

SwapPrompt: Test-Time Prompt Adaptation for Vision-Language Models

Xiaosong Ma¹, Jie Zhang¹, Song Guo², and Wenchao Xu¹

¹Department of Computing, The Hong Kong Polytechnic University (PolyU)

²Department of CSE, The Hong Kong University of Science and Technology (HKUST)



THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學



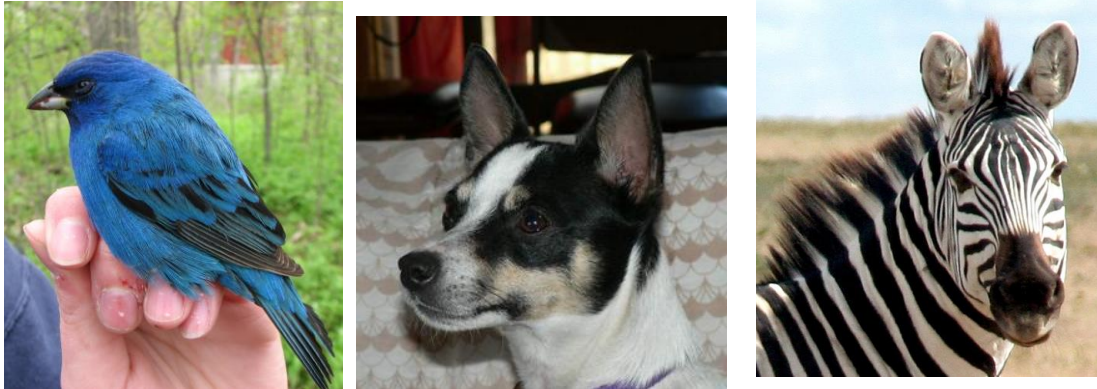
香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY



Test-time adaptation (TTA)

- **Test-time adaptation** (TTA) is a special and practical setting in unsupervised domain adaptation.
- **TTA** allows a pre-trained model in a source domain to adapt to unlabeled test data in another target domain

