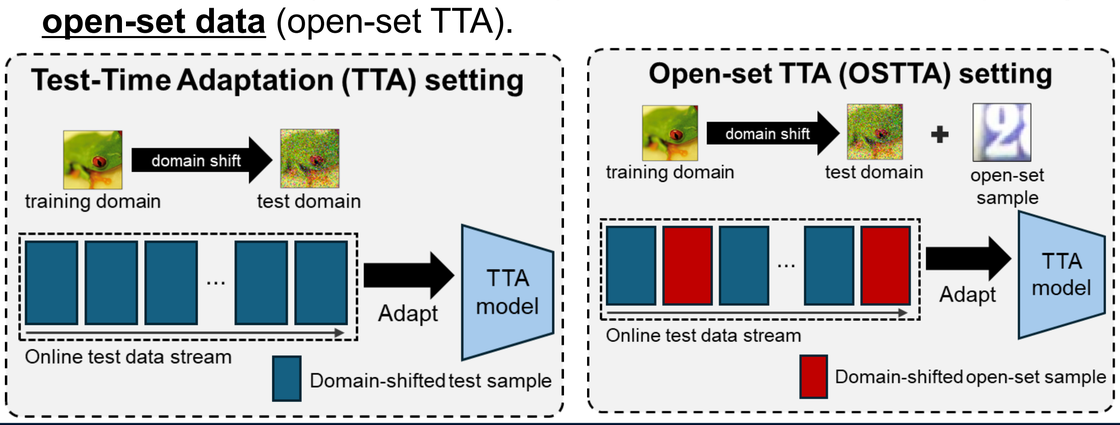
Stabilizing Open-Set Test-Time Adaptation via Primary-Auxiliary Filtering and Knowledge-Integrated Prediction

Byung-Joon Lee, Jin-Seop Lee, Jee-Hyong Lee†
Sungkyunkwan University, Republic of Korea

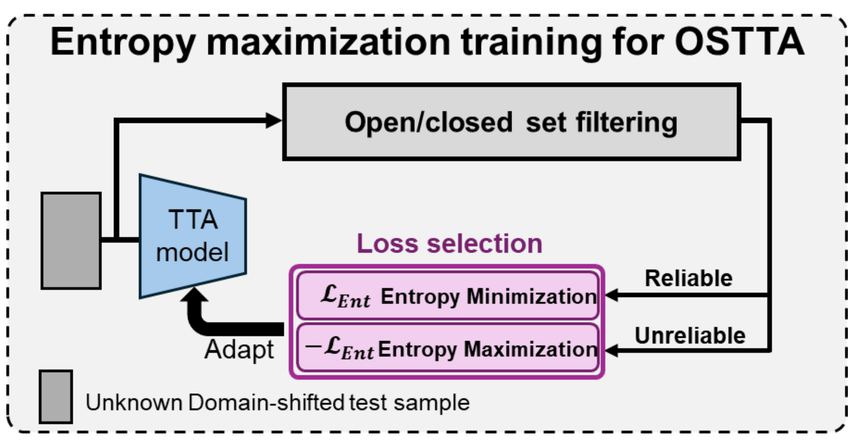
BMVC 2025

Introduction

- **Test-Time Adaptation** (TTA) aims to adapt the model to the test domain on the fly during inference.
- While most TTA studies assume that the training and test data share the same class set (closed-set TTA), real-world scenarios frequently include **open-set data** (open-set TTA).

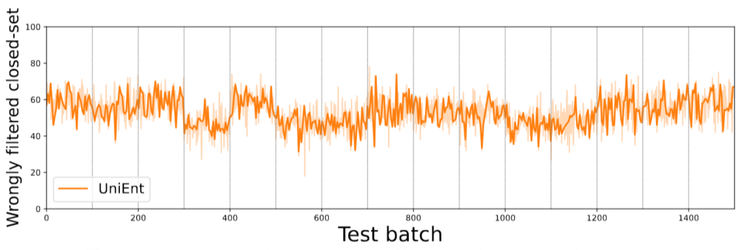


- While most OSTTA studies simply discard low-confidence data during adaptation, one research demonstrates that **maximizing the entropy of low-confidence data** can be effective.

