Confidence-Aware Contrastive Distillation for Test-time Prompt Tuning

Min Wang WANGMINWM@NUDT.EDU.CN
National University of Defense Technology, China

Qing Cheng[*] SGGGPS@163.COM
National University of Defense Technology, China

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

Pre-trained vision-language models like CLIP have shown strong performance on various visual recognition tasks but often suffer from poor generalization under distribution shifts. Test-Time Prompt Tuning (TPT) is a promising solution that adapts prompt embeddings during inference using entropy minimization on unlabeled test data, while keeping the vision and text encoders frozen. However, entropy-based tuning lacks structural regularization and can lead to overconfident misclassifications. In this paper, we introduce Confidence-Aware Contrastive Distillation (CaCoD), a lightweight and effective approach to improve the robustness and calibration of TPT. Our method leverages the confidence structure of test-time predictions by identifying high- and low-confidence samples, and aligning their feature representations through a contrastive distillation loss. This encourages semantically meaningful updates to the prompt embeddings without requiring labels or retraining. Experiments across 11 fine-grained datasets demonstrate that CaCoD consistently reduces calibration error and improves predictive reliability, while maintaining strong accuracy. Our approach is model-agnostic and easily pluggable into existing TPT pipelines.

Keywords: Test Time Prompt Tuning, Vision-Language Model, Confidence Calibration, Contrastive Language–Image Pre-training Model

1. Introduction

Pre-trained vision-language models, such as the Contrastive Language–Image Pre-training (CLIP) model (Radford et al., 2021), have achieved remarkable success across a broad spectrum of recognition tasks by aligning image and text representations within a shared embedding space. However, their performance often degrades significantly when deployed in scenarios involving distribution shifts, where the test-time data distribution differs from that of the training set. Such shifts are common in real-world applications and may arise from factors such as domain changes, image corruption, or variations in visual style (Ovadia et al., 2019).

Test-Time Prompt Tuning (TPT) (Shu et al., 2022) is a recent and effective approach designed to enhance CLIP's robustness by updating the prompt embeddings during inference while keeping the vision and text encoders frozen. Typically, TPT optimizes prompts via entropy minimization on unlabeled test data, aiming to encourage confident predictions.

However, we argue that relying solely on entropy as an optimization signal is insufficient for robust prompt adaptation. Entropy-based objectives treat each sample independently and lack structural regularization, which may lead the model to collapse into overconfident but incorrect predictions—especially when encountering ambiguous or unfamiliar inputs (Mukhoti and et al., 2020).

To address this issue, we propose a simple yet effective mechanism to enhance TPT: **Confidence-Aware Contrastive Distillation** (CaCoD). Our core idea is to exploit the internal structure of test-time data by distinguishing between high-confidence and low-confidence predictions. We assume that samples with high softmax confidence are more likely to be reliable, and thus their representations can serve as stable anchors for distilling knowledge to less certain samples.

Concretely, we partition test-time samples into high- and low-confidence groups based on the softmax output. Then, we extract their corresponding image features using the frozen vision encoder, and apply a contrastive distillation loss that encourages low-confidence features to align with their high-confidence counterparts. This additional regularization term complements entropy minimization by preserving semantic structure in the feature space and guiding prompt updates toward more consistent regions. Importantly, our method is entirely test-time and requires no extra labels, no retraining, and no modification to CLIP's frozen encoders. It introduces only a lightweight feature-level contrastive loss that is easy to integrate into existing TPT pipelines.

We evaluate our method on widely-used fine-grained classification benchmarks. Experimental results demonstrate that our method achieves better calibration and robustness than conventional entropy-based TPT, while maintaining competitive accuracy.

Our contributions are summarized as follows:

- We identify a limitation of entropy-only prompt tuning under test-time shift and propose a confidence-aware regularization mechanism.
- We design a contrastive distillation loss between high- and low-confidence test-time features to guide robust prompt adaptation.
- We show that our method improves both accuracy and calibration under various types of distribution shift, with minimal overhead.

2. Related Works

2.1. Test-Time Prompt Tuning

Prompt tuning has emerged as an effective strategy for adapting large-scale vision-language models (VLMs) without modifying their pre-trained backbones. CoOp (Zhou et al., 2022b) first proposed to learn class-specific continuous prompts for downstream tasks, followed by CoCoOp (Zhou et al., 2022a), which introduced a class-agnostic meta-prompt generator to improve generalization to unseen categories. More recently, test-time prompt tuning (TPT) (Shu et al., 2022; Zhang et al., 2023) has been developed to adapt prompts using unlabeled test data. These methods often leverage entropy minimization or confidence-based filtering to update prompts at test time while keeping the visual and text encoders frozen.