Source domain



Training phase

Target domain

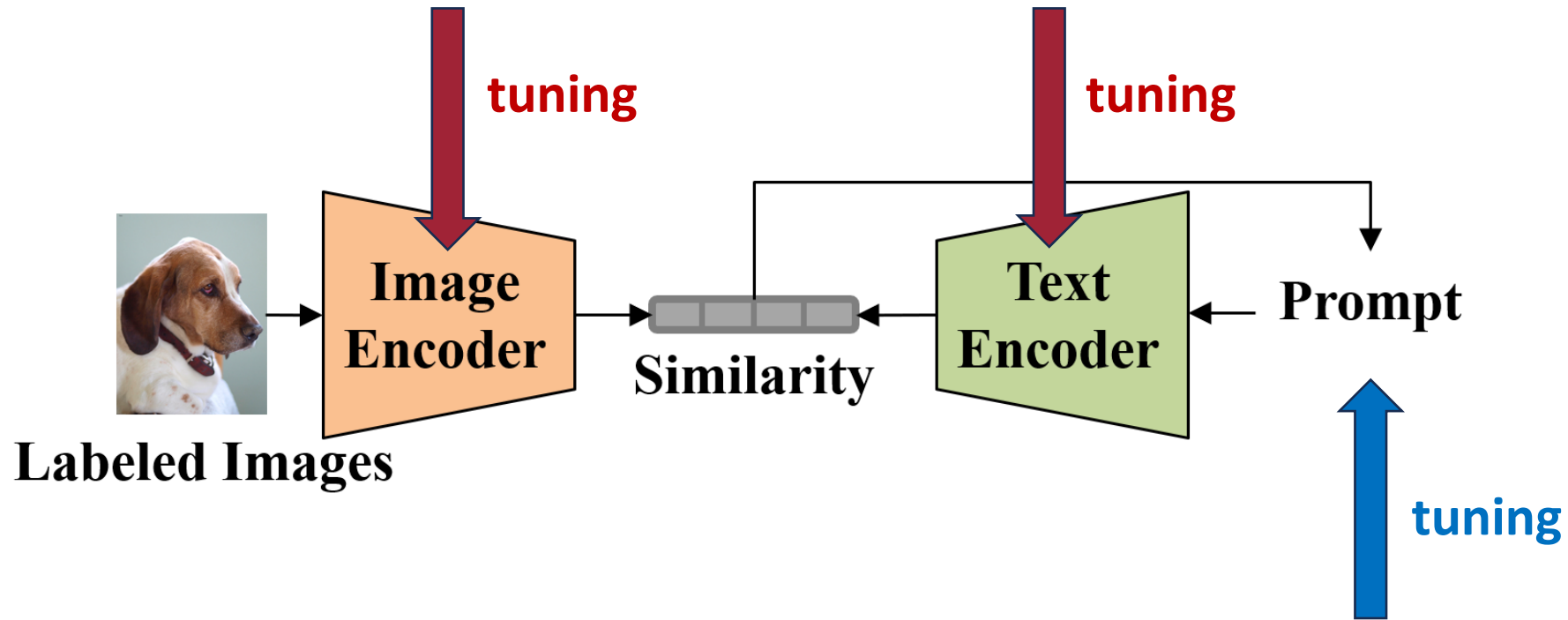


Testing phase

Test-time adaptation

Fine tuning in pre-trained vision-language models

- Traditional model-based TTA methods rely on **computationally intensive tuning** to the parameters of the **model backbone**.

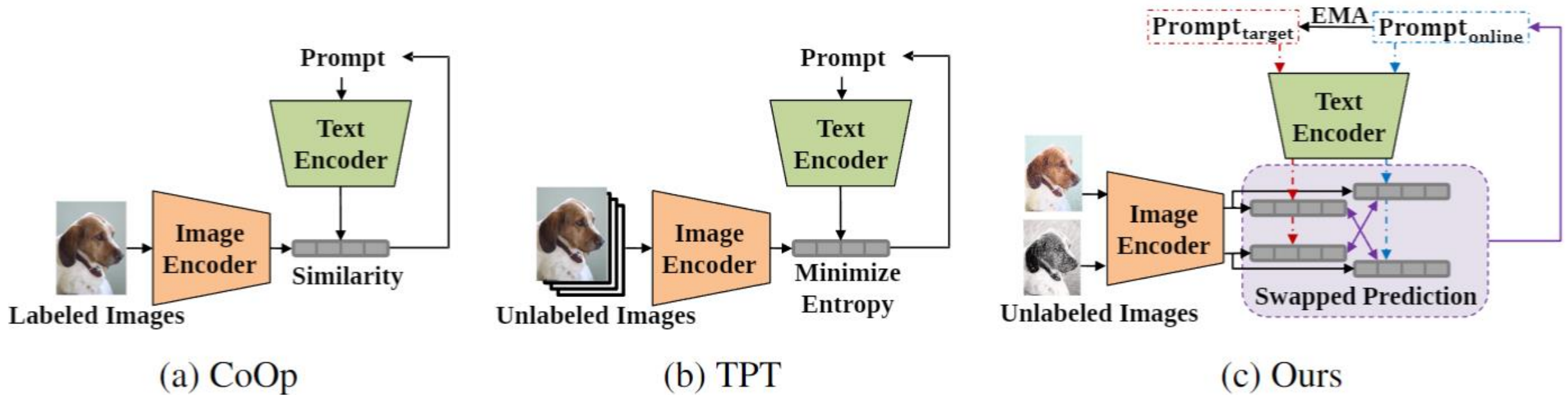


- The emerging pre-trained vision-language models (e.g., CLIP, CoOp) **only tune the run-time prompt** for target domains.

[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." ICML 2021.

[2] Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." IJCV 2022.

