

Supplementary Materials: Towards Robustness Prompt Tuning with Fully Test-Time Adaptation for CLIP’s Zero-Shot Generalization

A APPENDIX

A.1 Algorithm

In this section, we provide the pseudo code for SCP.

Algorithm 1 Self-Text Distillation with Conjugate Pseudo-Labels

Input: Test dataset \mathcal{X}^T , learnable prompt \mathbf{t} , fixed prompt list $\mathbf{t}_{\text{fixed}}$, CLIP image encoder $f(\cdot)$, text encoder $g(\cdot)$, threshold α

while $\mathbf{x}_t \in \mathcal{X}^T$ **do**

 Obtain fixed text feature $\mathbf{g}_{\text{fixed}}$ from text encoder $g(\mathbf{t}_{\text{fixed}})$.

 Obtain text feature \mathbf{g}_t from text encoder $g(\mathbf{t})$.

 Obtain image feature $\mathbf{f}(\mathbf{x}_t)$ from image encoder $f(\mathbf{x}_t)$.

 Calculate the teacher prediction $\mathbf{p}_{\text{teacher}}$ and student prediction $\mathbf{p}_{\text{student}}$ with $\cos(\mathbf{g}_{\text{fixed}}, \mathbf{f}(\mathbf{x}_t))$ and $\cos(\mathbf{g}_t, \mathbf{f}(\mathbf{x}_t))$, respectively.

 Calculate the text loss $\mathcal{L}_{\text{text}}$ from the eq.(14)

 Calculate the entropy of student prediction $H(\mathbf{p}(\mathbf{x}_t))$.

if $H(\mathbf{p}(\mathbf{x}_t)) < \alpha$ **then**

 Obtain the pseudo-label $\mathbf{y}_t^{\text{pseudo}}$.

 Calculate the cross-entropy loss \mathcal{L}_{ce} from eq.(5)

 Calculate the gradient loss $\mathcal{L}_{\text{grad}}$ from eq.(10)

 Calculate the total loss \mathcal{L} from eq.(16), update learnable prompt \mathbf{t}

end if

 Obtain the weighted prompt from eq.(15)

end while

Output: Prediction $p(\hat{y}|\mathbf{x}_t)$

A.2 Additional ablation study

In this section we implement the prompt token length, averaging method and fixed prompt list length ablation study.

Prompt token length. In Figure 1(a), we observe the impact of prompt length on the model’s accuracy for both the ResNet50 and ViT-B/16 architectures. Notably, a prompt length of 4 tokens yields the highest accuracy for both architectures, indicating an optimal balance between providing context and avoiding over-parameterization that can lead to confusion within the model.

Averaging methods. Figure 1(b) compares the influence of different weighting methods on model performance. Gaussian weighting demonstrates superior performance over EMA and BMA for both ResNet50 and ViT-B/16, suggesting that a higher emphasis on recent information beneficially impacts the model’s ability to adapt

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0686-8/24/10

<https://doi.org/10.1145/3664647.3681213>

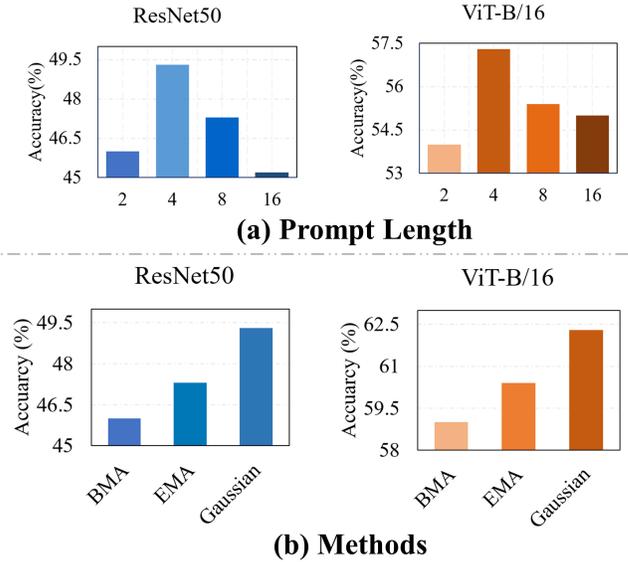


Figure 1: The ablation experiments for prompt length and averaging methods. The average results across three scenarios are analyzed on the ResNet50 and ViT-B/16 architectures.

to new data. **Fixed prompt list length.** In an analysis of the self-text distillation, we examine the influence of the length of the fixed prompt list on model performance. Figure 2 illustrates the ablation experiments conducted for various prompt lengths. The experiments were conducted with two architectures: RN50 and ViT-B/16. With the ViT-B/16 architecture, we observed an initial increase in accuracy as the prompt length increased from 0 to 10. However, further increases in the prompt length led to a gradual decrease in accuracy, with a peak at a prompt length of 30, followed by a plateau. Similarly, the ResNet50 model shows a sharp decline in accuracy as the prompt length increases up to 20, beyond which the accuracy stabilizes and slightly fluctuates. This suggests that, while there is a drop in performance with initial prompt lengthening, extending the prompt beyond a certain point does not significantly alter the accuracy. Therefore, we determined the optimal fixed prompt length to be 10.

