PROGRESSIVE CONFIDENCE-WEIGHTED MULTI-SOURCE PROMPT DISTILLATION

Yutian Shen, Yanru Wu, Enming Zhang, Zijie Zhao, Yang Li *
Tsinghua Shenzhen International Graduate School †
Tsinghua University
China
shenyutian552@gmail.com,{wu-yr21,zem24,zhaozj24}@mails.tsinghua.edu.cn,
yangli@sz.tsinghua.edu.cn

ABSTRACT

Prompt tuning is a parameter-efficient fine-tuning method designed to address the challenge of adapting large pre-trained models to downstream tasks. However, existing prompt tuning methods exhibit several limitations: (1) difficulties in knowledge transfer when significant domain gaps exist between source and target domains; (2) tendency to forget general knowledge contained in the source model; and (3) susceptibility to overfitting when target data is limited. To address these issues, we adopt the concept of multi-source domain transfer and propose **PCPrompt**, a **P**rogressive **C**onfidence-weighted Multi-source **Prompt** Distillation method for visual prompt tuning under multi-source and few-shot learning setting. By leveraging a confidence-weighted mechanism with knowledge distillation, our approach integrates teacher knowledge according to their respective contributions to the target task. Furthermore, we design dynamic training and progressive decay strategies to provide the student with coarse-to-fine guidance throughout the entire training process. Experimental results demonstrate that our method achieves superior performance compared to existing approaches.

1 INTRODUCTION

The past few years have witnessed the rapid development of large-scale pretrained models (Koroteev, 2021; Radford et al., 2019; Dosovitskiy et al., 2020), and fine-tuning these models on target tasks has become a prevailing practice for performance optimization.As a mainstream parameter-efficient fine-tuning method (Houlsby et al., 2019; Zaken et al., 2021; Hu et al., 2022; Karimi Mahabadi et al., 2021; Li & Liang, 2021), prompt tuning (Liu et al., 2024) only tunes the soft prompt,i.e., a set of trainable parameters added to the pre-trained model,while keeping the backbone fixed.Recent work further proposed prompt transfer (Vu et al., 2021),where prompts pre-trained on source-domains are used to initialize target-domain model before fine-tuning.However,existing approaches face certain limitations: (1) performance drop in data-limited scenario: learning new prompts from scratch (Jia et al., 2022) requires abundant labeled target data,while the initialize-then-finetune prompt transfer method overfits easily with limited training data,and (2) suboptimal knowledge transfer efficacy: directly tuning target prompts initialized with source prompts tend to forget general knowledge from source tasks,and transfer becomes difficult when domain gaps are large. Moreover,existing work taking these problems into account mostly focus on NLP tasks (Vu et al., 2021; Zhong et al., 2024; Asai et al., 2022; Peng et al., 2022),unimodal visual prompt transfer remains underexplored.

To address these challenges,we propose **PCPrompt**(**P**rogressive **C**onfidence-weighted Multi-source **Prompt** Distillation) for visual prompt tuning under multi-source,few-shot scenarios.Leveraging knowledge distillation,our approach not only mitigates catastrophic forgetting in initialize-based prompt transfer,but also provides a feasible solution for data-scarce settings.We use multi-teacher distillation mechanism to integrate knowledge from diverse task domains,avoiding transfer failures when the domain gap between student and a single teacher is too large.Furthermore,we propose a confidence-based weighting strategy that leverages knowledge from both teachers and the student to evaluate task similarity,thereby assigning proper weights to teacher supervision signals.We introduce a decay-controlled phase training strategy that dynamically adjusts teachers' weights while ensuring training stability.

We conduct few-shot transfer experiments on 19 datasets from the VTAB (Zhai et al., 2019) benchmark, where our method achieves superior average performance over all baseline methods, with analysis experiments validating the efficacy of each proposed module. To summarize, our contributions are as follows:

* Corresponding author. Email: yangli@sz.tsinghua.edu.cn
† Shenzhen Key Laboratory of Ubiquitous Data Enabling

- We propose a confidence-based weighting mechanism in prompt transfer with multi-source distillation, which adaptively integrate teacher knowledge depending on their task-specific contributions.
- We develop a dynamic phased training scheme with progressive decay strategy,enabling the student model to receive coarse-to-fine guidance throughout all training phases.
- We conduct comprehensive cross-domain benchmark experiments to demonstrate our method's superior performance over existing approaches,with extensive ablation studies to validate the effectiveness of each proposed component.

