Unified Entropy Optimization for Open-Set Test-Time Adaptation

Supplementary Material

6. Pseudo Code

For a better understanding of our proposed methods, we summarize UniEnt and UniEnt+ as Algorithm 1 and Algorithm 2, respectively.

Algorithm 1: UniEnt

Input: Source model f_{θ_0} pre-trained on the source domain dataset, testing samples $\mathcal{B}_t = \{\mathbf{x}\}, t = 1, \cdots, T$.

for $t \leftarrow 1$ **to** T **do**

  for $\mathbf{x} \in \mathcal{B}_t$ **do**

    Compute csOOD score for each testing sample via Eq. (3);

  end

  Obtain $\pi(x)$ via the EM algorithm;

  Split $\mathcal{B}_t$ into $\mathcal{B}_{t,\text{csID}}$ and $\mathcal{B}_{t,\text{csOOD}}$ via Eq. (5);

  Update model via Eq. (8);

end

Output: The predictions $\arg\max_c f_{\theta_t}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{B}_t, t = 1, \cdots, T$.

Algorithm 2: UniEnt+

Input: Source model f_{θ_0} pre-trained on the source domain dataset, testing samples $\mathcal{B}_t = \{\mathbf{x}\}, t = 1, \cdots, T$.

for $t \leftarrow 1$ **to** T **do**

  for $\mathbf{x} \in \mathcal{B}_t$ **do**

    Compute csOOD score for each testing sample via Eq. (3);

  end

  Obtain $\pi(x)$ via the EM algorithm;

  Update model via Eq. (9);

end

Output: The predictions $\arg\max_c f_{\theta_t}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{B}_t, t = 1, \cdots, T$.

7. More Analysis

Scalability of large-scale datasets. To demonstrate that our methods can be used for large-scale datasets, we conduct experiments on ImageNet-C [14]. Specifically, we use ResNet-50 [13] pre-trained with AugMix [17], the weights of which can be obtained from RobustBench [7]. For optimization, we use the SGD optimizer [38] with the learning rate of 0.00025 and the batch size of 64. We apply common corruptions and perturbations to ImageNet-O [18] through the official code of [14] to construct ImageNet-O-C as csOOD data. From Table 8, we can see that UniEnt and UniEnt+ consistently improve the performance of the existing baseline methods in the open-set setting.

Method	ImageNet-C			
	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑
Source [54]	28.21	49.63	94.74	19.81
BN Adapt [33]	43.57	55.89	93.39	30.42
CoTTA [46]	47.67	55.58	94.51	33.80
TENT [44]	45.82	51.34	96.47	30.33
+ UniEnt	**47.53** (+1.71)	**56.33** (+4.99)	95.21 (-1.26)	**34.42** (+4.09)
+ UniEnt+	46.87 (+1.05)	55.86 (+4.52)	**95.10** (-1.37)	33.73 (+3.40)
EATA [35]	51.40	53.10	95.18	34.87
+ UniEnt	49.60 (-1.80)	58.29 (+5.19)	93.63 (-1.55)	36.28 (+1.41)
+ UniEnt+	**51.57** (+0.17)	**59.45** (+6.35)	**93.60** (-1.58)	**38.27** (+3.40)
OSTTA [27]	47.91	52.93	96.15	32.77
+ UniEnt	**47.92** (+0.01)	**56.02** (+3.09)	95.23 (-0.92)	**34.47** (+1.70)
+ UniEnt+	47.47 (-0.44)	55.67 (+2.74)	**95.16** (-0.99)	34.03 (+1.26)

Table 8. Results of different methods on ImageNet-C. ↑ indicates that larger values are better, and vice versa. All values are percentages. The **bold** values indicate the best results, and the underlined values indicate the second best results. The values in parentheses indicate the improvements of our methods over the baseline methods.

Scalability of model architecture. Recently, Vision Transformer (ViT) [10] has demonstrated better performance than Convolutional Neural Network (CNN), we also perform experiments with ViT backbone on ImageNet-C. Specifically, we use DeiT-Base [41] designed in [40], which proposes many techniques in the training phase to improve the robustness of the model to common corruptions. The pre-trained weights are also available from RobustBench. We update the affine parameters of the model's layer normalization. Table 9 shows that our approaches are compatible with ViT.

Performance under long-term open-set test-time adaptation. Models deployed in real-world scenarios are exposed to test samples for long periods and need to make reliable predictions at any time. Recent work [27, 37] points out that most existing TTA methods perform poorly in long-term settings, even worse than non-updating models. Following [27], we simulate long-term TTA by repeating adaptation for 10 rounds. During adaptation, the domain changes continuously and the model is never reset. The results are summarized in Table 10. We observe that in most cases the performance degradation of our methods is very slight compared to the baseline methods.

