Medical Image Segmentation with InTEnt: Integrated Entropy Weighting for Single Image Test-Time Adaptation

Haoyu Dong Nicholas Konz Hanxue Gu
Maciej A. Mazurowski
Duke University
2127 Campus Drive, Durham, NC 27708
{haoyu.dong151, nicholas.konz, hanxue.gu, maciej.mazurowski}@duke.edu

1. Implementation Details of Integrating Other Methods

The strategy of integrating InTEnt with other methods depends on the methods to be integrated with. With general optimization-based methods, *i.e.,* TEnt [4], SAR[3], FullySeg[1], we propose to first obtain a collection of models via interpolation of the statistics, and apply these methods to each of the model. For the augmentation-based methods, *i.e.,* SITA [2], we follow its strategy to apply different augmentations to the test image to obtain various test statistics of the same image. The average of all variants is used as the final test statistics. Then we utilize our procedure to interpolate between the training and averaged test statistics.

2. Detailed results

2.1. Baseline and other TTA approaches with different batch normalization layer statistics

Due to space limitation, we present all detailed results of Main Table 3 in Appendix Table 1, 2, and 3. These tables show all results for each source/target split, rather than averaging over the domain shifts for each source domain. As we can observe, in general there are minor improvements or even degradation of other TTA methods given the instability in the single image TTA setting.

2.2. Detailed results: baseline model performance with different batch norm statistics, and integration strategies

Due to space limitations, we present all detailed results of Table 6 (main paper) in Appendix Table 4, 5, and 6. These tables show all results for each domain shift (columns), rather than averaging over the domain shifts for each source domain as in Table 6. As can be observed in the top blocks of these tables, there is no universal optimal value of λ used to determine the best batch norm. statistics to adapt the model, motivating our method to integrate over multiple λ. The bottom blocks also suggest that the performance ranking of integration strategies varies between different domain shifts/train-test splits. Our method's integration strategy, "Norm", gives the overall leading performance across the three datasets.

3. Effects of hyperparameter choices

Our method's performance dependence on the choice of ensembled adapted model count (governed by C) is provided in Fig. 1 for the SC dataset, over difference domain shifts. We report the change of Dice similarity score to highlight the affect of C. We observe no significant difference over a range of values from 2 ($C = 1$) to 50 ($C \simeq 0.02$) models. Our model uses $C = 0.2$ by default, or 6 models total.

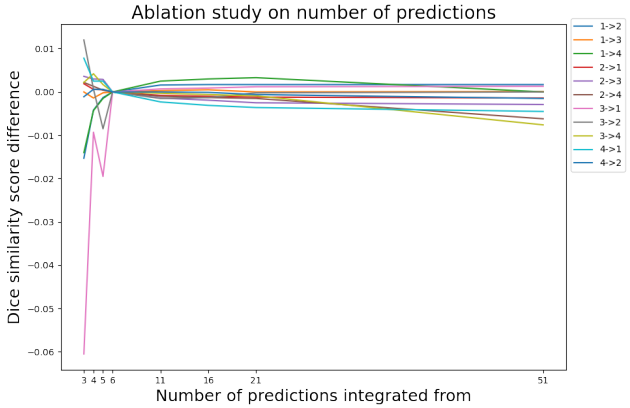

Figure 1. Our method's performance dependence on the choice of ensembled adapted model count (Eq. 5, main paper) for the SC dataset, over difference domain shifts.

References

[1] Minhao Hu, Tao Song, Yujun Gu, Xiangde Luo, Jieneng Chen, Yinan Chen, Ya Zhang, and Shaoting Zhang. Fully test-