Comparison on vision-language model architectures

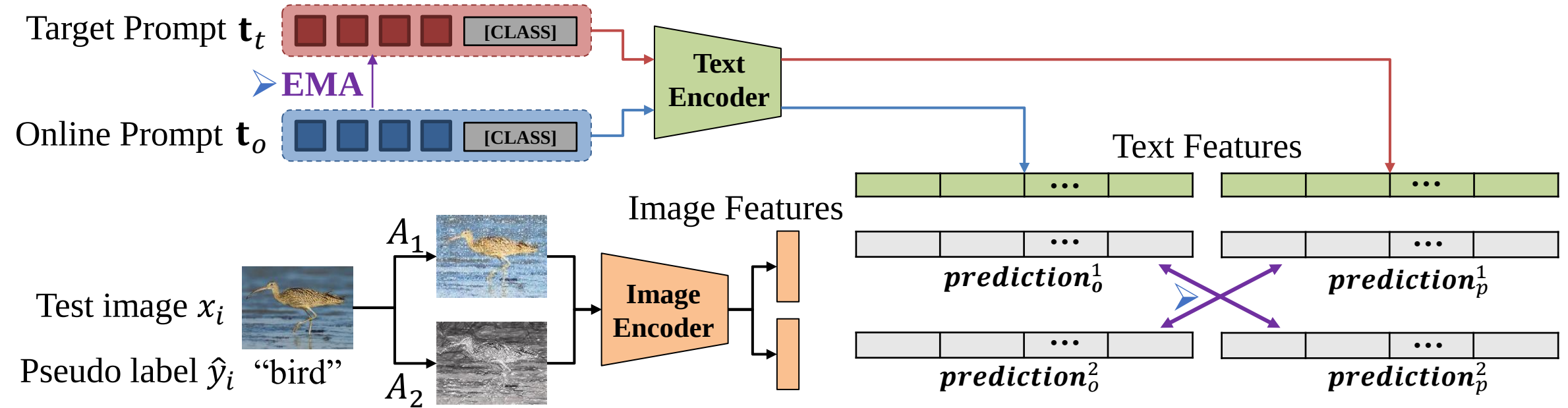


(a) CoOp adapts prompt on labeled data.

(b) TPT optimizes prompt by minimizing marginal entropy.

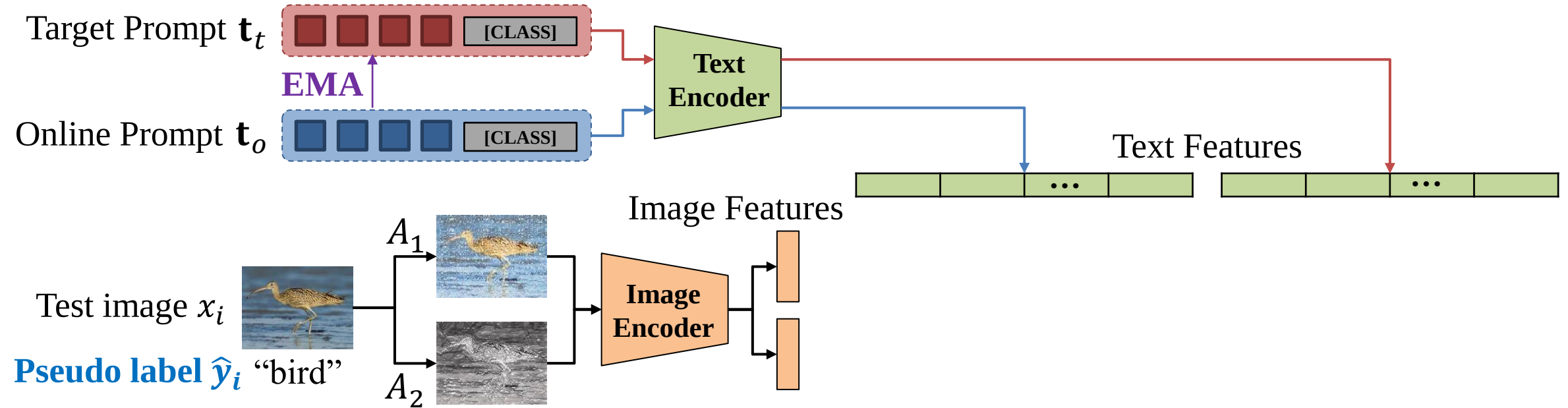
(c) Ours SwapPrompt leverages self-supervised contrastive learning to facilitate test-time prompt adaptation