Effects of learning rate and batch size. We explore the

Method	ResNet-50				DeiT Base			
	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑
Source [54]	28.21	49.63	94.74	19.81	56.59	56.01	91.55	36.13
CoTTA [46]	47.67	55.58	94.51	33.80	60.73	53.51	93.14	37.33
TENT [44]	45.82	51.34	96.47	30.33	**62.85**	59.51	93.47	43.52
+ UniEnt	**47.53** (+1.71)	**56.33** (+4.99)	95.21 (-1.26)	**34.42** (+4.09)	58.81 (-4.04)	**67.10** (+7.59)	90.90 (-2.57)	**47.40** (+3.88)
+ UniEnt+	46.87 (+1.05)	55.86 (+4.52)	**95.10** (-1.37)	33.73 (+3.40)	58.40 (-4.45)	66.69 (+7.18)	**90.43** (-3.04)	46.74 (+3.22)
EATA [35]	51.40	53.10	95.18	34.87	**65.38**	57.95	92.92	44.29
+ UniEnt	49.60 (-1.80)	58.29 (+5.19)	93.63 (-1.55)	36.28 (+1.41)	59.36 (-6.02)	**67.22** (+9.27)	91.63 (-1.29)	48.23 (+3.94)
+ UniEnt+	**51.57** (+0.17)	**59.45** (+6.35)	**93.60** (-1.58)	**38.27** (+3.40)	61.50 (-3.88)	66.96 (+9.01)	**89.99** (-2.93)	**48.79** (+4.50)
OSTTA [27]	47.91	52.93	96.15	32.77	**60.19**	60.69	92.42	43.19
+ UniEnt	**47.92** (+0.01)	**56.02** (+3.09)	95.23 (-0.92)	**34.47** (+1.70)	58.73 (-1.46)	**67.62** (+6.93)	90.51 (-1.91)	**47.64** (+4.45)
+ UniEnt+	47.47 (-0.44)	55.67 (+2.74)	**95.16** (-0.99)	34.03 (+1.26)	58.72 (-1.47)	67.28 (+6.59)	**90.02** (-2.40)	47.32 (+4.13)

Table 9. Results of different methods on ImageNet-C using diverse architectures.

Method	CIFAR-10-C				CIFAR-100-C				Average			
	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑
Source [54]	81.73	77.89	79.45	68.44	53.25	60.55	94.98	39.87	67.49	69.22	87.22	54.16
BN Adapt [33]	84.20	80.40	76.84	72.13	57.16	72.45	84.29	47.10	70.68	76.43	80.57	59.62
CoTTA [46]	85.77	85.89	72.40	77.26	56.46	77.04	80.96	48.95	71.12	81.47	76.68	63.11
TENT [44]	79.38	65.39	95.94	56.73	54.74	65.00	94.79	42.24	67.06	65.20	95.37	49.49
+ UniEnt	**84.31** (+4.93)	92.28 (+26.89)	36.74 (-59.20)	80.32 (+23.59)	**59.07** (+4.33)	89.28 (+24.28)	51.14 (-43.65)	56.26 (+14.02)	**71.69** (+4.63)	90.78 (+25.59)	43.94 (-51.43)	68.29 (+18.81)
+ UniEnt+	84.03 (+4.65)	**93.18** (+27.79)	**32.74** (-63.20)	**80.62** (+23.89)	58.58 (+3.84)	**91.39** (+26.39)	**41.09** (-53.70)	**56.36** (+14.12)	71.31 (+4.25)	**92.29** (+27.09)	**36.92** (-58.45)	**68.49** (+19.01)
EATA [35]	80.92	84.32	71.66	72.63	**60.63**	88.64	50.18	57.24	70.78	86.48	60.92	64.94
+ UniEnt	84.31 (+3.39)	**97.15** (+12.83)	**13.25** (-58.41)	82.99 (+10.36)	59.75 (-0.88)	93.42 (+4.78)	30.36 (-19.82)	57.99 (+0.75)	72.03 (+1.26)	95.29 (+8.81)	21.81 (-39.12)	70.49 (+5.55)
+ UniEnt+	**85.18** (+4.26)	96.97 (+12.65)	14.28 (-57.38)	**83.67** (+11.04)	59.71(-0.92)	**94.23** (+5.59)	**26.87** (-23.31)	**58.19** (+0.95)	**72.45** (+1.67)	**95.60** (+9.12)	**20.58** (-40.35)	**70.93** (+6.00)
OSTTA [27]	**84.44**	72.74	77.02	65.17	**60.03**	75.37	82.75	51.35	**72.24**	74.06	79.89	58.26
+ UniEnt	82.46 (-1.98)	96.20 (+23.46)	16.37 (-60.65)	80.51 (+15.34)	58.69 (-1.34)	94.84 (+19.47)	22.95 (-59.80)	57.28 (+5.93)	70.58 (-1.66)	95.52 (+21.47)	19.66 (-60.23)	68.90 (+10.64)
+ UniEnt+	84.30 (-0.14)	**97.38** (+24.64)	**11.56** (-65.46)	**82.91** (+17.74)	58.93 (-1.10)	**95.42** (+20.05)	**20.59** (-62.16)	**57.69** (+6.34)	71.62 (-0.62)	**96.40** (+22.35)	**16.08** (-63.81)	**70.30** (+12.04)