λ	Method	1→2	1→3	1→4	2→1	2→3	2→4	3→1	3→2	3→4	4→1	4→2	4→3	Avg.↑
1.0	UNet	71.2	17.4	58.5	97.2	39.4	82.2	26.2	1.3	74.5	86.0	67.2	72.7	**57.8**
	+Tent	70.5	16.8	57.4	**97.3**	38.9	82.1	**26.3**	**1.4**	**75.3**	**87.0**	**67.9**	**72.9**	**57.8**
	+SAR	**72.1**	**17.5**	**59.9**	97.2	**40.8**	**82.5**	26.2	1.3	70.6	85.1	66.6	72.7	57.7
	+FSeg	70.5	16.9	57.4	**97.3**	39.0	82.1	25.9	**1.4**	75.1	**87.0**	**67.9**	72.7	**57.8**
	+MEMO	69.9	17.0	56.4	97.2	38.7	81.5	25.6	**1.4**	74.5	86.8	67.0	72.5	57.4
0.5	UNet	85.5	28.4	72.7	96.6	34.5	80.3	35.2	11.2	85.1	80.4	76.6	74.8	63.4
	+Tent	**85.8**	27.5	72.4	96.8	33.8	80.3	**35.7**	**11.8**	**85.6**	82.1	78.2	**75.0**	**63.8**
	+SAR	85.0	28.8	73.2	96.4	**36.5**	80.4	32.6	9.6	82.5	79.0	75.0	74.8	62.8
	+FSeg	**85.8**	27.5	72.4	96.8	33.9	80.3	**35.7**	11.7	85.5	82.1	78.3	**75.0**	63.7
	+MEMO	85.7	27.7	71.5	**96.9**	33.7	79.5	35.4	11.2	85.3	**82.4**	78.1	74.6	63.5
	+SITA	85.5	**29.8**	**74.8**	96.3	36.0	**81.8**	34.2	8.0	82.5	81.7	**78.4**	74.3	63.6
0.0	UNet	**66.3**	33.3	68.3	94.4	28.2	74.8	28.1	24.9	89.1	44.8	70.7	74.1	58.1
	+Tent	63.1	31.5	65.1	95.2	26.9	75.0	28.3	25.0	89.2	46.3	73.0	**74.2**	57.7
	+SAR	63.8	33.0	68.0	92.9	**31.5**	74.3	28.0	24.8	88.2	43.2	67.6	74.1	57.5
	+FSeg	63.3	31.4	65.1	95.2	27.0	75.0	28.2	25.1	89.3	46.3	73.0	**74.2**	57.8
	+MEMO	63.9	31.5	66.0	**95.5**	27.4	73.4	**28.7**	**25.4**	**89.4**	46.8	72.9	73.9	57.9
	+SITA	59.1	**36.8**	**72.2**	91.3	**31.5**	**78.3**	26.1	23.2	85.0	**53.1**	**76.9**	72.8	**58.9**

Table 1. The performance (segmentation Dice coeff., avg. of 10 repeated experiments) of other TTA approaches for different domain/site shifts using different batch norm. layer statistic choices (λ), **on the Spinal Cord (SC) dataset.** The best performance for a given λ and domain shift is bolded. The statistics used largely determine performance, while TTA methods themselves affect little in the performance, and may have identical values due to roundup.

λ	Method	CHN→MCU	CHN→JSRT	MCU→CHN	MCU→JSRT	JSRT→CHN	JSRT→MCU	Avg.↑
1.0	Baseline	**86.2**	**95.2**	88.2	72.6	92.1	62.8	82.9
	+Tent	**86.2**	**95.2**	**88.5**	71.7	**92.5**	61.2	82.6
	+SAR	85.5	95.0	87.5	**73.1**	91.0	**67.6**	**83.3**
	+FSeg	**86.2**	**95.2**	**88.5**	71.7	**92.5**	61.3	82.6
	+MEMO	85.0	95.1	88.1	72.6	91.7	60.0	82.1
0.5	Baseline	95.2	**96.2**	92.5	90.4	93.8	92.6	93.4
	+Tent	95.3	**96.2**	92.6	**90.7**	94.0	**92.8**	**93.6**
	+SAR	94.8	96.1	92.2	89.6	93.7	92.1	93.1
	+FSeg	95.3	**96.2**	**92.7**	**90.7**	**94.1**	**92.8**	**93.6**
	+MEMO	95.0	96.1	92.5	90.1	94.0	92.7	93.4
	+SITA	**95.6**	96.1	92.1	90.5	93.7	92.7	93.5
0.0	Baseline	**95.6**	95.9	93.0	93.7	90.7	88.9	93.0
	+Tent	94.6	**96.0**	**93.2**	**93.8**	**91.2**	89.5	93.0
	+SAR	94.3	95.8	92.8	93.5	90.4	88.3	92.5
	+FSeg	94.6	**96.0**	**93.2**	**93.8**	**91.2**	89.5	**93.1**
	+MEMO	94.6	95.8	93.1	93.5	**91.2**	89.6	93.0
	+SITA	94.7	**96.0**	92.6	93.7	90.9	**90.0**	93.0