However, existing test-time prompt tuning methods generally treat all samples equally or only consider per-sample confidence scores for sample selection, overlooking the structural relations among samples. In contrast, our approach introduces a contrastive distillation framework that explicitly models high- and low-confidence samples during adaptation, enabling more robust prompt refinement under domain shift.

2.2. Calibration and Confidence Estimation

Modern deep neural networks tend to be poorly calibrated, often producing overconfident predictions (Guo et al., 2017). Numerous calibration techniques have been proposed to address this, such as temperature scaling (Guo et al., 2017), label smoothing, and post-hoc calibration methods (Mukhoti and et al., 2020). While effective in supervised scenarios, these methods typically require access to labeled validation data, making them impractical for test-time scenarios.

Recent test-time adaptation works (Shu et al., 2022) use entropy minimization to implicitly encourage confident predictions. However, these approaches do not directly exploit the internal confidence structure to guide the update. Our method bridges this gap by explicitly using confidence scores to separate reliable and unreliable samples, and then transferring representational structure via contrastive distillation for calibrated prompt refinement.

2.3. Contrastive Learning and Structure-Aware Adaptation

Contrastive learning has shown great success in unsupervised and semi-supervised representation learning (Chen et al., 2020; He et al., 2020). Methods such as SimCLR and MoCo learn features by maximizing agreement between different augmented views. In test-time adaptation, Tent (Wang et al., 2021) proposes entropy-based minimization without explicitly modeling sample-level structure, while others like SHOT (Liang et al., 2020) utilize pseudo-labeling and clustering.

Our approach builds upon this line by introducing a confidence-aware contrastive distillation mechanism that transfers structure from high-confidence representations to guide the adaptation of uncertain samples. To our knowledge, we are the first to leverage contrastive signal across confidence splits within a TPT framework for improving calibration and robustness without requiring additional test-time supervision or multiple views.

3. Methods

In this section, we present our proposed calibration-aware test-time prompt tuning strategy. We first review the standard formulation of Test-Time Prompt Tuning (TPT) and its entropy minimization objective. We then revisit the concept of calibration error and discuss its limitations in TPT. Finally, we introduce our contrastive distillation-based calibration loss that improves calibration without relying on labeled test data.

3.1. Test-Time Prompt Tuning

Given a pretrained vision-language model (e.g., CLIP), we denote the frozen image encoder $\phi_v(\cdot)$ and the frozen text encoder $\phi_t(\cdot)$. In zero-shot classification, the model predicts a label for an image x by computing the similarity between the image feature $\phi_v(x)$ and a set of class-specific text embeddings $\phi_t(c)$.

TPT introduces learnable continuous prompts $\mathbf{P} \in \mathbb{R}^{n \times d}$ appended to textual templates, yielding adapted class embeddings $\phi_t^{\mathbf{P}}(c)$. The prediction probability for class c is:

$$p(c|x) = \frac{\exp(\langle \phi_v(x), \phi_t^{\mathbf{P}}(c) \rangle)}{\sum_{c'} \exp(\langle \phi_v(x), \phi_t^{\mathbf{P}}(c') \rangle)} \tag{1}$$

Since test data are unlabeled, TPT minimizes the entropy of the predictions to encourage confident decisions:

$$\mathcal{L}_{\text{ent}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c} p(c|x_i) \log p(c|x_i) \tag{2}$$

However, this approach can lead to overconfidence on incorrectly predicted samples, especially under distribution shift, which harms model calibration.

3.2. Calibration Metrics and Limitations of TPT

Calibration refers to the alignment between predicted confidence and actual accuracy. A common metric is the Expected Calibration Error (ECE), which measures the discrepancy across confidence bins:

$$\text{ECE} = \sum_{b=1}^{B} \frac{|I_b|}{N} \left|\text{acc}(I_b) - \text{conf}(I_b)\right| \tag{3}$$

where I_b is the set of samples in confidence bin b, $\text{acc}(I_b)$ is the average accuracy in that bin, and $\text{conf}(I_b)$ is the average confidence. A low ECE indicates well-calibrated predictions.

Entropy minimization alone can artificially increase confidence without improving prediction accuracy, resulting in high ECE. Thus, there is a need for test-time regularization strategies that preserve semantic structure and mitigate overconfidence.