Threshold α In Figure 3, for RN50, the accuracy increases as the threshold α rises from 0.1 to 1, reaching a peak at 1. Beyond this value, accuracy declines slightly as α increases to 2.0. Similarly, the peak performance for ViT-B/16 occurs at $\alpha = 1$, after which the accuracy slightly declines but still remains above 62%. This indicates that while both models show optimal performance at an α of 1, they suggest better stability or robustness under varying conditions.

Method	Memory usage/GB	Inference Time/second			Accuracy/%		
	ImageNet	ImageNet	ImageNet-A	ImageNet-R	ImageNet	ImageNet-A	ImageNet-R
TENT	41	625	40.9	272	67.6	48.7	74.6
Pseudo label	43.7	647	42.5	280	66.9	48.6	74.1
MEMO	61.8	845	42.2	293	65.8	48.6	74.4
CoTTA	55.5	13160	443	5107	63.1	47.5	70.5
RMT	57.3	679	42.4	296	63.5	47.7	74.0
SAR	39	786	46.6	309	67.6	48.8	74.4
TPT	41.7	12600	2280	7650	68.4	47.8	77.0
SCP(Ours)	40.4	820	47.6	292	68.8	50.5	70.7

Table 1: Comparison of memory usage and inference time across all baselines, evaluated on ViT-B/16.

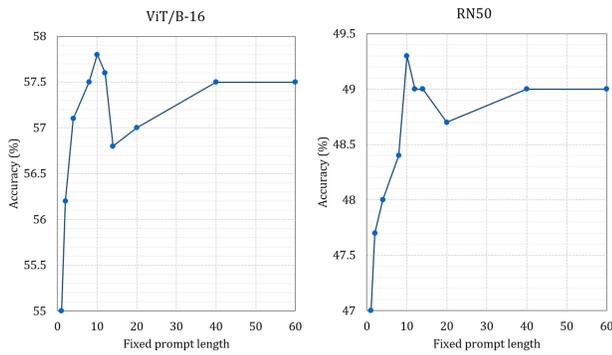


Figure 2: The ablation experiments for fixed prompt length. The average results across three scenarios are analyzed on the ResNet50 and ViT-B/16 architectures.

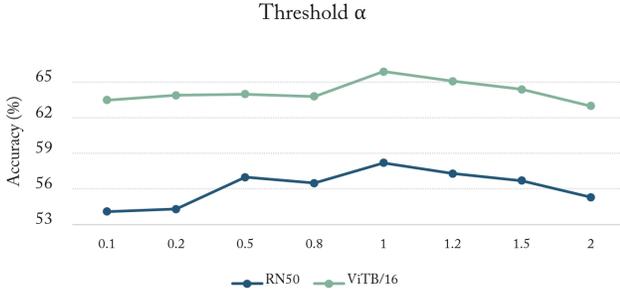


Figure 3: The ablation experiments for threshold α . The average results from the cross-domain generalization benchmark are analyzed using the ResNet50 and ViT-B/16 architectures.

A.3 Memory usage and inference time.

Table 1 results highlight trade-offs in memory usage, inference speed, and accuracy across methods using the ViT-B/16 model. TENT and SAR, which are entropy minimization methods, require less memory at 41 GB and 39 GB, respectively, but offer modest accuracy, with SAR having a slight edge at 67.6% on ImageNet. Teacher-student models such as RMT and CoTTA have higher memory usage, with CoTTA reaching 55.5 GB, and also exhibit prolonged

inference times, which reflects their computational intensity. Specifically, CoTTA’s inference time on ImageNet-A is notably lengthy, at 13160 seconds. TPT employs image augmentation and records the longest inference time at 12600 seconds on ImageNet, implying additional complexity from its enhancement steps. SCP, our proposed method, achieves an optimal balance with the lowest memory usage at 40.4 GB and high accuracy, particularly 70.7% on ImageNet-R. This demonstrates an effective management of distribution shifts with reduced computational demand and underscores the trade-offs between memory, speed, and accuracy across these methods.