Table 2. **Same as Table 1 but for the Chest X-ray dataset.**

time adaptation for image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021. 1

[2] Ansh Khurana, S. Paul, Piyush Rai, Soma Biswas, and Gaurav Aggarwal. Sita: Single image test-time adaptation. *ArXiv*, abs/2112.02355, 2021. 1

[3] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[4] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 1

λ	Method	CHASE→HRF	CHASE→RITE	HRF→CHASE	HRF→RITE	RITE→CHASE	RITE→HRF	Avg.↑
1.0	Baseline	52.3	40.5	61.9	53.6	55.5	**56.3**	53.3
	+Tent	52.1	39.6	61.2	53.1	54.4	**56.3**	52.8
	+SAR	**52.5**	**41.2**	**62.4**	**53.9**	**56.3**	**56.3**	**53.8**
	+FSeg	52.1	39.5	61.1	53.1	54.3	**56.3**	52.8
	+MEMO	52.1	39.5	61.1	53.0	54.3	**56.3**	52.7
0.5	Baseline	54.6	55.1	64.0	55.0	67.4	55.8	58.6
	+Tent	54.6	54.6	63.6	54.6	67.2	**55.9**	58.4
	+SAR	54.6	55.2	64.0	55.1	67.4	55.6	58.7
	+FSeg	54.5	54.5	63.6	54.6	67.2	**55.9**	58.4
	+MEMO	54.5	54.4	63.6	54.5	67.2	**55.9**	58.3
	+SITA	**54.7**	**56.0**	**64.4**	**55.9**	**67.5**	55.8	**59.0**
0.0	Baseline	54.2	60.7	**61.4**	55.2	67.6	54.4	58.9
	+Tent	**54.3**	60.4	**61.4**	54.9	**68.0**	**54.7**	58.9
	+SAR	54.2	60.7	61.2	55.3	67.3	54.2	58.8
	+FSeg	**54.3**	60.4	**61.4**	54.9	**68.0**	**54.7**	**59.0**
	+MEMO	**54.3**	60.4	**61.4**	54.9	**68.0**	**54.7**	58.9
	+SITA	54.2	**61.3**	60.3	**56.0**	67.2	54.5	58.9

Table 3. **Same as Table 1 but for the Retinal Fundus dataset.**

Method	BN stat.	1→2	1→3	1→4	2→1	2→3	2→4	3→1	3→2	3→4	4→1	4→2	4→3	Avg.↑
UNet	$\lambda=0.0$	71.2	17.4	58.5	97.2	39.4	82.2	26.2	1.3	74.5	86.0	67.2	72.7	57.8
	$\lambda=0.2$	84.1	21.6	67.6	97.3	38.1	81.7	30.9	2.9	79.2	84.8	73.2	74.0	61.3
	$\lambda=0.4$	86.6	26.5	71.8	96.9	35.7	80.9	34.1	7.4	83.3	82.6	76.0	74.7	63.0
	$\lambda=0.6$	83.2	29.8	72.8	96.3	33.2	79.7	35.1	15.7	86.7	77.4	76.8	74.9	63.5
	$\lambda=0.8$	76.4	31.3	71.4	95.6	30.6	77.8	34.0	25.9	88.8	66.8	75.3	74.8	62.4
	$\lambda=1.0$	66.3	33.3	68.3	94.4	28.2	74.8	28.1	24.9	89.1	44.8	70.7	74.1	58.1
	Integr. strat.													
InTEnt	Avg.	86.9	26.6	70.9	97.2	34.3	80.1	39.6	13.3	85.5	83.1	78.1	74.8	64.2
	Entropy	86.9	26.5	70.9	97.2	34.3	80.1	39.7	14.3	85.8	83.5	78.2	74.8	64.3
	Norm	85.8	26.4	70.0	97.4	33.2	80.0	38.8	21.1	87.7	84.4	78.4	74.9	64.8
	Min	64.6	27.7	64.3	97.5	31.5	79.2	29.4	24.4	89.1	83.0	72.1	74.2	61.4
	Top K	84.7	26.1	68.1	97.5	32.9	79.8	37.1	22.8	88.5	84.3	76.5	74.9	64.4
	Sharp.	85.1	27.4	70.8	97.2	35.0	79.5	38.5	15.8	86.1	80.0	78.0	74.9	64.0

