南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

A Versatile Framework for Continual Test-Time Domain Adaptation: Balancing Discriminability and Generalizability

Xu Yang*, Xuan Chen*, Moqi Li, Kun Wei, Cheng Deng†
School of Electronic Engineering, Xidian University, Xian 710071, China
{xuyang.xd, moqili14, weikunsk, chdeng.xd}@gmail.com,
x.chen@stu.xidian.edu.cn

CVPR 2024

Background

① First, we should explore supervisory signals without labels to improve performance in the current target domain, where widespread noisy pseudo labels are a crucial factor affecting performance.

② Then, the adaptation process means moving the initial source parameterization to a parameterization that better models the current target distribution, which carries the risk that predictions on the source distribution become inaccurate, causing catastrophic forgetting.

③ Finally, an excessive affinity for the existing domain will cause generalization to be lost for future domains when the current target distribution is narrow, especially under noisy pseudo labels.

Motivation

Existing continual test-time adaptation implementations may face a game between discrimination in the current domain and generalization in future domains. Meanwhile, the absence of label signals makes the model performance worse when facing domain shifts.

Thus, one research question is how to construct a novel pipeline that ensures generalization and improves discrimination, and the other is how to capture knowledge from the source pre-trained model.

Methods

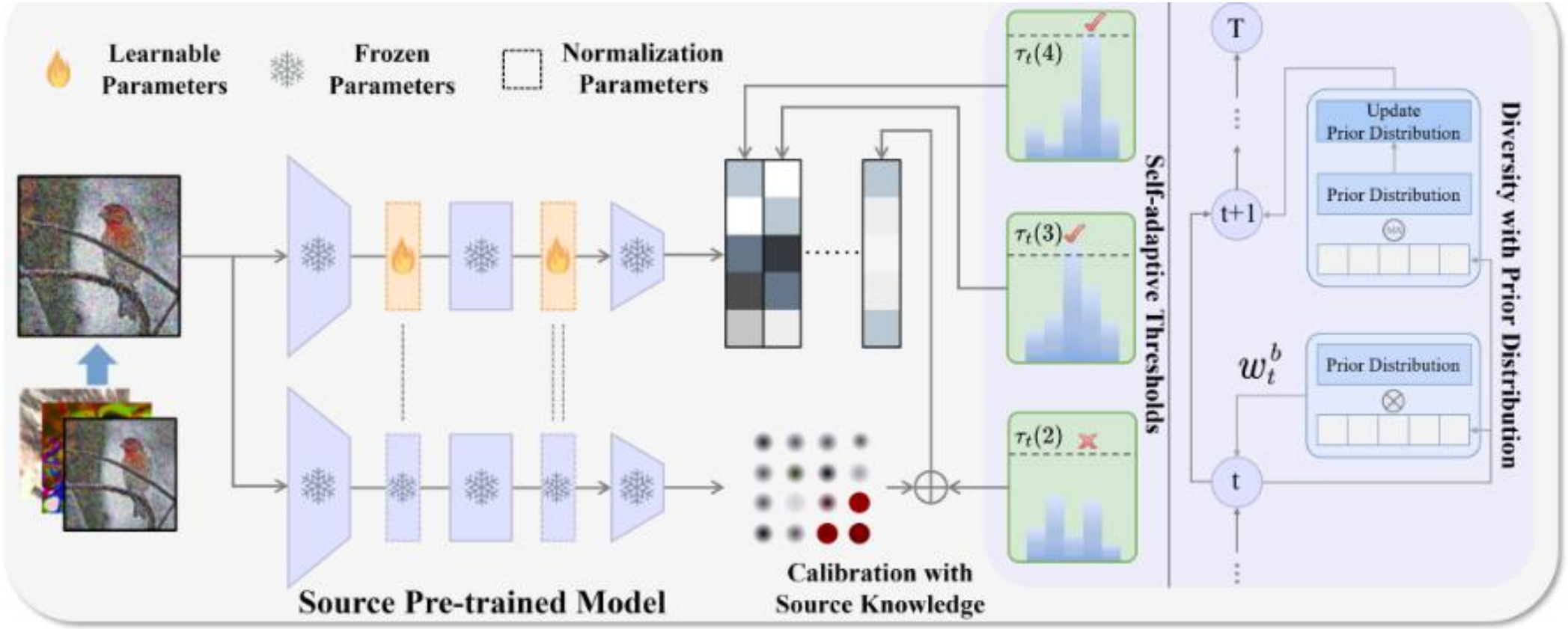


Figure 1. This is the flow of our method. We propose a novel pipeline to optimize the network's normalization parameters to ensure long-term generalization and improve instantaneous discrimination, and confidence thresholds are utilized in a self-adaptive manner to select reliable labels. Then, we explore various prior knowledge from the source pre-trained model to calibrate and enrich supervision signals. Moreover, we track the recent tendency of a model's prediction with an exponential moving average for a diversity score to ensure subsequent generalization. Finally, the learnable parameters are aligned with the source parameters in a soft-weighted manner to alleviate catastrophic forgetting.

