Learning Optimal Prompt Ensemble for Multi-source Visual Prompt Transfer

Enming Zhang[1]   Liwen Cao[2]   Yanru Wu[1]   Zijie Zhao[1]   Yang Li[1*]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University [†]

[2]Southeast University

Abstract

Prompt tuning has emerged as a lightweight adaptation strategy for adapting foundation models to downstream tasks, particularly in resource-constrained systems. As pretrained prompts have become valuable intellectual assets, combining multiple source prompts offers a promising approach to enhance generalization to new tasks by leveraging complementary knowledge from diverse sources. However, naive aggregation of these prompts often leads to representation collapse due to mutual interference, undermining their collective potential. To address these challenges, we propose HGPrompt, an adaptive framework for multi-source prompt transfer that learns optimal ensemble weights by jointly optimizing dual objectives: transferability and stability. Specifically, we first introduce an information-theoretic metric to evaluate the transferability of prompt-induced features on the target task, capturing the intrinsic alignment between the feature representations. Additionally, we propose a novel Gradient Alignment Regularization to mitigate gradient conflicts among prompts, enabling stable and coherent knowledge transfer from multiple sources while suppressing interference. Extensive experiments on the large-scale VTAB benchmark demonstrate that HGPrompt achieves state-of-the-art performance, validating its effectiveness in multi-source prompt transfer.

1. Introduction

With the development of expanding datasets, novel architectures, and improved training algorithms [7], a significant number of vision foundation models have been developed [9, 22, 24]. Transformer-based pre-trained vision models (PVMs) demonstrate exceptional efficacy across diverse tasks, including image classification and semantic segmentation. While these models exhibit impressive capability, adapting them to downstream applications still presents notable challenges. Full model fine-tuning becomes impractical given the substantial parameter volumes and challenges
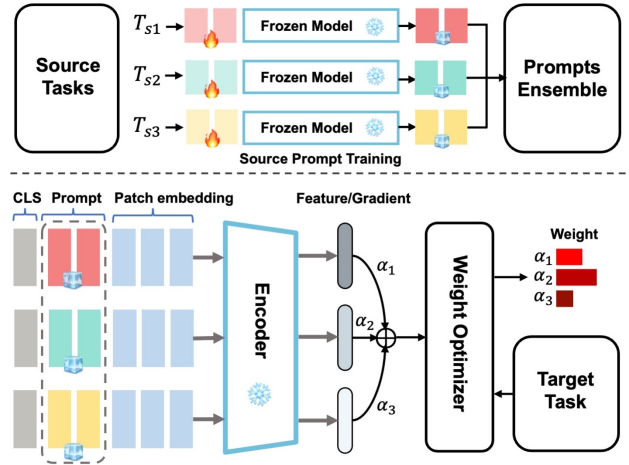

Figure 1. Multi-source prompt transfer framework. Task-specific prompts are tuned via a frozen backbone and statically aggregated for target initialization. Our approach dynamically optimizes source weights through single forward-backward propagation, learning prompt aggregation via an optimization module.

in low-data scenarios. This paradigm shift has made prompt tuning [14, 19] a key adaptation strategy. By freezing PVMs and adding learnable prompt tokens, it achieves competitive performance with only 0.4% parameter updates, significantly fewer than full fine-tuning.

The increasing sophistication of prompt learning has established well-generalized prompts as valuable intellectual assets [27]. This evolution has fostered a practical ecosystem where users can access provider task-specific prompts while maintaining data privacy and model integrity. With the availability of multiple prompts from the prompt pool, these prompts can be utilized in an ensemble way by concurrently assembling them and transferring them to a single pre-trained model (as illustrated in Fig. 1) [26, 35]. However, simply concatenating or averaging source prompts often proves suboptimal, as the knowledge encoded in different prompts may contribute unevenly to the target task and can even lead to negative transfer [28].

Previous work [2, 29, 33, 45] predominantly evaluates source prompt transferability as independent characteris-

---

tics, failing to treat prompt ensembles as a cohesive system. These approaches overlook the interactions emerging with the combination of multiple prompts, whose performance can significantly enhance or degrade performance depending on their compatibility. Moreover, existing methods primarily rely on parameter similarity measurement, ignoring the essential role of prompts in shaping latent feature representations [46] and guiding feature extraction processes. Simply aggregating prompts without consideration of preserving critical task-specific features may result in representation collapse, where essential discriminative information is lost. Consequently, two fundamental challenges remain unresolved: (1) How to systematically evaluate prompt transferability beyond simplistic parameter similarity metrics? (2) How to effectively achieve prompt ensembles that mitigate conflicts and negative interactions within prompts?

To address these challenges, we introduce an information-theoretic metric, H-score [38], as a novel assessment for assessing the transferability of weighted combinations of prompt-guided feature spaces. Grounded in information theory, the transferability metric provides a principled mechanism to quantify cross-prompt synergy and maximize their discriminability. Unlike conventional similarity-based assessments, which focus solely on parameter-space relationships, the metric evaluates the intrinsic informativeness of prompt-conditioned features. This approach offers a more explicit and interpretable transferability measure by directly assessing how well the combined prompts capture task-relevant information while minimizing redundancy and interference.

Furthermore, we propose a gradient alignment strategy to address cross-interference between prompts. This method resolves the problem of unstable optimization directions stemming from mismatched gradients [17, 40], which can disrupt training and hinder convergence. By aligning the gradient directions of all prompts, we ensure they collectively steer the model toward a unified optimization path. Specifically, we introduce the Gradient Alignment Loss, a metric that quantifies the variance in gradient directions across source prompts. Minimizing this loss promotes coherence among the source prompts, thereby enhancing their collaboration and transferability. Building upon these dual metrics, our framework unifies these objectives through a novel optimization formulation, advancing beyond conventional heuristic approaches. This approach dynamically balances feature discriminability with optimization stability, enabling the model to leverage complementary information across prompts and suppress interference simultaneously.