2 METHOD

We follow the setting in Visual Prompt Tuning (VPT for short, Jia et al., 2022) to build teacher and student model. We denote $\mathcal{D} = \{x_i, y_i\}_i^{N_s}$ as the target dataset, N_s as the number of samples, C as the number of target classes , K as the number of teachers. The input image x is firstly divided into a sequence of patches and embedded into d-dimensional latent space with positional encoding, then a [CLS] token is added to the sequence to obtain $\tilde{x_i}$. We use $e(\tilde{x_i})^k$ to represent the [CLS] token's embedding after the last Transformer layer of the k-th teacher and $e(\tilde{x_i})^s$ is that of student. M is the pre-trained backbone model without classification head, p^k and p^s is the prompt vectors for the k-th teacher and the student, $head^s$ is the student's classification head. The overall framework is presented in Figure1.

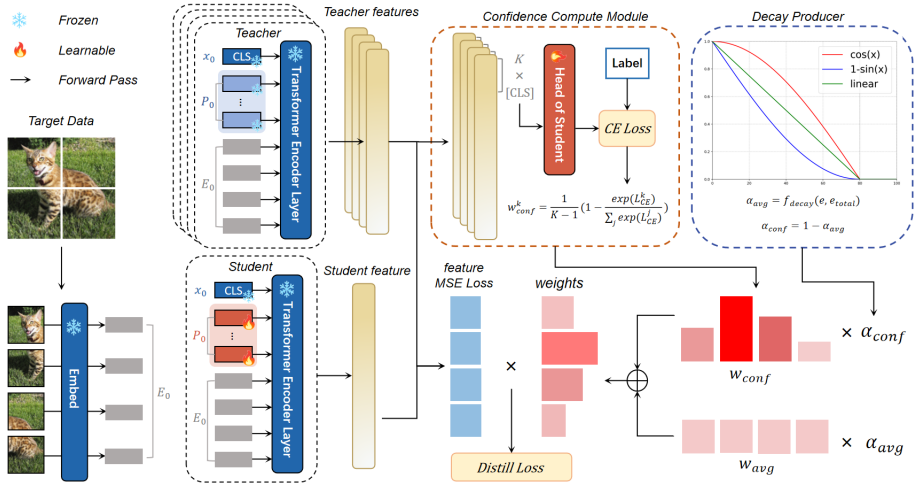

Figure 1: An overview of our **PCPrompt** framework to calculate distillation loss, which is then combined with student cross-entropy loss to form the final training loss.

2.1 CONFIDENCE-BASED TEACHER WEIGHTING

Recognizing that teacher models producing predictions closer to the ground-truth labels of the target task can provide more reliable guidance, we follow Zhang et al., 2022's work to use the cross-entropy loss of teacher predictions on the target data as the weighting factor. However, since teacher and student models serves for different tasks, directly applying the teacher model to obtain target task predictions is infeasible due to dimension mismatch in their classification heads. To address this issue, we leverage features extracted from the teacher's final transformer layer and process them through the student's classification head to generate target predictions.Then we evaluate the loss between such predictions and the ground-truth target labels to calculate the weighting coefficients of each teacher:

$$L_{CE_{conf}}^k = -\sum_{i \in N^s} logP(y_i | e(\tilde{x_i})^k; M, p^k, head^s) \tag{1}$$

$$w_{conf}^k = \frac{1}{K-1}\left(1 - \frac{exp(L_{CE_{conf}}^k)}{\sum_j exp(L_{CE_{conf}}^j)}\right) \tag{2}$$

Notably, in our method, the term *confidence* is not equivalent to the sharpness of the model's output probability distribution. In our approach, the features extracted by the teacher model are fed into the student's classification head to obtain predictions, and the cross-entropy loss is used to compute the teacher's contribution weight, referred to as *confidence* in our work. Consequently, regardless of the sharpness of the teacher's output probability distribution, teachers that produce more incorrect predictions on the target task will be assigned lower weights, thereby reducing their contribution to the training of the target task.

2.2 Decay-Controlled Phase Training

In early training stage, the student's classification head is underdeveloped, which makes the confidence-based weighting mechanism ineffective, reducing weight allocation discriminability. Therefore, we adopt a phased traing strategy: initially using average weighting to avoid bias from poor confidence estimation while ensuring balanced knowledge transfer from all teachers, then switching to confidence-based weighting after several epochs to emphasize more competent teachers.