3.3. Confidence-Aware Contrastive Calibration

To improve the calibration of Test-Time Prompt Tuning (TPT), we propose a confidence-aware contrastive distillation mechanism that leverages reliable predictions to guide uncertain ones in the feature space. The key insight is that high-confidence samples typically reside in well-structured regions of the embedding space, and thus can serve as implicit supervision to regularize the adaptation process for low-confidence samples.

Entropy-Based Prompt Adaptation. TPT optimizes the prompt embeddings during inference by minimizing the prediction uncertainty on unlabeled test data. This is typically done by minimizing the entropy of softmax outputs:

$$\mathcal{L}_{\text{ent}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} p_{ic} \log p_{ic}, \tag{4}$$

where p_{ic} is the softmax probability of class c for sample x_i, and C is the number of classes. This loss encourages confident predictions by reducing the entropy of the output distribution.

Confidence-Based Sample Partitioning. To add structural guidance to the adaptation process, we partition each test-time batch into a high-confidence subset $\mathcal{X}_h$ and a low-confidence subset $\mathcal{X}_l$, based on a confidence threshold τ derived from the model's own softmax outputs:

$$\mathcal{X}_h = \{x \mid \text{conf}(x) \geq \tau\}, \quad \mathcal{X}_l = \{x \mid \text{conf}(x) < \tau\}. \tag{5}$$

We then extract frozen visual features for each group: $F_h = \phi_v(\mathcal{X}_h)$, $F_l = \phi_v(\mathcal{X}_l)$, where $\phi_v(\cdot)$ denotes the CLIP visual encoder.

Contrastive Distillation Loss. To align the uncertain samples with more reliable structures, we introduce a contrastive distillation loss:

$$\mathcal{L}_{\text{distill}} = \frac{1}{|F_h|} \sum_{i=1}^{|F_h|} \text{CE}\left(\frac{F_h[i]^\top F_l}{\tau}, \mathbf{y}^{(i)} \right), \tag{6}$$

where CE is the cross-entropy loss, τ is the temperature scaling factor, and $\mathbf{y}^{(i)}$ denotes a one-hot identity label indicating alignment of the i-th high-confidence feature with its corresponding low-confidence features.

Overall Objective. The final objective combines the entropy minimization and the proposed contrastive regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ent}} + \lambda_{\text{distill}} \cdot \mathcal{L}_{\text{distill}}, \tag{7}$$

where λ_{distill} balances the two terms. This formulation enables prompt adaptation that not only reduces uncertainty but also improves calibration by enforcing semantic consistency in the learned features. Our approach is fully test-time, requires no labels, and integrates seamlessly into existing TPT pipelines.

4. Experiments

4.1. Datasets

We conduct experiments on 11 widely-used fine-grained classification benchmarks, covering a broad spectrum of visual domains. These datasets include animals and plants (OxfordPets, Flowers102), textures (DTD), food recognition (Food101), scenes (SUN397), vehicles and aircrafts (StanfordCars, FGVC-Aircraft), human actions (UCF101), satellite imagery (EuroSAT), and general object categories (Caltech101, ImageNet). We follow the standard test splits provided by Zhou et al. (2022b), ensuring a consistent and fair comparison with prior Test-Time Prompt Tuning (TPT) methods (Shu et al., 2022).

4.2. Baselines

We compare the following baselines:(1) Zero-Shot CLIP (Radford et al., 2021): No adaptation; directly uses handcrafted text prompts with frozen encoders. (2) TPT (Shu et al., 2022): Test-time prompt tuning based solely on entropy minimization. (3) C-TPT: A calibration-aware extension of TPT that includes a average text feature dispersion regularization term. (4) CaCoD (Ours): Our proposed **Confidence-Aware Contrastive Distillation** method that contrasts high- and low-confidence test samples at the feature level to guide prompt optimization.

All methods use CLIP-RN50 and CLIP-ViT-B/16 as the vision-language backbone. Both the visual and textual encoders are frozen during test-time tuning. Prompt embeddings are optimized using only unlabeled test samples. For a fair comparison, we use a prompt length of 4 tokens (`a photo of a`) and a single test-time tuning step (`TTA step = 1`) for all methods. We fix $\lambda_{\text{distill}} = 0.3$ across these datasets.

Table 1: Comparison on fine-grained classification using CLIP-RN50 and CLIP-ViT-B/16, evaluated by top-1 accuracy (↑) and ECE (↓). Bold entries indicate the best ECE per dataset.