Our method achieves state-of-the-art performance through extensive evaluations on the large-scale VTAB benchmark [42], consistently outperforming competitive strategies such as PANDA [45], SPoT, and ATTEMPT. Our approach establishes new benchmarks across a wide range of tasks, with robust results in challenging domains like remote sensing and 3D understanding, demonstrating the scalability and effectiveness of our optimization strategy.

2. Related Work

2.1. Parameter-efficient Transfer Learning

Parameter-efficient transfer learning has emerged as a critical research direction for adapting large pre-trained models. In NLP, various approaches have been proposed, including adapter modules [11], bias term tuning [5], and low-rank adaptation [13]. These methods achieve competitive performance while updating only 1-5% of model parameters. In computer vision, early parameter-efficient approaches focused on ConvNets through techniques like residual adapters [25]. With the recent paradigm shift towards vision Transformers [9], new challenges and opportunities for model adaptation have emerged. Some NLP methods, like adapters, have been directly applied to vision tasks [6]. Meanwhile, more vision-specific solutions have been explored. For example, Visual Prompt Tuning (VPT) [16] inserts learnable tokens in the input space, and Visual Prompting (VP) [3] applies learnable pixel-level perturbations. These methods show that minimal modifications to input embeddings can effectively adapt vision Transformers with high parameter efficiency.

2.2. Transferability Estimation

Understanding task transferability [41] is essential for prompt-based methods because prompts act as task-specific instructions that steer the behavior of a frozen model [10]. Thus, insights from task transferability directly inform the design and evaluation of prompt transferability. Prior work on task transferability has proposed various metrics based on task models and data distributions to estimate transferability [8, 32] quickly. Several information-theoretic metrics have been proposed to estimate task transferability, which can be adapted to evaluate prompt transferability. For instance, the H-score [4, 15, 37] uses a maximum correlation framework to measure the discriminability of features, providing a theoretical foundation for evaluating prompt effectiveness. Similarly, LEEP [1, 23] and LogME [39] assess transferability by predicting target task performance based on source model outputs or features. Recent advances in optimal transport have also been applied to transferability estimation. For example, OTCE [30, 31] combines optimal transport with conditional entropy to measure domain and task differences between source and target. The study of task transferability provides a strong theoretical and practical foundation for understanding prompt transferability.
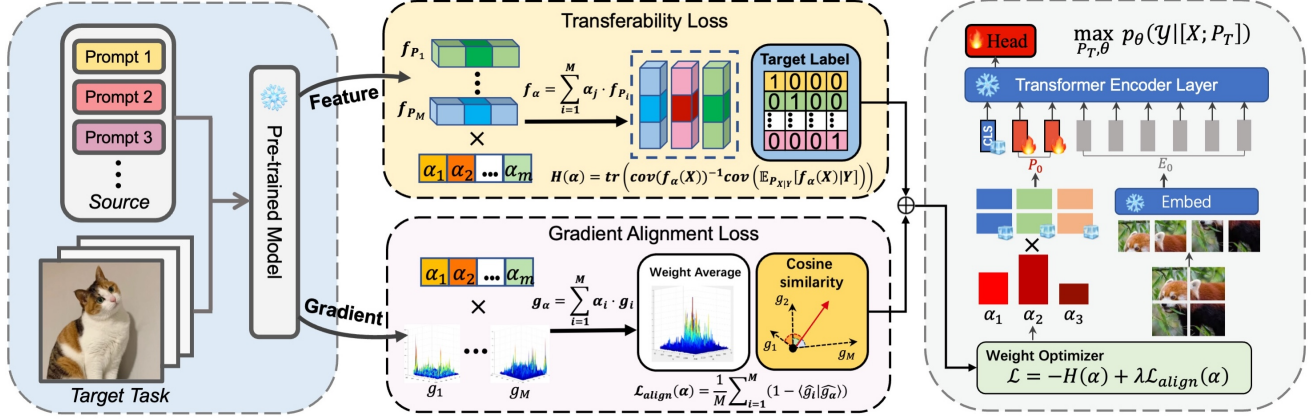
Figure 2. Overview of the framework. Given an input image X, the system generates M distinct feature representations $\{f_i\}_{i=1}^{M}$ and corresponding gradients $\{g_i\}_{i=1}^{M}$ through multiple source prompts. These features and gradients are fused using learnable weights $\boldsymbol{\alpha}$ to produce the final combined feature $f_{\boldsymbol{\alpha}}$ and gradient $g_{\boldsymbol{\alpha}}$. The Transferability Loss evaluates the fused feature distribution against the target class label, and the Gradient Alignment Loss measures consistency between individual gradients and their ensemble. Gradient directions from individual prompts are shown as dotted arrows, while the red arrow represents their weighted combination. The Weight Optimizer jointly optimizes these dual objectives to determine the optimal source weights $\boldsymbol{\alpha}$, which subsequently initialize the target prompt.

2.3. Multi-source Prompt Tuning

Prompt-tuning on smaller pre-trained models often underperforms and is highly sensitive to prompt initialization, as evidenced by prior studies [14, 19]. To address these limitations, Prompt Transfer (PoT) methods have been proposed [29, 33], which leverage soft prompts learned on source tasks to initialize prompts for target tasks, thereby improving tuning efficiency and performance. SPOT [33] explored the use of metrics to predict the best source tasks for prompt transfer, and in parallel, [29] emphasized how prompt-induced neuron activations play a crucial role in transferability. In addition to single-task transfer, PoT methods have been extended to multi-task settings. For example, ATTEMPT [2] proposed mechanisms to aggregate knowledge from multiple source tasks, using attention mechanisms strategies to initialize target prompts. PANDA [45] explicitly addresses the issue of prior knowledge forgetting by distilling task-specific knowledge into the target prompt.