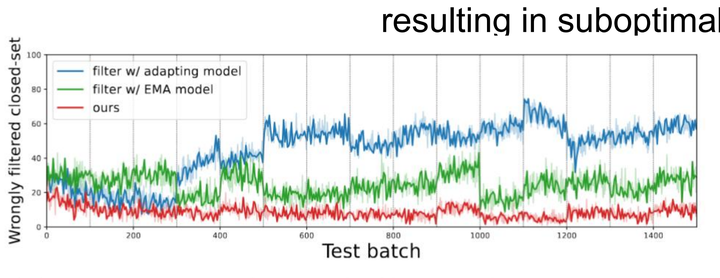


Motivation

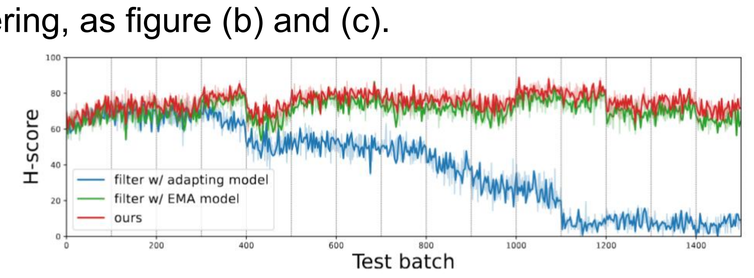
- Filtering of previous method rely solely on fixed source model.
- However, Incorporating **domain knowledge into filtering is crucial**, as domain-shifted test samples often exhibit low confidence when evaluated by a source model not trained on the target domain.
- In practice, prior studies have misclassified approximately 50% of closed-set data as open-set samples, as figure (a).
- One way to leverage domain knowledge into filtering is to utilize the adapting model itself.
- However, relying on the adapting model's confidence for loss selection leads to significant error accumulation (red curve).
- We found that the **exponential moving average (EMA) model** mitigates error accumulation but fails to promptly capture domain shifts, thereby resulting in suboptimal filtering, as figure (b) and (c).



(a). Percentage of wrongly filtered closed-set samples w/ prior study.



(b). Percentage of wrongly filtered closed-set samples with adapting, EMA, and our filter.



(c). H-score with adapting, EMA, and our filter.

Proposed Method

1. Primary-Auxiliary Filtering (PAF)

- Using the adapting/EMA model individually for filtering has distinct limitations.
- Therefore, **PAF** leverages both EMA model, and adapting model for filtering:

$$(1)\quad F_{pr}(x) = \mathbb{I}[H(f_{\theta_t}(\tilde{x})) < \tau] \qquad (2)\quad F_{aux}(x) = \mathbb{I}[H(f_t^{EMA}(\tilde{x})) < \tau]$$

(i). If $F_{pr}(x) = 0$, Sample x is identified as a potential open-set sample.

- However, directly applying entropy maximization leads to error accumulation.

➢ **Entropy maximization is applied only when $\mathbf{F_{aux}(x) = 0}$**. Otherwise, the sample is excluded from training.

(ii). If $F_{pr}(x) = 1$, sample x is identified as a closed-set sample.

- But we found that applying double filtering may lead to overly conservative training.
- Nevertheless, relying solely on $F_{pr}(\cdot)$ may still result in incorrect training.