Methodology: SwapPrompt



- **Exponential moving average (EMA) prompt** : The EMA of online prompt is used to update the target prompt
- **Prompt swapped prediction mechanism**: Based on an augmented view of image and the online prompt, predict the class assignment of an augmented view of the same image

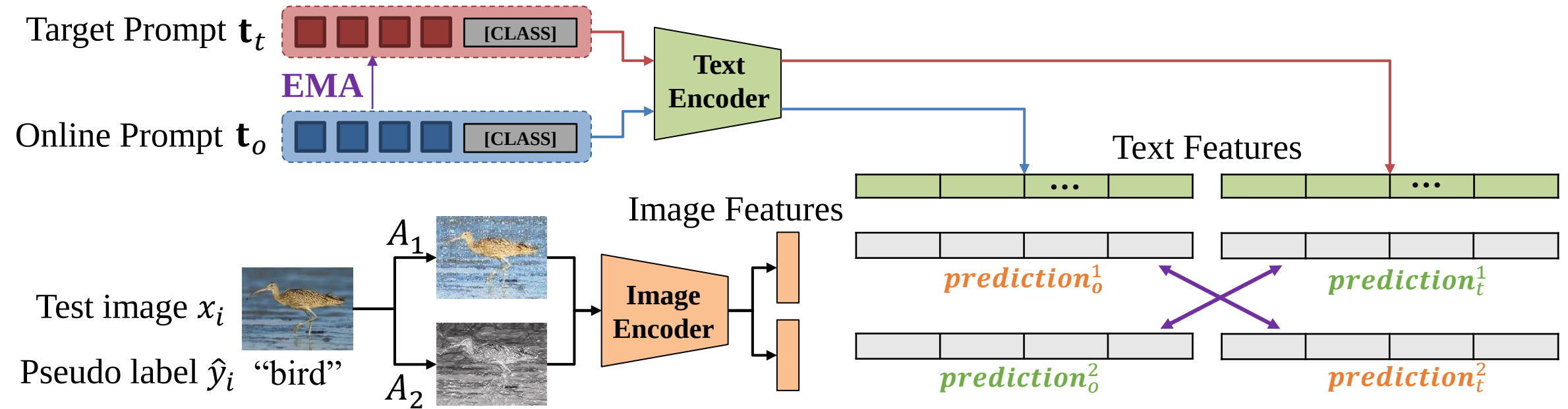
Workflow of FedoSSL



① **Obtain pseudo label:** perform inference on the test data with hand-crafted prompts (e.g., “a photo of a [CLS]”), obtaining their *pseudo-labels* and *classification confidences*

② **Data selection:** filter out potential noisy pseudo labels. For each class, only select the *top K* test data with the highest confidence