3. Preliminary

3.1. Visual Prompt Tuning

Visual Prompt Tuning (VPT) is a parameter-efficient transfer learning paradigm that adapts pre-trained vision transformers to downstream tasks by learning task-specific prompt embeddings while keeping the original model parameters frozen. This approach introduces a small set of learnable parameters in the form of prompt tokens, which are prepended to the input sequence, enabling efficient adaptation to new tasks without modifying the underlying model architecture. The key advantage of VPT lies in its ability to leverage the rich representations learned by large-

scale pre-trained models while requiring significantly fewer trainable parameters compared to full fine-tuning.

Formally, given a pre-trained Transformer with embedding dimension d, we introduce m learnable prompt tokens $P = [p_1, \ldots, p_m] \in \mathbb{R}^{m \times d}$. For an input image X with patch embeddings $E(X) \in \mathbb{R}^{n \times d}$, the combined input sequence becomes $[P; E(X)] \in \mathbb{R}^{(m+n) \times d}$, where m is the prompt length and n is the number of image patches. The model parameters θ remain fixed during training, with gradients only propagating through the prompt embeddings P. The prediction probability for class y is given by:

$$\Pr_{\theta}(y|x; P) = \frac{\exp(f_y([P; E(x)]; \theta))}{\sum_{i=1}^{C} \exp(f_i([P; E(x)]; \theta))}, \quad (1)$$

where C denotes the number of classes, and $f_i(\cdot)$ represents the pre-trained model's logit output for class i. This formulation allows the model to adapt to new tasks by learning task-specific context through the prompt tokens.

3.2. Multi-Source Prompt Transfer

In many real-world scenarios, we often have access to multiple source prompts that can be utilized for the target task. Multi-source prompt transfer aims to harness these related prompts to enhance performance on the target task. Given κ source tasks $\mathcal{S} = \{S_i\}_{i=1}^{\kappa}$ along with their corresponding optimized prompts $\{P_i\}_{i=1}^{\kappa}$, our goal is to construct a target prompt P_T for a new task T by optimally combining the source prompts based on their relevance to the target task.

Let $M \leq \kappa$ denote the number of selected source prompts. We formulate this as learning combination weights $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$ that maximize the target task performance:

$$\max_{\boldsymbol{\alpha}} \quad \mathbb{E}_{(x,y)\sim\mathcal{D}_T}\left[\log P_\theta(y|x; P_T)\right]$$

$$\text{s.t.} \quad P_T = \sum_{i=1}^{M}\alpha_i P_i \tag{2}$$

$$\sum_{i=1}^{M}\alpha_i = 1, \quad \alpha_i \geq 0 \quad \forall i$$

This formulation ensures that the target prompt is constructed as a convex combination of the source prompts, with the weights α_i reflecting the relative importance of each source task to the target task. The constraints $\sum_{i=1}^{M}\alpha_i = 1$ and $\alpha_i \geq 0$ ensure that the combination is interpretable and each source prompt contributes positively to the target prompt. The learned weights can also provide insights into the relationships between the source and target tasks, potentially revealing task similarities and transferability.

4. Methodology

To address cross-interference and unstable optimization in multi-source prompt transfer, we propose a dual-objective framework that maximizes feature discriminability and ensures gradient coherence. The framework integrates two key modules: the transferability loss, an information-theoretic metric for evaluating prompt-conditioned feature informativeness, and the Gradient Alignment Loss, which aligns gradient directions to stabilize optimization. By jointly optimizing these objectives, our approach mitigates cross-interference, refines the feature space, and leverages the collective potential of prompts for robust transfer.

4.1. Measuring Prompt Ensemble Transferability

To overcome the limitations of previous prompt ensemble strategies in efficiency or overlook of prompt interactions, we adopt an information theoretical metric to coherently measure the ensemble transferability of prompts. Specifically, we introduce a theoretically grounded metric of cross-prompt transferability based on H-score[38]. Unlike conventional assessments designed by assuming the relevance between transferability and parameter similarity, our proposed metric focuses on the intrinsic informativeness of prompt-conditioned features, explicitly revealing the effectiveness of prompt ensembles on the target task.

Given a frozen visual encoder f_θ and M source prompts $\{P_i\}_{i=1}^{M}$ pre-trained on different tasks, we construct features for a target task through a linear combination of source-specific features, where for an input image $X \in \mathcal{X}$, the feature extraction process given the i-th source prompt is defined as:

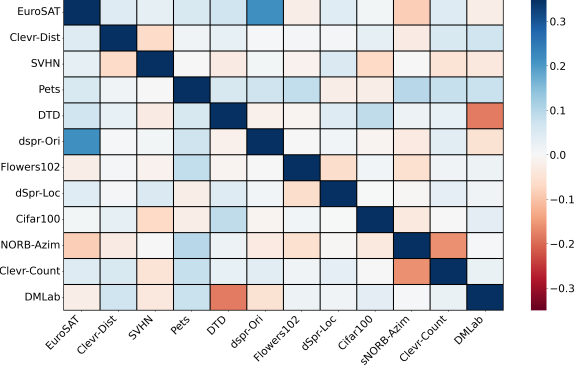$$f_{P_i}(X) = f_\theta\left([x_0; P_i; E(X)]\right) \in \mathbb{R}^h. \tag{3}$$

Figure 3. Heatmap visualizing the cosine similarity of gradients across different prompts.

