jiang, G., Ji, R.: Pseudo-label alignment for semi-supervised instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16337–16347 (2023) 4
25. Huang, Q., Dong, X., Chen, D., Zhang, W., Wang, F., Hua, G., Yu, N.: Diversity-aware meta visual prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10878–10887 (2023) 2, 4, 10
26. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022) 4, 9, 11
27. Khosla, A., Jayadevaprakash, N., Yao, B., Li, F.F.: Novel dataset for fine-grained image categorization: Stanford dogs. In: Proc. CVPR workshop on fine-grained visual categorization (FGVC). vol. 2. Citeseer (2011) 10
28. Korotin, A., Selikhanovych, D., Burnaev, E.: Neural optimal transport. arXiv preprint arXiv:2201.12220 (2022) 4
29. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009) 10
30. Kurmi, V.K., Namboodiri, V.P.: Looking back at labels: A class based domain adaptation technique. In: 2019 international joint conference on neural networks (IJCNN). pp. 1–8. IEEE (2019) 4
31. Le, T., Nguyen, T., Ho, N., Bui, H., Phung, D.: Lamda: Label matching deep domain adaptation. In: International Conference on Machine Learning. pp. 6043–6054. PMLR (2021) 4

32. Li, Y., Tsai, Y.L., Yu, C.M., Chen, P.Y., Ren, X.: Exploring the benefits of visual prompting in differential privacy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5158–5167 (2023) 2, 4
33. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022) 4
34. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys **55**(9), 1–35 (2023) 3
35. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: International Conference on Computer Vision. pp. 9992–10002 (2021) 10
36. Ma, X., Wang, Y., Liu, H., Guo, T., Wang, Y.: When visual prompt tuning meets source-free domain adaptive semantic segmentation. Advances in Neural Information Processing Systems **36** (2024) 3
37. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011) 10
38. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian conference on computer vision, graphics & image processing. pp. 722–729. IEEE (2008) 10
39. Pfeiffer, J., Rücklé, A., Poth, C., Kamath, A., Vulić, I., Ruder, S., Cho, K., Gurevych, I.: Adapterhub: A framework for adapting transformers. arXiv preprint arXiv:2007.07779 (2020) 2
40. Salimans, T., Zhang, H., Radford, A., Metaxas, D.: Improving gans using optimal transport. arXiv preprint arXiv:1803.05573 (2018) 4
41. Sang, J., Wang, Y., Yuan, L., Li, H., Jiang, X.: Multi-label transfer learning via latent graph alignment. World Wide Web **25**(2), 879–898 (2022) 4
42. Schick, T., Schütze, H.: It's not just size that matters: Small language models are also few-shot learners. arXiv preprint arXiv:2009.07118 (2020) 4
43. Sohn, K., Chang, H., Lezama, J., Polania, L., Zhang, H., Hao, Y., Essa, I., Jiang, L.: Visual prompt tuning for generative transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19840–19851 (2023) 4
44. Tsai, Y.Y., Chen, P.Y., Ho, T.Y.: Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources. In: International Conference on Machine Learning. pp. 9614–9624. PMLR (2020) 2, 4
45. Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., Belongie, S.: Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 595–604 (2015) 10
46. Villani, C., et al.: Optimal transport: old and new, vol. 338. Springer (2009) 4
47. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011) 10
48. Wu, J., He, J.: Continuous transfer learning with label-informed distribution alignment. arXiv preprint arXiv:2006.03230 (2020) 4
49. Zhao, Y., Dai, G., Borghini, G., Zhang, J., Li, X., Zhang, Z., Aricò, P., Di Flumeri, G., Babiloni, F., Zeng, H.: Label-based alignment multi-source domain adaptation for cross-subject eeg fatigue mental state evaluation. Frontiers in Human Neuroscience **15**, 706270 (2021) 4