➢ We leverage the **confidence of EMA model** into entropy minimization by **loss weighting** term $w_{soft}(x)$:

$$(3)\quad \omega_{soft}(x) = \frac{1}{\exp[\,H(f_t^{EMA}(\tilde{x})) - \tau]}$$

Final optimization objective:

$$(4)\quad \min_{\theta_t} \mathcal{L}_{PAF}(x) \sum_{x \in X_t} \mathbb{I}[F_{pr}(x) = 1] \cdot \omega_{soft}(x) \cdot H\left(f_{\theta_t}(\tilde{x})\right) - \alpha \cdot \mathbb{I}\left[(F_{pr}(x) = 0) \wedge (F_{aux}(x) = 0)\right] \cdot H\left(f_{\theta_t}(\tilde{x})\right)$$

2. Knowledge-Integrated Prediction (KIP)

- Despite PAF significantly improves filtering accuracy, some samples are still incorrectly filtered.
- Although the source model is not adapted to the target domain, it has never encountered open-set samples.

➢ We **ensemble** outputs of **adapting/EMA/source models** by **comparing their confidence** to the average across all models:

$$(5)\quad c_i(x) = 1/3 + \gamma \cdot (\max(p_i(x)) - \Sigma_{i=1}^{3} \max(p_j(x))/3)$$

$$(6)\quad z_{kip}(x) = \Sigma_{i=1}^{3} c_i(x) \cdot z_i(x)$$

Pseudo-code

Input: Source pre-trained model f_{θ_0}, test domain dataset $D_t = \{x_i, y_i\}_{i=1}^{N_i}$, batch size $B = |B_t|$.
for t ← 1 to T **do**
  for $x_i \in B_t$ **do**
    Compute $F_{pr}(x_i)$ via eq. (1) // Primary filter
    if $F_{pr}(x_i) = 1$ **then**
      Compute $w_{soft}(x_i)$ via eq. (3) // Soft filtering
    if $F_{pr}(x_i) = 0$ **then**
      Compute $F_{aux}(x_i)$ via eq. (2) // Hard filtering
  Compute $\sum_{i=1}^{B} \mathcal{L}_{PAF}(x_i)/B$ via eq. (4)
  Update θ_t
  Compute $z_{kip}(x_i)$ via eq. (6) // KIP
  Update f_t^{EMA}
  return $z_{kip}(x_i)$

Experimental Results

➢ **CIFAR10/100-C (closed-set) + SVHN/Tiny-ImageNet/Places365/Textures-C (open-set)**

CIFAR 10-C	SVHN-C			TinyImageNet-C			Places365-C			Textures-C		
	ACC	AUROC	H-S	ACC	AUROC	H-S	ACC	AUROC	H-S	ACC	AUROC	H-S
Source	81.73	77.89	79.76	81.73	80.43	81.07	81.73	83.60	82.65	81.73	82.60	82.16
TENT	78.05	67.64	72.47	56.91	62.44	59.55	49.31	55.83	52.37	65.92	64.68	65.29
CoTTA	85.23	82.91	84.05	85.88	79.20	82.40	85.01	82.71	83.84	83.17	79.44	81.26
EATA	84.47	82.24	83.34	84.97	78.07	81.37	84.65	80.63	82.59	82.45	76.27	79.24
RoTTA	85.08	82.97	84.01	85.12	80.09	82.53	85.09	82.33	83.69	84.80	80.89	82.80
OSTTA	84.11	71.98	77.57	85.29	71.18	77.60	84.69	73.09	78.46	83.07	66.79	74.05
STAMP	82.95	73.27	77.81	86.83	79.72	83.12	85.65	80.19	82.83	81.48	66.09	72.98
UniEnt	84.56	90.21	87.29	85.19	84.43	84.81	84.74	89.42	87.02	82.36	85.60	83.95
Ours	**87.49**$_{+2.26}$	**97.66**$_{+4.72}$	**92.30**$_{+4.07}$	**88.26**$_{+1.43}$	**90.61**$_{+5.08}$	**89.42**$_{+4.57}$	**88.35**$_{+2.70}$	**94.24**$_{+3.48}$	**91.20**$_{+3.82}$	**87.80**$_{+3.00}$	**97.71**$_{+9.35}$	**92.49**$_{+7.86}$