$E(X) \in \mathbb{R}^{n\times d}$ denotes the image patch embeddings, $x_0 \in \mathbb{R}^d$ the [CLS] token, and h the feature dimension. The fused feature representation is obtained by a weighted fusion of these source-specific features:

$$f_{\boldsymbol{\alpha}}(X) = \sum_{i=1}^{M}\alpha_i f_{P_i}(X), \tag{4}$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_M)$ are learnable combination weights satisfying $\sum_{i=1}^{M}\alpha_i = 1$ and $\alpha_i \geq 0$ for all i.

To evaluate the quality of the derived feature, we employ the H-score based metric, which measures the discriminative power of the feature representation. Given the fused features, our metric is defined as:

$$H(\boldsymbol{\alpha}) = \text{tr}\left(\text{cov}(f_{\boldsymbol{\alpha}}(X))^{-1}\text{cov}\left(\mathbb{E}_{X|Y}[f_{\boldsymbol{\alpha}}(X)|Y]\right)\right). \tag{5}$$

In this formulation, $\text{cov}(f_{\boldsymbol{\alpha}}(X))$ captures the global variability of the features across all inputs. At the same time, the conditional expectation $\mathbb{E}_{X|Y}[f_{\boldsymbol{\alpha}}(X)|Y]$ represents the label-conditioned average of the features. The covariance of this conditional expectation measures how informative the features are for distinguishing different labels. A high score value indicates that the prompt successfully conditions the model to produce task-effective features, thereby implying superior transferability. Leveraging this metric, we provide a principled and interpretable evaluation for prompt transferability, enabling a capture of interior prompt interactions that naive similarity measures usually overlook.

4.2. Gradient Alignment Regularization

Each prompt encodes task-specific knowledge. However, directly aggregating these prompts often leads to suboptimal performance due to two critical challenges: (1) cross-interference between prompts, where independent evaluation fails to account for their synergistic or conflicting interactions, and (2) unstable optimization directions, where mismatched gradients from different prompts induce conflicting updates, as illustrated in Fig.3. To address these

issues, we propose aligning the gradient directions of all prompts, ensuring they collectively guide the model toward a unified optimization trajectory.

To establish the theoretical foundation for gradient alignment in prompt ensembles, we analyze how combining multiple prompts geometrically shapes the model's optimization trajectory, inspired by [18, 20, 21]. Consider augmenting a base prompt P_i with an additional source prompt P_j. The first-order Taylor expansion reveals how the model's output $f(x; P_i)$ evolves under this ensemble:

$$f(x; P_i + \alpha_j P_j) - f(x; P_i) \approx \alpha_j \nabla_{P_i} f(x; P_i)^\top P_j, \quad (6)$$

where $\nabla_{P_i} f$ captures the directional sensitivity of the model to prompt interactions. Extending this to an ensemble of M prompts with convex weights $\boldsymbol{\alpha}$, the combined prompt $P_T = \sum_{i=1}^{M} \alpha_i P_i$ induces an output shift governed by the weighted gradient contributions:

$$\Delta f \approx \sum_{i=1}^{M} \alpha_i \nabla_{P_i} f(x; P_i)^\top P_i. \quad (7)$$

This formulation exposes a critical dependency: misaligned gradients across prompts ($\nabla_{P_i} f$) create conflicting contributions to the ensemble's feature space, destabilizing optimization.

To ensure coherent parameter updates across multiple source prompts, we propose a gradient alignment mechanism that enforces directional consistency in the optimization process. Given a batch of training samples (X, Y), let $\mathcal{L}(f_{P_i}(X), Y)$ denote the task loss for the i-th source prompt. For each source prompt P_i, we compute its normalized gradient direction:

$$\hat{g}_i = \frac{\nabla_{P_i} \mathcal{L}(f_{P_i}(X), Y)}{\|\nabla_{P_i} \mathcal{L}(f_{P_i}(X), Y)\|_2}. \quad (8)$$

The ensemble gradient is defined as:

$$g_{\boldsymbol{\alpha}} = \sum_{i=1}^{M} \alpha_i \hat{g}_i \in \mathbb{R}^{p \times d}, \quad (9)$$

reflecting the collective optimization direction. The gradient alignment regularization is formulated as follows:

$$\mathcal{L}_{\text{align}}(\boldsymbol{\alpha}) = \frac{1}{M} \sum_{i=1}^{M} (1 - \langle \hat{g}_i, \widehat{g}_{\boldsymbol{\alpha}} \rangle), \quad (10)$$

where $\widehat{g}_{\boldsymbol{\alpha}} = g_{\boldsymbol{\alpha}} / \|g_{\boldsymbol{\alpha}}\|_2$ is the normalized consensus gradient, and $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product. This loss quantifies directional incoherence by measuring deviations between individual gradients and the ensemble gradient. Minimizing this loss enforces geometric consistency across prompts, ensuring their gradients cooperatively steer the model toward shared minima.