Considering that directly switching the training strategy at intermediate stage may compromise training stability, we propose a gradual transition strategy to ensure stable training. Specifically, we introduce a decay factor α that progressively decreases from 1 to 0 according to a decay function. We choose cosine as our decay function to ensure greater importance assigned to the average weighting in early stages, and set the decay endpoint at end of the total training epochs. We denote q as current epoch and n as total epochs:

$$\alpha(q) = cosine\left(\frac{min(q,n)}{n} \times \frac{\pi}{2}\right) \tag{3}$$

Furthermore, we allocate the decay coefficients to both weighting methods above. When $\alpha = 1$ (initial state), the average weighting strategy is fully applied; when $\alpha = 0$ (end of the decay period), the confidence-based weighting strategy is used exclusively. We obtain the final weight coefficient of each teacher as follows:

$$w_{total} = \alpha(q) \cdot w_{avg} + (1 - \alpha(q)) \cdot w_{conf} \tag{4}$$

2.3 Training Student Model

The loss function during model training consists of two parts: student loss and distillation loss. Specifically, the student loss is computed as the cross-entropy loss between the predicted values and the ground-truth labels:

$$L_{CE}(p^s, head^s) = -\sum_{i \in N^s} logP(y_i | e(\tilde{x_i})^s; M, p^s, head^s) \tag{5}$$

For distillation loss, given multiple teacher models pre-trained on different source tasks, we first feed the target task data into both teacher models and student model to obtain the output feature vectors of the last Transformer layer. We then calculate the mean squared error (MSE) between teacher features and the student feature, which serves as the distillation loss for individual teacher,where L_{KD}^k denotes the distillation loss between the student and the k-th teacher:

$$L_{KD}^k(p^s, head^s) = -\sum_{i \in N^s} \left(M(\tilde{x_i}|p^k) - M(\tilde{x_i}|p^s)\right)^2 \tag{6}$$

Subsequently, the weighting coefficients for each teacher are calculated using Equation 4. The ultimate distillation loss is obtained by multiplying these final weights with the individual teacher distillation losses computed in Equation 6. The overall training loss for the student model is formulated as follows,where λ is a factor to adjust the transfer ratio:

$$L_{total}(p^s, head^s) = L_{CE}(p^s, head^s) + \lambda \cdot L_{KD}(p^s, head^s) \tag{7}$$

3 Experiment

3.1 Comparison with Baselines

We use 19 tasks of the **VTAB** (Zhai et al., 2019) dataset to test the effect of our method selecting one task from the VTAB dataset as the target task while using all remaining tasks as source tasks. We randomly selects 5 samples per class to construct the few-shot training set. We compare our method with: **VPT** (Jia et al., 2022), **SPoT** (Vu et al., 2021), **ATTEMPT** (Asai et al., 2022) and **PanDa** (Zhong et al., 2024). More details are shown in AppendixB.

Results and Analysis. We report the *top*-1 classification accuracy on these tasks in Table 1. Overall, our method performs the best among all competitors. Compared to the second best method(ATTEMPT), our **PCPrompt** outperforms it with 3.14% average improvement.

For further analysis, we conduct t-SNE (Van der Maaten & Hinton, 2008) feature visualization experiments to examine the class-wise discrimination capacity of the model. We compute t-SNE projections using the [CLS] token obtained from the final Transformer layer output, and compares our method with the second best method(ATTEMPT), shown in Figure2. The figure demonstrates that **PCPrompt** has superior feature clustering ability with tighter intra-class groupings and clearer inter-class separation.

Table 1: Top-1 Accuracy Comparison on tasks of VTAB (5-shot).

Method	Cifar100	Caltech	DTD	Flowers	Pets	SVHN	Sun397	Patch	EuroSAT	Resisc45	Retino	Clevr/Cnt	Clevr/Dist	DMLab	KITTI	dSpr/loc	dSpr/ori	sNORB/azi	sNORB/ele	Average
VPT	32.60	**87.59**	30.73	93.47	3.73	12.79	60.82	66.14	**66.61**	51.37	69.02	16.54	19.82	20.77	42.05	7.18	8.42	**6.72**	11.25	37.24
SPoT	53.59	86.39	44.84	95.22	63.07	14.46	60.28	63.03	26.93	52.17	73.25	15.20	15.47	20.83	41.77	6.26	**8.63**	5.82	12.75	39.99
ATTEMPT	59.61	86.28	**49.26**	95.84	73.13	13.86	60.00	63.59	55.78	37.57	**73.60**	27.05	24.85	22.09	38.96	6.24	7.59	6.67	12.69	42.87
PanDa	56.97	85.32	21.81	92.39	72.66	16.40	60.46	64.69	63.94	50.49	72.81	21.21	**26.71**	23.22	40.51	6.16	7.60	5.57	**15.69**	42.34
PCPrompt	**63.32**	86.87	46.70	**97.51**	**78.99**	**18.15**	**61.09**	**69.05**	66.48	**57.41**	**73.60**	**27.58**	23.17	**24.04**	**45.01**	**7.37**	7.69	6.30	13.90	**46.01**