Table 4. **Top:** Baseline UNet performance with different batch norm. statistics S^{λ}_{mix}, averaged over all domain shifts, for the **Spinal Cord (SC) Dataset**. **Bottom:** Integrated performance of the top block using different integration strategies.

Method	BN stat.	CHN→MCU	CHN→JSRT	MCU→CHN	MCU→JSRT	JSRT→CHN	JSRT→MCU	Avg.↑
UNet	$\lambda=0.0$	86.2	95.2	88.2	72.6	92.1	62.8	82.9
	$\lambda=0.2$	92.4	95.7	90.7	81.4	93.8	88.5	90.4
	$\lambda=0.4$	94.7	96.1	92.0	88.1	94.0	92.7	92.9
	$\lambda=0.6$	95.3	96.2	92.8	92.1	93.5	92.0	93.7
	$\lambda=0.8$	95.1	96.2	93.2	93.9	92.5	90.6	93.6
	$\lambda=1.0$	95.6	95.9	93.0	93.7	90.7	88.9	93.0
	Integr. strat.							
InTEnt	Average	95.1	96.2	92.7	90.8	93.8	92.8	93.6
	Entropy	95.1	96.2	92.7	91.1	93.9	93.0	93.7
	Norm	95.3	96.2	93.1	92.7	94.0	93.4	94.1
	Min	93.9	96.1	93.4	93.7	94.1	90.3	93.6
	Top K	95.1	96.2	93.3	93.6	94.1	92.9	94.2
	Sharp	95.1	96.2	92.8	91.8	93.8	92.8	93.7

Table 5. **Same as Table 4 but for Chest Dataset.**

Method	BN stat.	CHASE→HRF	CHASE→RITE	HRF→CHASE	HRF→RITE	RITE→CHASE	RITE→HRF	Avg.↑
UNet	$\lambda=0.0$	52.3	40.5	61.9	53.6	55.5	56.3	53.3
	$\lambda=0.2$	53.7	48.1	63.6	54.3	62.7	56.2	56.5
	$\lambda=0.4$	54.4	53.1	64.0	54.8	66.3	56.1	58.1
	$\lambda=0.6$	54.7	56.7	63.8	55.1	68.0	55.5	59.0
	$\lambda=0.8$	54.6	59.2	63.0	55.3	68.2	54.9	59.2
	$\lambda=1.0$	54.2	60.7	61.4	55.2	67.6	54.4	58.9
InTEnt	Integr. strat.							
	Average	54.6	54.8	64.6	55.1	67.5	56.0	58.8
	Entropy	54.6	54.8	64.6	55.1	67.5	56.0	58.8
	Norm	54.4	55.1	64.5	54.8	65.9	56.1	58.5
	Min	52.5	53.9	62.7	53.5	56.3	56.3	55.9
	Top K	53.8	55.1	63.7	54.3	62.7	56.2	57.6
	Sharp	54.7	56.7	64.4	55.0	67.3	56.1	59.0

Table 6. **Same as Table 4 but for Retinal Fundus Dataset.**