Algorithm 1 Ensemble weight optimization with gradient regularization

---

Input: Target data $\mathcal{D}_T = \{(x_i, y_i)\}_{i=1}^{N}$, source prompts $\{P_j\}_{j=1}^{M}$, learning rate η, trade-off parameter λ
Output: Optimal weights $\boldsymbol{\alpha}^*$
1: Initialize $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \dots, \alpha_M\}$ with $\sum_{j=1}^{M} \alpha_j = 1$
2: **for** epoch = 1 to K **do**
3:     $\mu_y(\boldsymbol{\alpha}) = \mathbb{E}_{X|Y}[f_{\boldsymbol{\alpha}}(X)|Y = y]$
4:     $H(\boldsymbol{\alpha}) = \text{tr}(\text{cov}(f_{\boldsymbol{\alpha}})^{-1}\text{cov}(\{\mu_y\}))$
5:     Compute normalized gradient directions: $\{\hat{g}_j\}_{j=1}^{M}$ via Eq.(10)
6:     Evaluate gradient alignment loss: $\mathcal{L}_{\text{align}}(\boldsymbol{\alpha})$ via Eq.(11)
7:     Compute total loss: $\mathcal{L}(\boldsymbol{\alpha}) = -H(\boldsymbol{\alpha}) + \lambda\mathcal{L}_{\text{align}}(\boldsymbol{\alpha})$
8:     Update weights: $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \eta\nabla_{\boldsymbol{\alpha}}\mathcal{L}$
9: **end for**

From a statistical perspective, the ensemble's Fisher Information Matrix (FIM) decomposes into:

$$\mathcal{I}_{\text{ens}} = \sum_{i=1}^{M} \alpha_i \mathcal{I}(P_i) + \sum_{i \neq j} \alpha_i \alpha_j \text{Cov}(\hat{g}_i, \hat{g}_j), \quad (11)$$

where $\mathcal{I}(P_i) = \mathbb{E}[\hat{g}_i \hat{g}_i^\top]$ encodes individual prompt sensitivity. The cross-term $\text{Cov}(\hat{g}_i, \hat{g}_j)$ captures gradient interactions: misaligned directions increase covariance terms, ill-conditioning the FIM and destabilizing training [36]. By minimizing the alignment loss $\mathcal{L}_{\text{align}}$, we suppress these covariance terms, thereby conditioning $\mathcal{I}_{\text{ens}}$ to ensure stable convergence. This geometric alignment ensures that all prompts collectively refine the feature space in a coherent manner, avoiding destructive interference while amplifying complementary knowledge. In our formulation, the alignment loss quantifies the variance in gradient directions across source prompts, and minimizing it enhances the coherence of the aggregated gradient. This synergy-aware training approach not only stabilizes optimization but also unlocks the full potential of prompt-based learning by fostering collaborative and consistent knowledge transfer.

4.3. Optimization Procedure

The final objective function integrates feature discriminability and gradient alignment through a weighted combination:

$$\mathcal{L}(\boldsymbol{\alpha}) = -H(\boldsymbol{\alpha}) + \lambda\mathcal{L}_{\text{align}}(\boldsymbol{\alpha}), \quad (12)$$

Where λ is a hyperparameter that balances the trade-off between maximizing the H-score and minimizing the Gradient Alignment Loss, this formulation ensures that the optimization process simultaneously enhances the discriminative power of the aggregated features while maintaining consistent gradient directions across different source prompts.

The optimization problem can be efficiently solved using gradient-based methods, leveraging the convexity of α to ensure reliable convergence, as detailed in Algorithm 1.

5. Experiments

We evaluate the proposed approach for a wide range of downstream recognition tasks with pre-trained Transformer backbones. We first describe our experimental setup in Section 5.1, including the pre-trained backbone and downstream tasks and a brief introduction to alternative transfer learning methods.

5.1. Setup

Datasets. We experiment on a collection of 13 datasets from V-tab-1k [42]. VTAB is a collection of dieverse visual classification tasks, which encompasses three distinct categories of tasks: Natural, featuring images taken with conventional cameras; Specialized, containing data acquired through specialized devices, such as medical imaging or satellite sensors; and Structured, which demands spatial reasoning, like counting objects.

Implementation Details. We implement all experiments on NVIDIA A800-80GB GPUs. For a fair comparison, all methods use a ViT-B/16 backbone pre-trained on ImageNet-21k, and the number of prompt tokens is set to 50. We follow the original configurations, *e.g.* number of image patches divided, existence of `[CLS]`, etc. We train the prompt on all the source tasks for 10 epochs for source prompt training. We use 2000 samples from each source task for each target task to compute the transferability loss and gradient alignment loss. Following [42], we use the provided 800-200 split of the train set to determine hyperparameters and run the final evaluation using the full training data.

Baselines. We compare our approach to three recent prompt tuning methods: (1) **Visual prompt tuning (VPT** [16]), where target prompt embeddings are initialized by random (2) **SPoT** [34], where target prompts are initialized by source prompt embeddings trained on other tasks; (3) **ATTEMPT** [2], which mixes pre-trained source prompts and target prompts via attention mechanism. (4) **PANDA** [45], which uses a new metric to predict transferability and employs knowledge distillation. (5) **Adapter** [12], which inserts new MLP modules with residual connection in the side Transformers layers. (6) **SIDETUNE** [43], which trains a "side network and linear interpolate between pre-trained features and side-tuned features before fed into the head. (7) **PARTIAL-k** [44], which fine-tunes the last k layers of backbones while freezing the others. (8) **MLP-k**, which utilizes a multilayer perceptron (MLP) with k lay-

ers, instead of a linear, as classification head. (9) **BIAS** [5], which fine-tunes only the bias terms of a pre-trained backbone.

5.2. Main Results

Our experimental evaluation across 13 diverse vision tasks, as detailed in Tab. 1, demonstrates HGPrompt's superiority over 13 baselines using a ViT-B/16 backbone pre-trained on ImageNet-21k. The proposed method achieves state-of-the-art performance with an average accuracy of 59.6%, surpassing prior multi-source prompt transfer approaches. HGPrompt excels in fine-grained recognition tasks, achieving top results on Flowers102 and Oxford Pets. It also outperforms all baselines in texture analysis on DTD and maintains competitive performance on CIFAR100. Notably, the method establishes new state-of-the-art results in geometric reasoning tasks, including sNORB-Azimuth and dSprite-Orientation, with significant improvements in complex visual reasoning tasks like Clevr-Count. However, performance bottlenecks emerge in specific domains. The 32.1% accuracy on SVHN suggests limitations in handling low-resolution digit recognition. While SPoT retains an advantage in SVHN, HGPrompt exhibits a more balanced and robust performance across all task categories, highlighting its effectiveness.