Methods

1、High-quality Supervision Generator

$$p_t^b = \texttt{Softmax}(F_{\theta_t}(x_t^b)),$$
$$\mathcal{L}_{ce}(X_t) = -\frac{1}{B}\sum_{b=1}^{B} y_t^b \log p_t^b, \tag{1}$$

where p_t^b represents classification result of the sample b at time t, and y_t^b is the supervision signal of the ith sample.

global threshold

We set the global threshold τ_t as the average confidence from the model, and estimate the global confidence at each stage t. τ_t is defined and adjusted as:

$$\tau_t = \frac{1}{B}\sum_{b=1}^{B} \max(y_t^b). \tag{2}$$

local threshold

Except for the global threshold, the local threshold is utilized to modulate the global threshold in a class-specific fashion to account for the intra-class diversity and the possible class adjacency. We compute the expectation of the model's predictions on each class to estimate the class-specific learning status:

$$\xi_t(c) = \frac{1}{B}\sum_{b=1}^{B} y_t^{b,c}, \tag{3}$$

where $c \in C$ is the number of classes. After integrating the global and local thresholds, we can obtain the final self-adaptive threshold of each class c.

$$\tau_t(c) = \frac{\xi_t(c)}{\max\{\xi_t(c) : c \in C\}}\tau_t. \tag{4}$$

Based on such thresholds, the samples at the current batch can be divided into two parts, the reliable part $N_{rel}(t) = \{b|b \in B, \max(y_t^b) \geq \tau_t(\arg\max y_t^b)\}$ and the unreliable one $N_{unrel}(t) = \{b|b \in B, \max(y_t^b) < \tau_t(\arg\max y_t^b)\}$.

Methods

2、Calibration with Source Knowledge

we attempt to distill knowledge from the source pre-trained model to calibrate the unreliable signals.

$$f_t^b = \texttt{Softmax}(F_\theta(x_t^b)), s_t^{b,d} = \text{sim}(f_t^b, f_t^d), \quad (5)$$

where f_t^b and f_t^d are the representations of the sample b and d at time t. $\text{sim}(\cdot)$ represents the consine similarity. Here, the set of the K nearest neighbors $N_{neg}^b(t), b \in N_{unrel}(t)$ are selected by $s_t^{b,d}$ for the sample, and the calibrated pseudo-labels are calculated.

$$y_t^b = \frac{1}{\sum_{d \in N_{neg}^b(t)} s_t^{b,d} + 1} \sum_{d \in N_{neg}^b(t)} s_t^{b,d} * y_t^d + y_t^b. \quad (6)$$

1、**校准预测**：通过比较当前样本与相似样本的预测结果，可以校准当前样本的伪标签。这个过程有助于捕捉多样化的监督信号，从而提高模型在新领域的泛化能力。

2、**捕获多样化监督信号**：通过校准伪标签，模型可以更好地适应新数据的分布，捕获更丰富的监督信息。这有助于模型在面对新领域的数据时，减少过拟合的风险，提高其预测的准确性和鲁棒性。

3、**相似性矩阵**：通过建立相似性矩阵，可以衡量批次中样本之间的相似度。这有助于识别哪些样本是相似的，哪些是不相似的，从而为不可靠的样本找到可靠的邻居样本，用于计算伪标签。

Methods

3、Diversity with Prior Distribution

The output results of the network may become biased or collapse to a trivial solution after a narrow distribution during test time

diversity weighting is employed by tracking the recent tendency of a model's prediction with an exponential moving average.

$$\bar{y}_{t+1} = \alpha \bar{y}_t + \frac{1-\alpha}{B} \sum_{i=1}^{B} y_t^b, \tag{7}$$

where $\alpha = 0.9$. To determine a diversity weight for each test sample, the cosine similarity between the current model output y_t^b and the tendency of the recent outputs $\bar{y}_t$ is calculated as follows.

$$u_t^b = 1 - \frac{\bar{y}_t^\top y_t^b}{\|y_t^b\| \|\bar{y}_t^b\|}. \tag{8}$$

u_t^b has the advantage that if the model output is uniform, uncertain predictions receive a smaller weight, mitigating errors in the model. More importantly, certainty weighting based on negative entropy is employed to avoid bias towards specific classes.

$$v_t^b = y_t^b \log y_t^b. \tag{9}$$

We normalize the certainty and diversity weights to be within the unit range, and exponentiate the product of diversity and certainty weights, scaled by a temperature τ. Thus, the weight of each sample can be obtained.