Method		Imag.	Calt.	Pets	Cars	Flow	Food	Airc.	SUN	DTD	SAT	UCF	**Average**
Backbone		CLIP-RN50											
Zero Shot	Acc.	58.1	85.8	83.8	55.7	61.0	74.0	15.6	58.6	40.0	23.7	58.4	55.9
	ECE	2.09	4.33	5.91	4.70	3.19	3.11	6.45	3.54	9.91	15.4	3.05	5.61
TPT	Acc.	60.7	87.0	84.5	58.0	62.5	74.9	17.0	61.1	41.5	28.3	59.5	57.7
	ECE	11.4	5.04	3.65	3.76	13.4	5.25	16.1	9.24	25.7	22.5	12.4	11.7
C-TPT	Acc.	60.2	86.9	84.1	56.5	65.2	74.7	17.0	61.0	42.2	27.8	59.7	57.8
	ECE	3.01	**2.07**	2.77	1.94	4.14	1.86	10.7	**2.93**	19.8	15.1	3.83	6.20
CaCoD	Acc.	59.8	86.8	83.6	56.4	65.5	74.8	16.3	60.5	41.6	25.6	59.7	57.1
	ECE	**2.73**	2.13	**2.34**	**1.23**	**3.53**	**1.43**	**9.47**	3.27	**18.55**	**14.90**	**3.04**	**5.96**
Backbone		CLIP-ViT-B/16											
Zero Shot	Acc.	66.7	92.9	88.0	65.3	67.3	83.6	23.9	62.5	44.3	41.3	65.0	63.7
	ECE	2.12	5.50	4.37	4.25	3.00	2.39	5.11	2.53	8.50	7.40	3.59	4.43
TPT	Acc.	69.0	93.8	87.1	66.3	69.0	84.7	23.4	65.5	46.7	42.4	67.3	65.0
	ECE	10.6	4.51	5.77	5.16	13.5	3.98	16.8	11.3	21.2	21.5	13.0	11.6
C-TPT	Acc.	68.5	93.6	88.2	65.8	69.8	83.7	24.0	64.8	46.0	43.2	65.7	64.8
	ECE	3.15	4.24	1.90	1.59	5.04	3.43	4.36	5.04	11.9	13.2	**2.54**	5.13
CaCoD	Acc.	68.3	93.4	88.7	65.7	70.2	83.0	23.6	64.1	46.1	43.0	64.8	64.6
	ECE	**3.08**	**3.91**	**1.35**	**1.68**	**4.56**	**3.23**	**4.12**	**4.37**	**10.57**	**9.39**	2.85	**4.46**

4.3. Evaluation Protocol

We report the following metrics: (1) Top-1 Accuracy (%): Classification accuracy on the test set. (2) Expected Calibration Error (ECE) (Naeini et al., 2015): Measures the gap between model confidence and accuracy. Lower values indicate better calibration. ECE is computed using 20-bin reliability diagrams. All results are averaged over the entire test set. No labeled data is used for adaptation.

4.4. Results

Table 1 summarizes the results on the 11 fine-grained datasets. For backbone CLIP-RN50, our method achieves the best average ECE (↓ **5.96**) across all datasets, outperforming both the standard TPT and the C-TPT baseline. Notably, we observe consistent gains in calibration on datasets with high domain shift such as DTD, EuroSAT, and UCF101. For backbone CLIP-ViT-B/16, our method achieves the best average ECE (↓ **4.46**). These results demonstrate that incorporating feature-level contrastive signals between confident and uncertain test samples provides meaningful regularization for robust prompt adaptation.

5. Conclusion

We present CaCoD, a simple test-time calibration method that enhances prompt tuning by leveraging confidence-aware contrastive distillation. Unlike entropy-only TPT, CaCoD encourages low-confidence features to align with high-confidence anchors, improving calibration without modifying CLIP's frozen encoders. Our method is lightweight, label-free, and seamlessly integrates into existing TPT pipelines. Experiments on fine-grained benchmarks demonstrate that CaCoD achieves superior calibration and robustness under distribution shift, while preserving accuracy.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020.

Jishnu Mukhoti and et al. Calibrated adversarial training improves robustness and uncertainty calibration. In *ICML*, 2020.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Yaniv Ovadia, Elad Fertig, Jae Ren, et al. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *NeurIPS*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Ting Chen Shu, Roozbeh Mottaghi, Alexander Schwing, et al. Test-time prompt tuning for zero-shot generalization in vision-language models. In *CVPR*, 2022.

Dequan Wang, Fisher Yu, Evan Shelhamer, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Ronghang Zhang, Kun Han, Yixuan Wang, et al. Tip-adapter: Training-free adaptation of clip for few-shot classification. In *ICLR*, 2023.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.