5.3. Task Transferability Patterns

Fig. 4 illustrates the prompt weights computed in the 13 downstream tasks for four MPT methods. Our proposed metric excels in precisely capturing intrinsic prompt transferability between functions—for example, semantically related tasks such as dsp-Loc (Object location) and dspr-Ori (Object orientation) exhibit significantly higher mutual transferability. In contrast, SPoT does not show any appearant structure in the pairwise prompt transferability, unlike the other three methods that shows self-transfers(diagonal) are more effective than cross-task transfer in general. ATTEMPT and PANDA, on the other hand, produce sparser score distributions, with limited emphasis on specific task pairs such as Clevr-Dist (Distance estimation) and Clevr-Count (Object counting). While these methods capture some task affinities, their overall patterns lack the granularity and consistency observed in our approach. In summary, our method is the only approach that effectively captures task transferability, demonstrating its superiority in modeling prompt transferability.

5.4. Alation Studies

The ablation study in Tab.2 highlights the contributions of each component in the framework. Without H-score or Gradient Loss, the model relies on average weighted source prompts, achieving an average accuracy of 64.4%. Introducing the H-score loss alone improves accuracy to

Method	Cifar100	DTD	Flowers102	Pets	SVHN	EuroSAT	DMLab	sNORB-Azim	sNORB-Ele	dSpr-Loc	dspr-Ori	Clevr-Count	Clevr-Dist	Average
Linear	61.7	58.6	96.6	83.9	32.7	83.9	30.6	12.2	20.3	12.6	18.2	32.1	28.6	44.0
PARTIAL-1 [44]	64.4	60.3	97.5	86.0	36.3	87.8	32.5	16.5	21.8	31.3	_39.2_	41.3	32.1	49.8
MLP-2	39.3	43.0	88.5	76.3	28.0	80.4	29.7	12.5	20.3	24.5	30.8	31.5	29.5	41.1
MLP-3	41.9	46.2	90.5	78.4	30.3	83.9	30.7	14.1	21.5	25.9	33.1	33.8	30.2	43.1
MLP-5	38.1	44.1	90.8	79.1	28.8	81.2	30.5	13.9	20.4	22.5	33.2	33.0	29.1	41.9
MLP-9	38.6	46.1	92.1	81.2	28.0	84.2	31.0	14.7	22.9	19.7	33.2	39.0	28.3	43.0
Adapter [12]	73.8	61.7	97.5	86.6	32.7	85.3	29.4	11.9	19.5	22.4	20.8	40.1	35.1	47.4
SIDETUNE [43]	53.5	58.7	93.4	77.2	17.6	37.2	26.7	10.6	15.1	13.2	13.6	20.3	19.4	35.1
BIAS [5]	70.8	57.5	97.2	85.1	45.3	89.7	31.2	13.5	23.2	63.3	**39.7**	49.1	54.5	56.2
VPT [16]	56.0	57.4	97.3	82.5	61.4	88.9	40.7	15.3	14.1	42.8	37.5	34.8	51.0	52.3
SPoT [34]	**75.6**	_63.7_	_97.7_	_86.3_	_70.4_	_92.1_	37.3	_19.4_	23.3	65.0	36.0	41.5	52.8	58.5
ATTEMPT [2]	67.8	62.1	96.1	85.1	69.0	91.0	36.2	17.9	23.5	61.2	35.0	_43.5_	51.2	56.9
PANDA [45]	74.1	61.3	96.5	86.2	**71.2**	90.8	_37.8_	_19.4_	_24.0_	_67.7_	37.3	42.8	**53.9**	_58.7_
HGPrompt	_75.2_	**64.3**	**97.9**	**87.3**	65.1	**92.5**	**38.2**	**20.2**	**25.1**	**68.2**	38.4	**49.3**	_52.5_	**59.6**

Table 1. Performance comparison across diverse vision tasks using a Vision Transformer (ViT-B/16) backbone pre-trained on ImageNet-21k. The second-best results are underlined, while the best results are highlighted in bold. All reported values represent the average accuracy obtained from three independent runs, with the highest average accuracy achieved by our method.

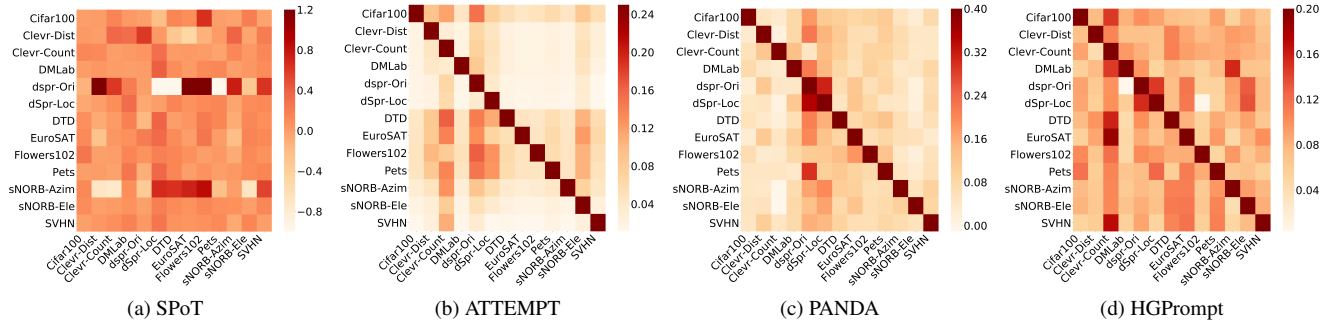(a) SPoT   (b) ATTEMPT   (c) PANDA   (d) HGPrompt

Figure 4. The four heatmaps illustrate the prompt weights between various downstream tasks across different MPT methods.

