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

Supplementary Material

Method	Backbone	Mean	Gain
HSG	ResNet	15.6	-
HSG+PA		14.8	+0.8
HSG+SPA		14.5	+1.1
HSG	Vit-Base	11.1	-
HSG+PA		9.8	+1.3
HSG+SPA		8.9	+2.2

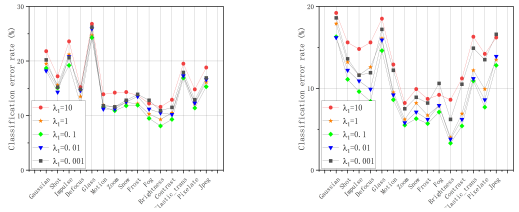
Table 1. Ablation experiments of Soft-weighted parameter alignment for the CIFAR10-to-CIFAR10C task. 'HSG' is the proposed High-quality Supervision Generator. 'PA' is parameter alignment in an average manner, and 'SPA' is the proposed Soft-weighted Parameter Alignment.

1. Ablation Studies

We verify the impact of the soft-weighted parameters alignment. Instead of calibrating the parameters using the source pre-trained model randomly, we attempt to calibrate the model using the soft-weighted parameters alignment, and the influence of the soft weights is shown in Table 1. The results demonstrate that average parameter alignment performs poorly many times compared with the soft-weighted alignment module. The latter layers in a network are much more sensitive to label noise, while their former counterparts are quite robust [1]. The weights control the similarity of the adapted model to the source one with the depth of layers, allowing noise-robust former layers to be adjusted more and noise-sensitive latter ones to be adjusted less.

We further conduct ablation experiments with the same supervision signals to prove the effectiveness of the proposed framework in Table 2. For the convenience of expression, 'SST' represents the label selection with self-adaptive thresholds, and the unreliable part is discarded directly. Then, such a module is combined with label calibration (Calibration with Source Knowledge, CSK) and diversity reweighting (Diversity with Prior Distribution, DPD), respectively. Ultimately, These three will form a versatile supervisory signal generator. SPA is the Soft-weighted Parameters Alignment module. the pseudo-label after selection and calibration strategies can effectively suppress noisy labels and improve performance. By contrast, diversity with prior distribution is vital for the model. Such modules work together to build high-quality supervision signals. Moreover, parameter alignment improves by nearly 1% in ImageNet-to-ImageNet-C, which indicates a large amount of generalization knowledge in the source model.

The results in Table 3 represent that the performance of CoTTA slightly degrades when only updating normalization parameters, mainly because CoTTA does not alleviate the effects of noisy signals, and ignores the diversity of supervision signals. Our method significantly improves such problems, which has higher computational efficiency.



(a) (b)

Figure 1. Results with different hyper-parameters in terms of classification error rate (%) for the standard CIFAR10-to-CIFAR10C. a) λ_1 with ViT-base model; b) λ_1 with ResNet model.

2. Parameters Analysis

We evaluate different hyper-parameters in terms of classification error rate (%) for the standard CIFAR10-to-CIFAR10C. We explored how the model varies with the hyper-parameter λ_1. The results shown in Figure 1 represent that our method is not sensitive to λ_1 at range $[0.01, 1]$.

References

[1] Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *NeurIPS*, 34:24392–24403, 2021. 1

Time	$t \longrightarrow$															
Method	*Gaussian*	*Shot*	*Impulse*	*Defocus*	*Glass*	*Motion*	*Zoom*	*Snow*	*Frost*	*Fog*	*Brightness*	*Contrast*	*Elastic_trans*	*Pixelate*	*Jpeg*	Mean
Source	50.1	45.5	45.4	58.3	59.6	46.6	52.9	42.6	41.3	44.3	26.9	58.6	48.2	36.6	36.8	36.8
SST	48.9	44.2	45.1	58.2	58.2	45.5	51.5	41.0	40.2	43.1	26.2	57.2	47.6	35.9	36.1	45.2
SST+CSK	48.2	43.7	43.9	57.5	57.5	45.5	51.0	40.6	39.5	42.5	28.8	57.6	47.1	35.5	35.2	44.7
SST+DPD	47.8	43.2	43.5	56.8	57.1	45.2	50.2	39.8	39.1	41.4	24.3	57.5	46.2	34.8	34.4	44.1
SST+CSK+DPD	47.5	43.4	42.8	56.1	56.5	44.9	49.2	39.8	38.4	40.9	24.5	57.5	45.8	34.2	34.1	43.7
SST+CSK+SPA	47.2	43.2	42.5	56.2	56.0	44.5	48.8	39.2	38.0	40.2	24.2	56.8	45.2	33.9	33.5	43.3
SST+DPD+SPA	47.4	42.8	42.0	56.2	55.5	45.1	48.6	39.0	38.2	39.4	23.8	57.2	44.8	33.7	32.8	43.1
SST+CSK+DPD+SPA	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

Table 2. Ablation experiments of the framework in standard ImageNet-to-ImageNet-C dataset. '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.

Time		$t \longrightarrow$															
Method	Backbone	*Gaussian*	*Shot*	*Impulse*	*Defocus*	*Glass*	*Motion*	*Zoom*	*Snow*	*Frost*	*Fog*	*Brightness*	*Contrast*	*Elastic_trans*	*Pixelate*	*Jpeg*	Mean
CoTTA	ResNet	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
CoTTA w/ BN	ResNet	24.7	23.4	28.7	12.9	31.1	14.1	11.9	17.2	16.9	15.0	8.4	12.9	22.9	19.0	22.9	18.8
Ours	ResNet	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
CoTTA	ViT	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
CoTTA w/ BN	ViT	70.3	69.4	61.5	28.9	48.2	36.6	31.9	19.1	19.6	43.8	8.5	75.0	42.9	37.7	42.0	42.4
Ours	ViT	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

Table 3. 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.