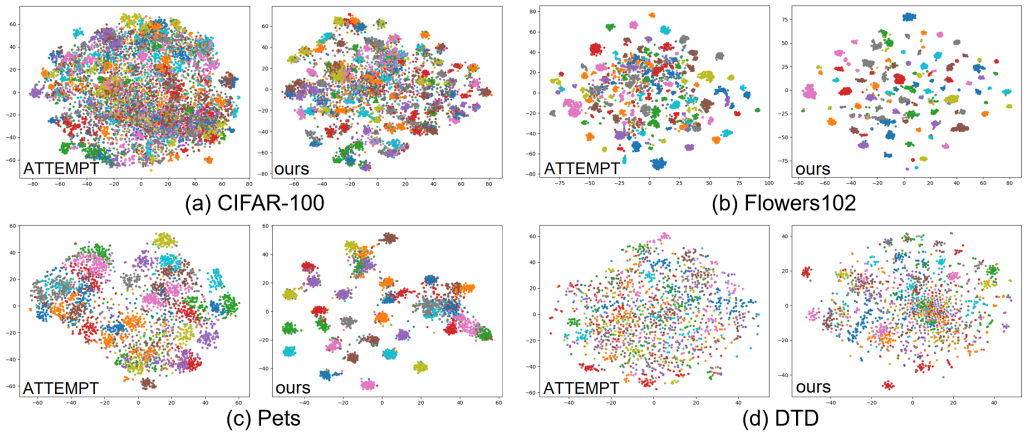

Figure 2: t-SNE visualizations of the final [CLS] embedding of 4 VTAB tasks, **PCPrompt** has a more discriminative feature extracting ability.

3.2 Ablation Study

Through the progressive removal of confidence weighting and phased decay mechanisms, we construct the following method variants:

- w/o confidence-based weighting: only use average teacher weighting strategy to combine teacher features.
- w/o phase training: only use confidence-based weighting to aggregate teacher features.
- w/o decay coefficient: employ only truncated phased distillation, with the first 50% of epochs using average weighting and the remaining 50% using confidence-based weighting.

The results are shown in Table2 of AppendixB. The table shows that our confidence-based weighting trick contributes +1.88% to the average accuracy, and the decay-controlled phase weighting contribures to +0.54% to the performance.

4 Conclusion

In this paper, we introduce a novel **PCPrompt**(Progressive Confidence-weighted Multi-source Prompt Distillation) method to achieve stable and efficient knowledge transfer for unimodal image tasks. We calculate confidence-based weights for teachers to balance their contributions while using distillation loss, alleviating general knowledge forgetting in traditional initialize-based prompt transfer approaches. Moreover, we proposed a decay-controlled phase training strategy to ensure training stability. Comprehensive experiments were conducted on standardized benchmarks to validate the model efficacy, demonstrating the superior performance of our method, offering a new solution for few-shot multi-source transfer learning.

Future Work. Our future research will further explore the application of multi-source prompt distillation to a broader range of challenges. For instance, when addressing tasks with varying levels of difficulty, the optimal prompt length may differ significantly. A promising direction involves leveraging the advantage of knowledge distillation to facilitate knowledge transfer across prompt sequences of different lengths. Additionally, we will further investigate the application of prompt distillation methods in scenarios where the source and target tasks differ in model architectures or even modalities.

REFERENCES

Akari Asai, Mohammadreza Salehi, Matthew E Peters, and Hannaneh Hajishirzi. Attempt: Parameter-efficient multi-task tuning via attentional mixtures of soft prompts. *arXiv preprint arXiv:2205.11961*, 2022.

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

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*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

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*, pp. 709–727. Springer, 2022.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035, 2021.

Mikhail V Koroteev. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*, 2021.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 5:208–215, 2024.