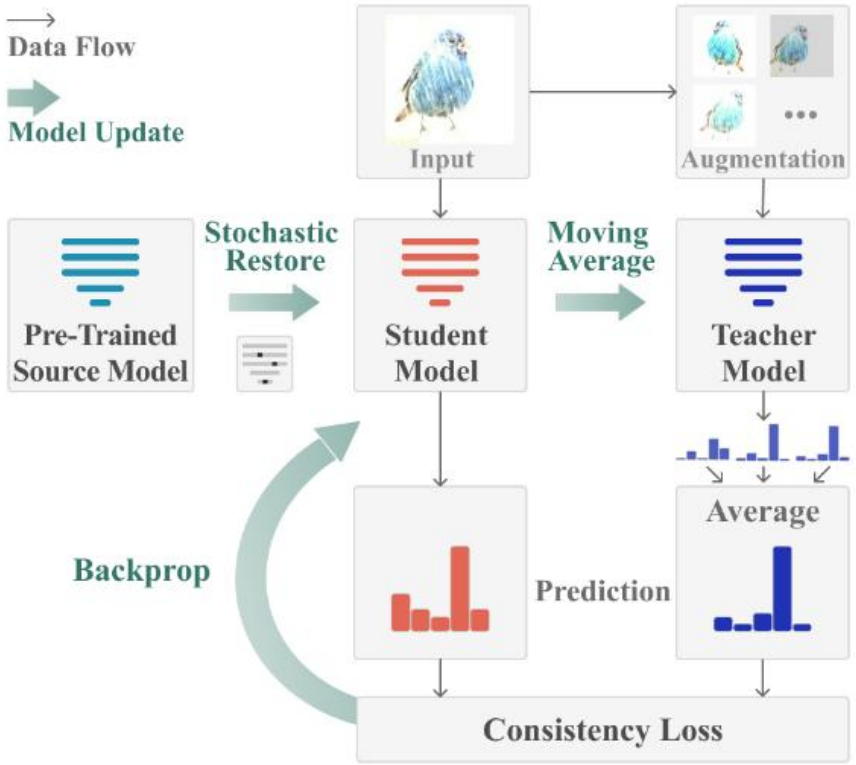
$$w_t^b = \exp\left(\frac{u_t^b \cdot v_t^b}{\tau}\right). \tag{10}$$

After the above selection, calibration and weighting, the objective in Eq. 1 can be reconstructed as follows.

$$\mathcal{L}_{ce}(X_t) = -\frac{1}{B} \sum_{b=1}^{B} w_t^b y_t^b \log p_t^b, \tag{11}$$

Methods

4、Soft-weighted Parameter Alignment



we construct a soft parameter alignment function and incorporate it into the loss function to optimize the network. This ensures that the network parameters are highly correlated with those of the source pre-trained model during loss optimization, rather than being overwritten afterward

$$\mathcal{L}_{pa}(\theta_t) = \sum_l \mathbf{1}[l \in \mathrm{BN}] \cdot \beta^l \left\| \theta_t^l - \theta^l \right\|_2^2, \qquad (12)$$

where l is the layer of the network and β^l repressents the similarity strength of l-th layer. We set $\beta^l = \frac{1-e^{-10l}}{1+e^{-10l}}$, which is increased with the deeper layers.

3.3. Overall

The overall objective of our method is as follows.

$$\mathcal{L}(X_t) = \mathcal{L}_{ce}(X_t) + \lambda_1 \mathcal{L}_{pa}(\theta_t), \qquad (13)$$

where λ_1 is the hyperparameter. In general, we do not directly use the results of pre-trained and adapted models as supervision signals, but apply them as prior knowledge to calibrate pseudo-labels, and design a soft-weighted parameter alignment method to prevent excessive parameter deviation.