(a) Results after 1 round of adaptation (*i.e.*, short-term test-time adaptation).

Method	CIFAR-10-C				CIFAR-100-C				Average			
	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑
Source [54]	81.73	77.89	79.45	68.44	53.25	60.55	94.98	39.87	67.49	69.22	87.22	54.16
BN Adapt [33]	84.20	80.40	76.84	72.13	57.16	72.45	84.28	47.09	70.68	76.43	80.57	59.62
CoTTA [46]	35.90	47.27	97.52	19.95	13.34	48.34	91.61	8.19	24.62	47.81	94.57	14.07
TENT [44]	32.61	60.86	93.24	20.86	37.49	53.73	95.07	25.02	35.05	57.30	94.16	22.94
+ UniEnt	84.07 (+51.46)	**88.53** (+27.67)	**51.48** (-41.76)	**77.87** (+57.01)	**57.93** (+20.44)	90.62 (+36.89)	46.18 (-48.89)	**55.67** (+30.65)	71.00 (+35.95)	**89.58** (+32.28)	**48.83** (-45.33)	**66.77** (+43.83)
+ UniEnt+	**84.17** (+51.56)	88.21 (+27.35)	52.57 (-40.67)	77.75 (+56.89)	57.92 (+20.43)	**90.63** (+36.90)	**45.10** (-49.97)	55.59 (+30.57)	**71.05** (+36.00)	89.42 (+32.13)	48.84 (-45.32)	66.67 (+43.73)
EATA [35]	40.94	64.52	88.41	29.07	48.75	73.26	80.83	41.27	44.85	68.89	84.62	35.17
+ UniEnt	**81.22** (+40.28)	91.05 (+26.53)	30.59 (-57.82)	76.42 (+47.35)	57.07 (+8.32)	**98.59** (+25.33)	**5.85** (-74.98)	56.70 (+15.43)	69.15 (+24.30)	94.82 (+25.93)	**18.22** (-66.40)	66.56 (+31.39)
+ UniEnt+	80.41 (+39.47)	**92.49** (+27.97)	**30.00** (-58.41)	**77.00** (+47.93)	**58.02** (+9.27)	98.05 (+24.79)	7.92 (-72.91)	**57.47** (+16.20)	**69.22** (+24.37)	**95.27** (+26.38)	18.96 (-65.66)	**67.24** (+32.07)
OSTTA [27]	**83.83**	71.93	76.12	63.90	57.39	75.46	82.47	49.61	**70.61**	73.70	79.30	56.76
+ UniEnt	80.74 (-3.09)	88.94 (+17.01)	35.66 (-40.46)	74.52 (+10.62)	56.13 (-1.26)	95.20 (+19.74)	21.15 (-61.32)	54.89 (+5.28)	68.44 (-2.18)	92.07 (+18.38)	28.41 (-50.89)	64.71 (+7.95)
+ UniEnt+	82.42 (-1.41)	**90.15** (+18.22)	**31.18** (-44.94)	**76.46** (+12.56)	**57.45** (+0.06)	**95.91** (+20.45)	**17.33** (-65.14)	**56.32** (+6.71)	69.94 (-0.67)	**93.03** (+19.34)	**24.26** (-55.04)	**66.39** (+9.64)

(b) Results after 10 rounds of adaptation (*i.e.*, long-term test-time adaptation).

Table 10. Results of different methods on CIFAR benchmarks.

impact of learning rate and batch size on our approaches in Table 11. A learning rate that is too large or too small can hurt performance, while a larger batch size results in better performance. Compared to TENT [44] and EATA [35], our methods are more robust to learning rate and batch size. Nonetheless, our methods share the same limitation as the baseline methods: they rely on a large batch size to estimate the distribution accurately. Moreover, we observe that OSTTA [27] is less sensitive to learning rate and batch size.