Xiangyu Peng, Chen Xing, Prafulla Kumar Choubey, Chien-Sheng Wu, and Caiming Xiong. Model ensemble instead of prompt fusion: a sample-specific knowledge transfer method for few-shot prompt tuning. *arXiv preprint arXiv:2210.12587*, 2022.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008.

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.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021.

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.

Hailin Zhang, Defang Chen, and Can Wang. Confidence-aware multi-teacher knowledge distillation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4498–4502. IEEE, 2022.

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*, 2024.

A Model Framework

An overview of our **PCPrompt** framewok is shown in Figure1, where the blue part of the model is kept frozen and the red part is learnable. Given an input image, we firstly divide it into a sequence of image patches before embedded into d-dimensional latent space with positional encoding. Then we insert a [CLS] token into the sequence to form the embedding sequence E_0. Given a set of pre-trained teacher models, we keep all the components of the model fixed, and obtain the teacher features of the last Transformer layer using E_0. Then we pass the [CLS] embeddings of teacher features through the head of student to compute cross-entropy loss with ground-truth labels, which is further calculated as confidence weight of each teacher. The decay producer outputs decay coefficients according to current training epoch. After that, we multiply the weight coefficient with decay coefficient of both confidence-based weights and average weights, then calculate the element-wise addition to obtain the final weights. We compute the MSE loss of student feature and each teacher feature, followed by calculating the Hadamard product of losses and weights to get the final distillation loss. Finally, distillation loss together with the student CE loss forms the overall training loss to update the prompt and head parameters of student.

B Experiment Details

Datasets. We use 19 tasks of the **VTAB** (Zhai et al., 2019) dataset to test the effect of our method. VTAB comprises 19 distinct visual classification tasks, categorized into three groups: *Natural*, *Specialized*, and *Structured* tasks. Each task provides 1,000 training samples. We use the 800-200 split of training set to tune model parameters and compute the classification accuracy on test set.

Task Setting. We adopt a multi-source few-shot learning setting. Specifically, we select one task from the VTAB dataset as the target task while using all remaining tasks as source tasks, with a uniform model architecture applied throughout. Teacher models are trained on complete source datasets. For target task fine-tuning, we employ a k-shot sampling strategy (where k=5 in our experiments), randomly selecting k samples per class to construct the few-shot training set. For tasks with faster convergence characteristics (e.g., CIFAR-100), we employ a shorter training schedule of 30 epochs, while other tasks training 100 epochs.

Baselines. We compare our method with: **VPT** (Jia et al., 2022), a baseline approach that directly tunes prompts on the target task with visual prompt; **SPoT** (Vu et al., 2021), a prompt transfer method that initializes target prompts using source prompts before fine-tuning; **ATTEMPT** (Asai et al., 2022), an attention-based approach that aggregates source prompts for target task initialization; and **PanDa** (Zhong et al., 2024), the first work to incorporate knowledge distillation into prompt tuning framework.Although SPoT, ATTEMPT, and PanDa were originally developed for NLP tasks, they share most significant similarities with our task setting. To adapt them for visual tasks, we implement them using a pre-trained ViT-B/16 backbone while strictly preserving their original design specifications for prompt

Table 2: Top-1 Accuracy Comparison (5-shot) for Ablation Study.

Method	Cifar100	Caltech	DTD	Flowers	Pets	SVHN	Sun397	Patch	EuroSAT	Resisc45	Retino	Clevr/Cnt	Clevr/Dist	DMLab	KITTI	dSpr/loc	dSpr/ori	sNORB/azi	sNORB/ele	Average
PCPrompt	**63.32**	**86.87**	**46.70**	**97.51**	**78.99**	**18.15**	**61.09**	**69.05**	**66.48**	**57.41**	**73.60**	**27.58**	23.17	**24.04**	45.01	**7.37**	**7.69**	6.30	13.90	**46.01**
w/o conf.	61.73	86.65	45.48	96.16	74.33	13.80	60.67	65.03	64.2	52.22	70.32	24.23	23.63	21.05	**45.85**	6.39	5.38	**6.69**	**14.72**	**44.13**
w/o phase	57.33	86.31	44.84	95.97	75.58	11.83	60.62	65.42	64.46	51.67	69.40	18.65	19.37	20.66	42.62	5.85	5.04	6.16	14.39	42.95
w/o decay	61.73	86.64	45.48	96.16	74.33	13.78	60.43	65.03	64.20	52.16	70.32	24.21	**23.65**	21.05	33.33	6.38	6.20	6.62	**14.72**	43.49