Workflow of FedoSSL

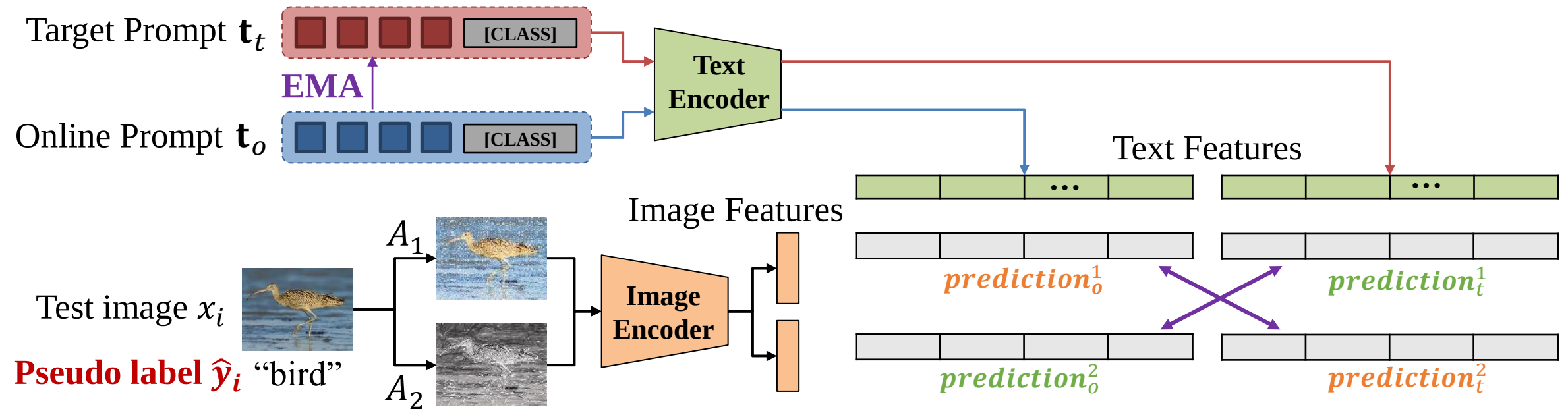


③ Prompt swapped prediction loss function:

$$L_{\text{swap}}(x_i) = \ell(\text{prediction}_o^1, \text{prediction}_t^2) + \ell(\text{prediction}_o^2, \text{prediction}_t^1)$$

We use the text features generated by the target prompt as prototypes and assign the image feature of an augmented view of an image to these prototypes to obtain a soft class assignment. The online prompt is trained to predict this class assignment with a different augmented view of the same image. The EMA of online prompt is used to update the target prompt.

Workflow of FedoSSL



④ Pseudo Label loss function:

$$L_{\text{pseudo}}(x_i) = \ell(\text{prediction}_o^1, \hat{\mathbf{y}}_i) + \ell(\text{prediction}_o^2, \hat{\mathbf{y}}_i)$$

Considering the inherent generalization ability of pre-trained vision-language model, pseudo-labels encapsulate the most pre-trained knowledge. SwapPrompt combines L_{swap} and L_{pseudo} resulting in improved performance compared to using L_{pseudo} alone.

Evaluation Setup

Dataset:

- ImageNet and its four variants: ImageNet-V2, ImageNet-A, ImageNet-R, ImageNet-sketch
- Nine image classification dataset: Caltech101, DTD, Flowers102, Oxford-Pets, UCF101, StanfordCars, Food101, EuroSAT, SUN397

Baselines:

- 1) zero-shot CLIP
- 2) TPT: a test-time prompt tuning method that minimizes the marginal entropy of test data
- 3) UPL: an unsupervised prompt learning approach and we make some modifications on it to suit the test-time setting
- 4) CoOp: a supervised few-shot prompt tuning method, be used as an upper bound performance of test-time prompt adaptation

Performance Comparison to SOTA Baselines

- Comparison of test-time adaptation methods on 14 datasets. Δ denotes SwapPrompt's gain over the better one of UPL and TPT. '+ Online' denotes SwapPrompt with online test data.

Method	Caltech101	DTD	Flowers102	Oxford-Pets	UCF101	StanfordCars	Food101	EuroSAT	SUN397	ImageNet	ImageNet-V2	ImageNet-A	ImageNet-R	ImageNet-Sketch
CoOp [17]	88.76	54.62	83.98	87.44	66.71	61.83	73.79	61.68	64.33	61.23	55.29	23.41	56.96	35.64
CLIP [16]	85.13	42.16	65.40	83.05	61.15	55.65	74.23	37.60	58.55	58.18	51.36	21.69	55.98	33.33
UPL [24]	86.37	45.04	67.11	88.53	63.63	58.46	74.38	41.40	61.07	61.19	52.07	23.59	57.09	36.40
TPT [19]	87.22	42.17	65.42	84.60	61.18	58.49	74.88	43.82	61.46	60.74	54.35	26.24	58.72	35.02
SwapPrompt	89.90	47.34	70.22	89.14	65.66	59.60	75.08	46.64	63.93	61.80	53.94	24.46	60.88	38.21
Δ	+2.68	+2.30	+3.11	+0.61	+2.03	+1.11	+0.20	+2.82	+2.47	+0.61	-0.41	-1.78	+2.16	+1.81
+ Online	89.69	46.40	68.12	88.97	64.52	58.88	75.66	42.45	63.36	61.41	52.93	24.42	60.25	38.13