66.7%, demonstrating its effectiveness in enhancing feature discriminability through information-theoretic optimization. The Gradient Loss alone achieves 66.4% accuracy and stabilizes optimization by aligning gradient directions. Combining both components yields the best performance (67.6%), proving their complementary roles in achieving optimal results.

5.5. Visualization of Learned Features

To analyze the effect of prompt ensemble on the learned representation of the target test data, we present t-SNE visualizations of ViT feature embeddings in four different source prompt setups in Fig.5. In each experiment, three source prompts, pre-trained on CIFAR-100, DTD and Pets datasets respectively, are used to initialize prompts, while our method initializes the target prompt by ensembling these three source prompts. The target domain features used for visualization are computed on the Eurosat dataset. As shown in Fig.5.D, our prompt ensemble method shows better class discriminability than single source prompt transfer. Instead of scattered clusters, objects from the same category form tightly grouped regions with clear separation boundaries despite the vit model never being exposed to EuroSAT images during training. The visualization underscores the effectiveness of our method in constructing a coherent and well-structured feature space for transfer learning.

Table 2. Ablation Study on Framework Components

H-score	Gradient Loss	Cifar100	DTD	Flowers	Pets	SVHN	Euro	DMLab	sN-A	Average
✗	✗	71.5	61.1	93.5	83.8	63.2	89.8	35.5	16.8	64.4
✓	✗	74.3	63.2	97.1	86.5	64.1	91.4	37.1	19.5	66.7
✗	✓	74.1	62.9	96.8	86.2	63.8	91.1	36.8	19.2	66.4
✓	✓	**75.2**	**64.3**	**97.9**	**87.3**	**65.1**	**92.5**	**38.2**	**20.2**	**67.6**

(A) Source Prompt: Cifar100   (B) Source Prompt: DTD

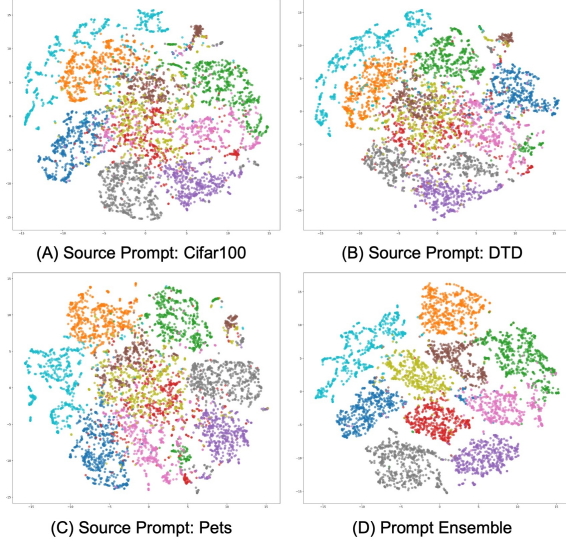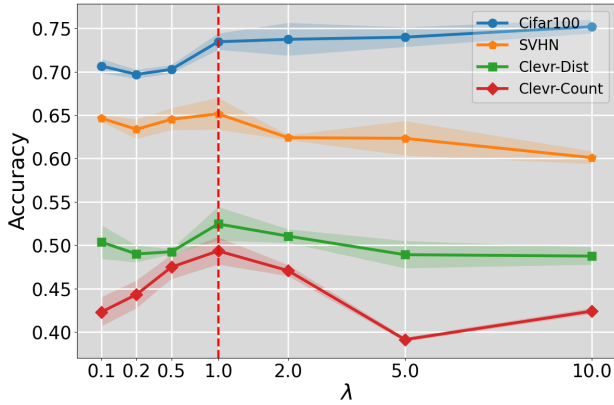(C) Source Prompt: Pets   (D) Prompt Ensemble

Figure 5. t-SNE Visualization of Feature Embeddings with varied source prompt initialization on EuroSAT (10 Classes). Each color corresponds to a distinct land cover class.

Figure 6. Accuracy changes as λ varies from 0.1 to 10.

5.6. Parameter Analysis

To investigate the effect of parameter λ, the coefficient of the alignment regularizer in the total loss function Eq.12, we analyze the model performance in four benchmark data sets: Cifar100, SVHN, Clevr-Dist and Clevr-Count, randomly selected to represent various task domains, as λ varies from

0.1 to 10. Fig.6 illustrates the accuracy trends for each dataset under varying λ, revealing consistent sensitivity patterns.

For Cifar100, accu3racy steadily increases as λ rises from 0.1 to 1.0, stabilizing around 75% for $\lambda \geq 1.0$. However, the other three datasets exhibit a decline in accuracy beyond $\lambda = 1.0$. A critical inflection point occurs at $\lambda = 1.0$ (marked by a red dashed line), where all datasets exhibit either peak performance or trend reversals. This observation suggests that equal weighting of the H-score and gradient alignment regularization generally provide a robust initialization for prompt ensemble tuning. These findings validate the necessity of our dual-objective formulation in balancing feature discriminability and gradient coherence for effective prompt-based transfer learning.