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Time		t →																
Method	Backbone	*Gaussian*	*Shot*	*Impulse*	*Defocus*	*Glass*	*Motion*	*Zoom*	*Snow*	*Frost*	*Fog*	*Brightness*	*Contrast*	*Elastic_trans*	*Pixelate*	*Jpeg*	Mean	Gain
Source	ResNet	72.3	65.7	72.9	46.9	54.3	34.8	42.0	25.1	41.3	26.0	9.3	46.7	26.6	58.5	30.3	43.5	-
BN Stats Adapt		28.1	26.1	36.3	12.8	35.3	14.2	12.1	17.3	17.4	15.3	8.4	12.6	23.8	19.7	27.3	20.4	+23.1
Pseudo-Label		26.7	22.1	32.0	13.8	32.2	15.3	12.7	17.3	17.3	16.5	10.1	13.4	22.4	18.9	25.9	19.8	+23.7
TENT-continual [ICLR'21]		24.8	20.5	28.5	14.5	31.7	16.2	15.0	19.2	17.6	17.4	11.4	16.3	24.9	21.6	26.0	20.4	+23.1
CoTTA [CVPR'22]		24.6	21.9	26.5	11.9	27.8	12.4	10.6	15.2	14.4	12.8	**7.4**	11.1	18.7	13.6	17.8	16.5	+27.0
NOTE [NeurIPS'22]		**7.3**	**7.4**	**12.5**	20.9	**13.8**	15.5	34.2	34.2	39.6	25.0	11.6	24.2	29.9	14.1	**12.7**	20.1	+23.4
RoTTA [CVPR'23]		30.3	25.4	34.6	18.3	34.0	14.7	11.0	16.4	14.6	14.0	8.0	12.4	20.3	16.8	19.4	19.3	+24.2
RMT [CVPR'23]		24.1	20.2	25.7	13.2	25.5	14.7	12.8	16.2	15.4	14.6	10.8	14.0	18.0	14.1	16.6	17.0	+26.5
ROID [2023.6.1]		23.7	18.7	26.4	**11.5**	28.1	12.4	**10.1**	14.7	14.3	12.0	7.5	9.3	19.8	14.5	20.3	16.2	+27.3
Ours		20.7	17.1	20.2	12.1	24.3	**11.6**	10.9	**13.8**	**12.9**	**10.5**	8.1	**9.3**	**17.9**	**13.4**	15.3	**14.5**	**+29.0**
Source	ViT-base	60.1	53.2	38.3	19.9	35.5	22.6	18.6	12.1	12.7	22.8	5.3	49.7	23.6	24.7	23.1	28.2	-
CoTTA [CVPR'22]		58.7	51.3	33.0	20.1	34.8	20.0	15.2	11.1	11.3	18.5	4.0	34.7	18.8	19.0	17.9	24.6	+3.6
VDP [AAAI'23]		57.5	49.5	31.7	21.3	35.1	19.6	15.1	10.8	10.3	18.1	4.0	27.5	18.4	22.5	19.9	24.1	+4.1
ViDA [2023.6.7]		52.9	47.9	19.4	11.4	31.3	13.3	7.6	7.6	9.9	12.5	3.8	26.3	14.4	33.9	18.2	20.7	+7.5
ROID [2023.6.1]		20.8	14.5	10.5	9.3	20.3	10.2	8.3	7.9	7.4	9.6	4.1	9.2	13.0	10.9	15.5	11.4	+16.8
Ours		**16.3**	**11.1**	**9.6**	**8.4**	**14.6**	**8.6**	**5.5**	**6.3**	**5.7**	**7.1**	**3.3**	**5.4**	**10.9**	**7.7**	**12.8**	**8.9**	**+19.3**

Table 1. Classification error rate (%) for the standard CIFAR10-to-CIFAR10C continual test-time adaptation task. All results are evaluated with the largest corruption severity level 5 in an online fashion. **Bold** text indicates the best performance. Blue is the suboptimal solution.