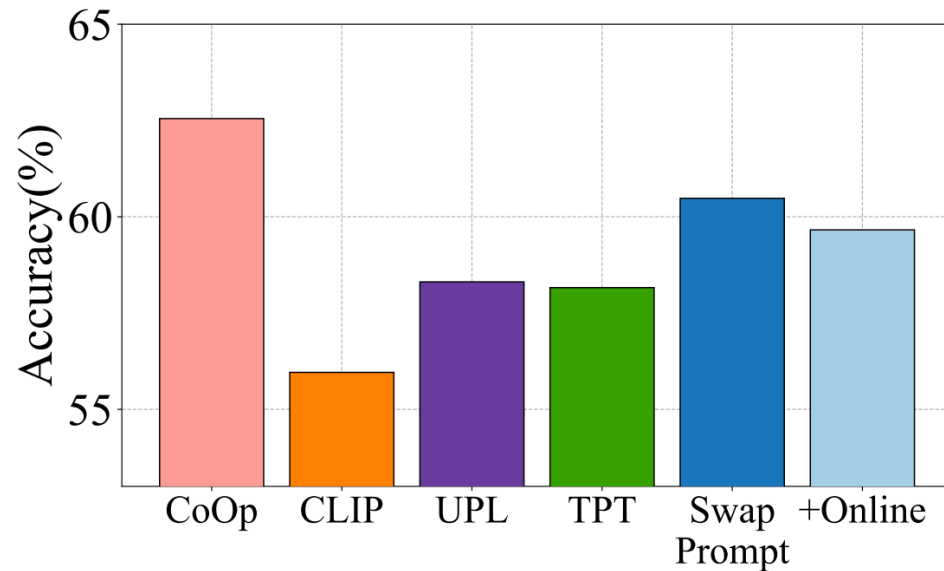
SwapPrompt vs. SOTA Baselines:

- Outperforms TPT by 2.31% accuracy
- Outperforms UPL by 2.17% accuracy
- Very close even outperform CoOp on many datasets

Performance Comparison to SOTA Baselines

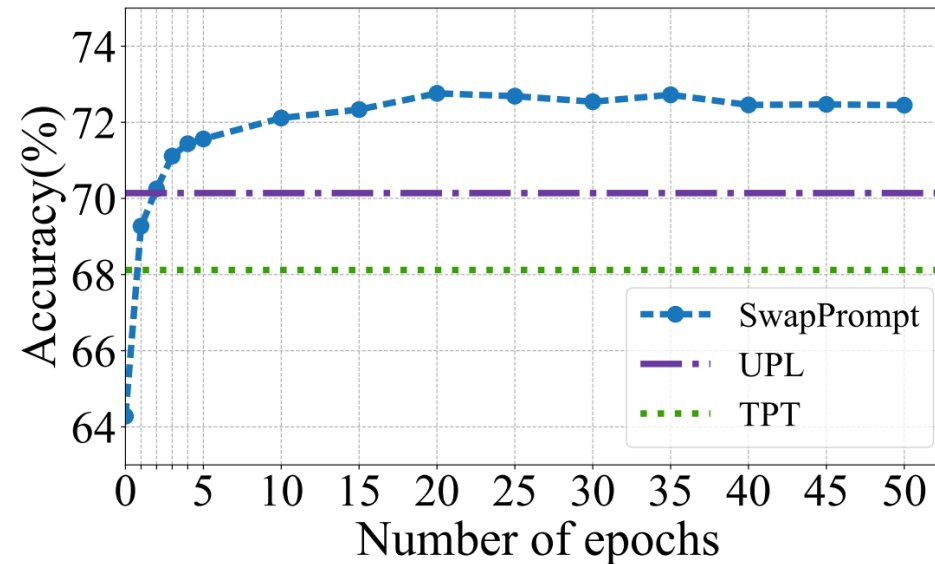
Accuracy and Efficiency of SwapPrompt:

(a) final accuracy and (b) the relationship between the epoch and the accuracy



(a) Average accuracy on 14 datasets

- The average accuracy on all 14 datasets, CoOp is compared as an upper bound.



(b) Accuracy of different epochs on 5 datasets

- The average accuracy of SwapPrompt on 5 datasets with different adaptation epochs, the accuracy of UPL and TPT is the final epoch average accuracy.

Thank you!

xiaosong16.ma@connect.polyu.hk