Effects of OOD score. We use the energy score [32] to measure the model's detection performance on csOOD data. From Table 12, we can make two observations. First, our methods consistently improve the performance using different OOD scores. Second, compared with MSP [15], using Max Logit [19] and Energy yields better detection performance.

Method	Learning rate				Δ
	0.005	0.001	0.0005	0.0001	
Source [54]	39.87	39.87	39.87	39.87	0.00
BN Adapt [33]	47.10	47.10	47.10	47.10	0.00
TENT [44]	10.60	42.24	42.38	48.36	37.76
+ UniEnt	53.82 (+43.22)	56.20 (+13.96)	56.06 (+13.68)	54.51 (+6.15)	2.38
+ UniEnt+	**54.44** (+43.84)	**56.36** (+14.12)	**56.27** (+13.89)	**54.65** (+6.29)	1.92
EATA [35]	40.96	57.00	56.91	53.60	16.04
+ UniEnt	**49.36** (+8.40)	57.76 (+0.76)	57.10 (+0.19)	**53.63** (+0.03)	8.40
+ UniEnt+	49.05 (+8.09)	**58.07** (+1.07)	**57.39** (+0.48)	53.40 (-0.20)	9.02
OSTTA [27]	49.43	51.35	51.98	52.37	2.94
+ UniEnt	51.41 (+1.98)	56.93 (+5.58)	57.22 (+5.24)	55.58 (+3.21)	5.81
+ UniEnt+	**53.39** (+3.96)	**57.69** (+6.34)	**57.68** (+5.70)	**56.06** (+3.69)	4.30

(a) OSCR w.r.t. learning rate

Method	Batch size				Δ
	64	32	16	8	
Source [54]	39.87	39.87	39.87	39.87	0.00
BN Adapt [33]	46.38	45.25	42.94	38.61	7.77
TENT [44]	33.27	8.10	2.51	0.95	32.32
+ UniEnt	**55.17** (+21.90)	53.05 (+44.95)	48.87 (+46.36)	**31.47** (+30.52)	23.70
+ UniEnt+	55.17 (+21.90)	**53.13** (+45.03)	**49.27** (+46.76)	28.35 (+27.40)	26.82
EATA [35]	53.09	47.78	40.57	31.57	21.52
+ UniEnt	**57.08** (+3.99)	**54.52** (+6.74)	**50.71** (+10.14)	**43.89** (+12.32)	13.19
+ UniEnt+	56.79 (+3.70)	54.29 (+6.51)	50.49 (+9.92)	43.17 (+11.60)	13.62
OSTTA [27]	50.35	48.82	46.07	39.75	10.60
+ UniEnt	54.54 (+4.19)	50.49 (+1.67)	44.97 (-1.10)	36.72 (-3.03)	17.82
+ UniEnt+	**55.76** (+5.41)	**52.66** (+3.84)	**47.94** (+1.87)	**41.45** (+1.70)	14.31

(b) OSCR w.r.t. batch size

Table 11. OSCR of different methods on CIFAR-100-C with diverse learning rates and batch sizes. Δ is the difference between the largest and smallest values. Smaller Δ values represent better robustness.

Method	OOD score			Δ
	MSP [15]	Max Logit [19]	Energy [32]	
Source [54]	39.65	40.24	39.87	0.59
BN Adapt [33]	48.75	48.04	47.10	1.65
CoTTA [46]	49.44	49.73	48.99	0.74
TENT [44]	36.86	41.79	42.24	5.38
+ UniEnt	**55.42** (+18.56)	56.20 (+14.41)	56.26 (+14.02)	0.84
+ UniEnt+	55.24 (+18.38)	**56.31** (+14.52)	**56.36** (+14.12)	1.12
EATA [35]	55.20	57.52	57.55	2.35
+ UniEnt	56.94 (+1.74)	57.88 (+0.36)	57.87 (+0.32)	0.94
+ UniEnt+	**57.37** (+2.17)	**58.33** (+0.81)	**58.33** (+0.78)	0.96
OSTTA [27]	49.14	51.42	51.35	2.28
+ UniEnt	56.52 (+7.38)	57.23 (+5.81)	57.25 (+5.90)	0.73
+ UniEnt+	**57.12** (+7.98)	**57.69** (+6.27)	**57.69** (+6.34)	0.57

Table 12. OSCR of different methods on CIFAR-100-C using diverse OOD scores.