CIFAR 100-C	SVHN-C			TinyImageNet-C			Places365-C			Textures-C		
	ACC	AUROC	H-S	ACC	AUROC	H-S	ACC	AUROC	H-S	ACC	AUROC	H-S
Source	53.25	60.55	56.67	53.25	68.91	60.08	53.25	66.92	59.31	53.25	62.37	57.45
TENT	20.76	67.69	31.77	20.91	56.87	30.58	19.85	59.50	29.77	22.86	61.13	33.28
CoTTA	55.44	74.07	63.42	57.61	68.97	62.78	55.21	71.94	62.47	52.13	65.01	57.86
EATA	52.81	70.79	60.49	54.25	68.00	60.35	53.69	70.81	61.07	49.75	57.88	53.51
RoTTA	59.74	75.53	66.71	60.00	70.74	64.93	60.05	70.14	64.70	59.27	72.33	65.15
OSTTA	60.24	75.68	67.08	61.60	71.20	66.05	60.35	72.72	65.96	59.57	66.43	62.81
STAMP	59.51	92.61	72.46	**64.25**	73.16	68.42	62.89	74.23	68.09	54.32	39.79	45.93
UniEnt	59.24	88.91	71.10	60.34	73.87	66.42	59.11	78.10	67.29	57.02	73.91	64.38
Ours	**62.59**$_{+2.35}$	**97.61**$_{+5.00}$	**76.27**$_{+3.81}$	63.79$_{-0.46}$	**84.16**$_{+9.95}$	**72.57**$_{+3.81}$	**62.94**$_{+0.05}$	**85.25**$_{+5.98}$	**72.42**$_{+4.33}$	**62.81**$_{+3.24}$	**94.54**$_{+19.99}$	**75.48**$_{+10.33}$

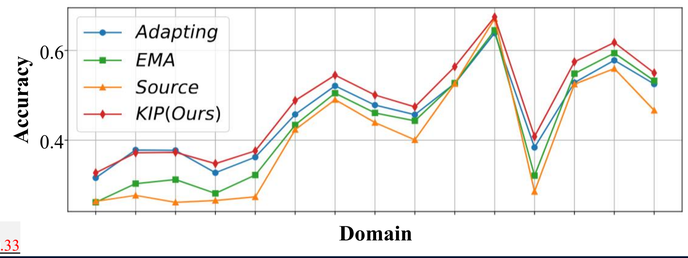
➢ **ImageNet-C + Places365/Textures-C**

Image Net-C	Places365-C			Textures-C		
	ACC	AUROC	H-S	ACC	AUROC	H-S
Source	28.22	67.86	39.86	28.22	71.55	40.48
TENT	41.28	65.91	50.77	35.12	47.30	40.31
CoTTA	41.41	72.64	52.75	40.81	70.05	51.57
EATA	46.99	78.95	58.91	43.42	65.62	52.26
RoTTA	44.15	72.14	54.78	45.13	77.60	57.07
OSTTA	47.05	75.62	58.01	45.03	61.73	52.07
STAMP	46.88	80.11	59.15	47.49	80.59	59.76
UniEnt	46.16	78.81	58.22	44.46	64.01	52.47
Ours	**48.22**$_{+1.17}$	**84.18**$_{+4.07}$	**61.32**$_{+2.17}$	**47.96**$_{+0.47}$	**82.91**$_{+2.32}$	**60.77**$_{+1.01}$

➢ **Ablation study of PAF and KIP**

F_{pr}	F_{aux}	KIP	ACC	AUROC	H-S
✗	✗	✗	20.76	67.69	31.77
✓	✗	✗	26.27	82.98	39.91
✗	✓	✗	55.83	97.11	70.90
✓	✓	✗	58.27	**97.61**	72.98
✓	✓	✓	**62.59**	**97.61**	**76.27**

➢ **Accuracy of KIP v.s. independent models**



➢ **Inference Time v.s. H-Score**