6. Conclusion

In this work, we presented a novel framework for multi-source prompt transfer that addresses the inherent limitations of existing approaches by explicitly optimizing the aggregation of multiple-source prompts. Our methodology determines optimal source weights by maximizing the H-score and minimizing gradient variance, effectively measuring the transferability of the source prompt ensemble. This dual-level evaluation, encompassing feature-level and gradient-level analysis, directly addresses the limitations of existing methods by accounting for cross-interference among prompts and their collective impact on latent feature distributions. Looking forward, future research could explore the extension of our framework to even more diverse sets of source prompts, as well as its application to other modalities beyond vision. Overall, our contributions lay a solid foundation for advancing multi-source prompt transfer, offering both theoretical and practical insights in enhancing foundation model adaptability.

References

[1] Andrea Agostinelli, Jasper Uijlings, Thomas Mensink, and Vittorio Ferrari. Transferability metrics for selecting source model ensembles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7936–7946, 2022. 2

[2] Akari Asai, Mohammadreza Salehi, Matthew Peters, and

Hannaneh Hajishirzi. ATTEMPT: Parameter-efficient multi-task tuning via attentional mixtures of soft prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6655–6672, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 1, 3, 6, 7

[3] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 2

[4] Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE international conference on image processing (ICIP)*, pages 2309–2313. IEEE, 2019. 2

[5] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bit-Fit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland, 2022. Association for Computational Linguistics. 2, 6, 7

[6] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 2

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 1

[8] Yuhe Ding, Bo Jiang, Aijing Yu, Aihua Zheng, and Jian Liang. Which model to transfer? a survey on transferability estimation. *arXiv preprint arXiv:2402.15231*, 2024. 2

[9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[10] Lingyun Feng. Learning to predict task transferability via soft prompt. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8829–8844, Singapore, 2023. Association for Computational Linguistics. 2

[11] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 2

[12] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 6, 7

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[14] Yukun Huang, Kun Qian, and Zhou Yu. Learning a better initialization for soft prompts via meta-learning. *arXiv preprint arXiv:2205.12471*, 2022. 1, 3

[15] Shibal Ibrahim, Natalia Ponomareva, and Rahul Mazumder. Newer is not always better: Rethinking transferability metrics, their peculiarities, stability and performance. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 693–709. Springer, 2022. 2

[16] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 2, 6, 7

[17] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[18] Sungyoon Lee, Hoki Kim, and Jaewook Lee. Graddiv: Adversarial robustness of randomized neural networks via gradient diversity regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2645–2651, 2022. 5

[19] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 1, 3

[20] Kaizhao Liang, Jacky Y Zhang, Boxin Wang, Zhuolin Yang, Sanmi Koyejo, and Bo Li. Uncovering the connections between adversarial transferability and knowledge transferability. In *International Conference on Machine Learning*, pages 6577–6587. PMLR, 2021. 5

[21] Xuannan Liu, Yaoyao Zhong, Yuhang Zhang, Lixiong Qin, and Weihong Deng. Enhancing generalization of universal adversarial perturbation through gradient aggregation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4435–4444, 2023. 5

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[23] Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*, pages 7294–7305. PMLR, 2020. 2

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[25] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30, 2017. 2

[26] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Ar-

naud Stiegler, Teven Le Scao, Arun Raja, et al. Multi-task prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021. 1

[27] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online, 2021. Association for Computational Linguistics. 1

[28] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International conference on machine learning*, pages 9120–9132. PMLR, 2020. 1

[29] Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, Lei Hou, Maosong Sun, and Jie Zhou. On transferability of prompt tuning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3949–3969, Seattle, United States, 2022. Association for Computational Linguistics. 1, 3

[30] Yang Tan, Yang Li, and Shao-Lun Huang. Otce: A transferability metric for cross-domain cross-task representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15779–15788, 2021. 2

[31] Yang Tan, Enming Zhang, Yang Li, Shao-Lun Huang, and Xiao-Ping Zhang. Transferability-guided cross-domain cross-task transfer learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 2

[32] Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1395–1405, 2019. 2

[33] Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. Spot: Better frozen model adaptation through soft prompt transfer. *arXiv preprint arXiv:2110.07904*, 2021. 1, 3

[34] Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou', and Daniel Cer. SPoT: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5039–5059, Dublin, Ireland, 2022. Association for Computational Linguistics. 6, 7

[35] Zhen Wang, Rameswar Panda, Leonid Karlinsky, Rogerio Feris, Huan Sun, and Yoon Kim. Multitask prompt tuning enables parameter-efficient transfer learning. *arXiv preprint arXiv:2303.02861*, 2023. 1

[36] Xiaodong Wu, Wenyi Yu, Chao Zhang, and Phil Woodland. An improved empirical fisher approximation for natural gradient descent. *Advances in Neural Information Processing Systems*, 37:134151–134194, 2024. 5

[37] Yanru Wu, Jianning Wang, Weida Wang, and Yang Li. H-ensemble: An information theoretic approach to reliable few-shot multi-source-free transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15970–15978, 2024. 2

[38] Xiangxiang Xu, Shao-Lun Huang, Lizhong Zheng, and Gregory W Wornell. An information theoretic interpretation to deep neural networks. *Entropy*, 24(1):135, 2022. 2, 4

[39] Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*, pages 12133–12143. PMLR, 2021. 2

[40] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020. 2

[41] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 2

[42] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 2, 6

[43] Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: a baseline for network adaptation via additive side networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 698–714. Springer, 2020. 6, 7

[44] Richard Yi Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *Computer Vision - 14th European Conference, ECCV 2016, Proceedings*, pages 649–666, Germany, 2016. Springer. 6, 7

[45] Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. Panda: Prompt transfer meets knowledge distillation for efficient model adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 36(9):4835–4848, 2024. 1, 2, 3, 6, 7

[46] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2