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Time		t →																
Method	Backbone	*Gaussian*	*Shot*	*Impulse*	*Defocus*	*Glass*	*Motion*	*Zoom*	*Snow*	*Frost*	*Fog*	*Brightness*	*Contrast*	*Elastic_trans*	*Pixelate*	*Jpeg*	Mean	Gain
Source	ResNet	73.0	68.0	39.4	29.3	54.1	30.8	28.8	39.5	45.8	50.3	29.5	55.1	37.2	74.7	41.2	46.4	-
BN Stats Adapt		42.1	40.7	42.7	27.6	41.9	29.7	27.9	34.9	35.0	41.5	26.5	30.3	35.7	32.9	41.2	35.4	+11.0
Pseudo-Label		38.1	36.1	40.7	33.2	45.9	38.3	36.4	44.0	45.6	52.8	45.2	53.5	60.1	58.1	64.5	46.2	+0.2
TENT-continual [ICLR'21]		37.2	35.8	41.7	37.7	50.9	48.5	48.5	58.2	63.2	71.4	72.0	83.1	88.6	91.6	95.1	61.6	-15.2
CoTTA [CVPR'22]		40.1	37.7	39.7	26.8	38.0	27.9	26.5	32.9	31.7	40.4	24.6	26.8	32.5	28.1	33.8	32.5	+13.9
NOTE [NeurIPS'22]		**28.4**	32.7	36.4	44.4	42.9	42.2	65.8	61.1	70.8	51.6	34.4	45.4	62.7	39.9	36.4	43.3	+3.1
RoTTA [CVPR'23]		49.1	44.9	45.5	30.2	42.7	29.5	26.1	32.2	30.7	37.5	24.7	29.1	32.6	30.4	36.7	34.8	+11.6
RMT [CVPR'23]		40.2	36.2	36.0	27.9	33.9	28.4	26.4	28.7	28.8	31.1	25.5	27.1	28.0	26.6	29.0	30.2	+16.2
ROID [2023.6.1]		36.5	31.9	33.2	**24.9**	34.9	26.8	**24.3**	28.9	28.5	31.1	**22.8**	24.2	30.7	26.5	34.4	29.3	+17.1
Ours		33.5	**31.8**	**31.2**	25.9	**30.9**	**25.2**	25.9	**27.9**	**27.4**	**30.6**	25.2	**23.5**	**26.6**	**26.2**	**27.2**	**27.9**	**+18.5**
Source	ViT-base	55.0	51.5	26.9	24.0	60.5	29.0	21.4	21.1	25.0	35.2	11.8	34.8	43.2	56.0	35.9	35.4	-
CoTTA [CVPR'22]		55.0	51.3	25.8	24.1	59.2	28.9	21.4	21.0	24.7	34.9	11.7	31.7	40.4	55.7	35.6	34.8	+0.6
VDP [AAAI'23]		54.8	51.2	25.6	24.2	59.1	28.8	21.2	20.5	23.3	33.8	7.5	11.7	32.0	51.7	35.2	32.0	+3.4
ViDA [2023.6.7]		50.1	40.7	22.0	21.2	45.2	21.6	**16.5**	17.9	16.6	25.6	11.5	29.0	29.6	34.7	27.1	27.3	+8.1
ROID [2023.6.1]		45.7	32.2	20.5	22.2	37.8	24.6	17.2	16.8	15.8	23.2	**10.6**	28.3	29.1	33.2	26.2	25.6	+9.8
Ours		**38.2**	**31.8**	**18.2**	**20.8**	**34.3**	**20.3**	17.5	**14.9**	**16.2**	**22.9**	11.5	**27.5**	**28.2**	**32.5**	**25.3**	**24.0**	**+11.4**

Table 2. Classification error rate (%) for the standard CIFAR100-to-CIFAR100C continual test-time adaptation task. All results are evaluated with the largest corruption severity level 5 in an online fashion. **Bold** text indicates the best performance. Blue is the suboptimal solution.

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Time		t →																
Method	Backbone	*Gaussian*	*Shot*	*Impulse*	*Defocus*	*Glass*	*Motion*	*Zoom*	*Snow*	*Frost*	*Fog*	*Brightness*	*Contrast*	*Elastic_trans*	*Pixelate*	*Jpeg*	Mean	Gain
Source	ResNet	97.8	97.1	98.2	81.7	89.8	85.2	78.0	83.5	77.0	75.9	41.3	94.5	82.5	79.3	68.5	82.0	-
CoTTA [CVPR'22]		84.5	82.0	80.4	81.8	79.5	69.2	58.8	60.8	61.1	48.5	36.5	67.5	47.8	41.8	45.9	63.1	+18.9
RoTTA [CVPR'23]		88.3	82.8	82.1	91.3	83.7	72.9	59.4	66.2	64.3	53.3	35.6	74.5	54.3	48.2	52.6	67.3	+14.7
RMT [CVPR'23]		79.9	76.3	73.1	75.7	72.9	64.7	56.8	56.4	58.3	49.0	40.6	58.2	47.8	43.7	44.8	59.9	+22.1
ViDA [2023.6.7]		79.3	74.7	73.1	76.9	74.5	65.0	56.4	59.8	62.6	49.6	38.2	66.8	49.6	43.1	46.2	61.2	+20.8
ROID [2023.6.1]		71.7	62.2	62.2	69.6	66.5	57.1	49.3	52.3	57.4	43.5	**33.4**	59.1	45.4	41.8	46.2	54.5	+27.5
Ours		**70.8**	**60.3**	**60.5**	**65.8**	**55.2**	**55.5**	**46.7**	**49.0**	**50.1**	**40.3**	34.1	**56.1**	**42.8**	**40.2**	**43.9**	**51.4**	**+30.6**
Source	ViT-base	53.0	51.8	52.1	68.5	78.8	58.5	63.3	49.9	54.2	57.7	26.4	91.4	57.5	38.0	36.2	55.8	-
CoTTA [CVPR'22]		52.9	51.6	51.4	68.3	78.1	57.1	62.0	48.2	52.7	55.3	25.9	90.0	56.4	36.4	35.2	54.8	+1.0
VDP [AAAI'23]		52.7	51.6	50.1	58.1	70.2	56.1	58.1	42.1	46.1	45.8	**23.6**	70.4	54.9	34.5	36.1	50.0	+5.8
ViDA [2023.6.7]		47.7	42.5	42.9	**52.2**	56.9	45.5	48.9	38.9	42.7	40.7	24.3	52.8	49.1	**33.5**	33.1	43.4	+12.4
ROID [2023.6.1]		57.6	51.5	52.2	55.1	**52.4**	46.5	**47.2**	45.6	39.5	**36.0**	26.0	**45.0**	**43.8**	39.7	36.3	45.0	+10.8
Ours		**47.5**	**42.1**	**41.6**	55.5	55.4	**44.5**	47.9	**38.8**	**37.8**	39.6	**23.6**	57.0	44.4	**33.5**	**32.3**	**42.7**	**+13.1**

Table 3. Average error of standard ImageNet-to-ImageNet-C experiments over 10 diverse corruption sequences. All results are evaluated with the largest corruption severity level 5 in an online fashion. **Bold** text indicates the best performance. Blue is the suboptimal solution.

Experiments

Time	t →															
Method	*Gaussian*	*Shot*	*Impulse*	*Defocus*	*Glass*	*Motion*	*Zoom*	*Snow*	*Frost*	*Fog*	*Brightness*	*Contrast*	*Elastic_trans*	*Pixelate*	*Jpeg*	Mean
Source	28.1	26.1	36.3	12.8	35.3	14.2	12.1	17.3	17.4	15.3	8.4	12.6	23.8	19.7	27.3	20.4
SST	23.3	20.4	25.0	13.8	30.5	13.9	12.8	15.5	14.6	15.4	8.0	12.4	22.4	18.2	19.4	17.7
SST+CSK	27.5	24.8	28.9	12.0	32.8	13.6	11.2	16.9	12.8	10.2	7.9	12.2	20.5	13.8	17.5	17.4
SST+DPD	25.8	22.2	27.0	11.3	29.5	13.1	10.6	15.8	12.0	10.1	7.8	12.0	19.6	13.5	15.5	16.1
SST+CSK+DPD	21.3	17.8	22.7	13.2	26.8	13.2	11.5	14.7	13.2	10.9	8.0	10.2	18.8	14.5	16.8	15.8
SST+CSK+SPA	25.5	19.1	22.2	12.1	28.3	11.9	10.9	14.7	11.9	10.6	8.5	11.6	18.5	13.2	15.9	15.6
SST+DPD+SPA	22.1	18.1	21.2	12.6	25.1	11.9	10.5	14.3	12.2	9.8	8.1	10.6	18.3	13.9	15.6	14.8
SST+CSK+DPD+SPA	20.7	17.1	20.2	12.1	24.3	11.6	10.9	13.8	12.9	10.5	8.1	9.3	17.9	13.4	15.3	14.5

Table 5. Ablation experiments of the framework for the CIFAR10-to-CIFAR10C task. 'SST' represents the label selection with self-adaptive thresholds, and the unreliable part is discarded directly. 'CSK' is the Calibration with Source Knowledge, and 'DPD' is the Diversity with Prior Distribution module. SPA is the Soft-weighted Parameters Alignment. All results are evaluated on the ResNet.

Thanks