

# Supplementary Materials: From Question to Exploration: Can Classic Test-Time Adaptation Strategies Be Effectively Applied in Semantic Segmentation?

Anonymous Authors

**Table 1: Results based on DeepLabv3+, a CNN-based architecture. Compared to the counterpart that is based on Transformer (Table 1 of the main manuscript), the performance drops about 10% in average.**

Method	A-fog	A-night	A-rain	A-snow	CS-fog	CS-rain	Avg.
SO	58.9	32.0	48.1	45.7	65.9	49.8	50.1
BN adapt	36.2	28.9	37.7	36.1	60.5	55.4	47.5
TENT	59.2	31.9	48.8	46.6	64.9	53.1	50.7

## 1 TRANSFORMER-BASED ARCHITECTURES ARE PREFERRED

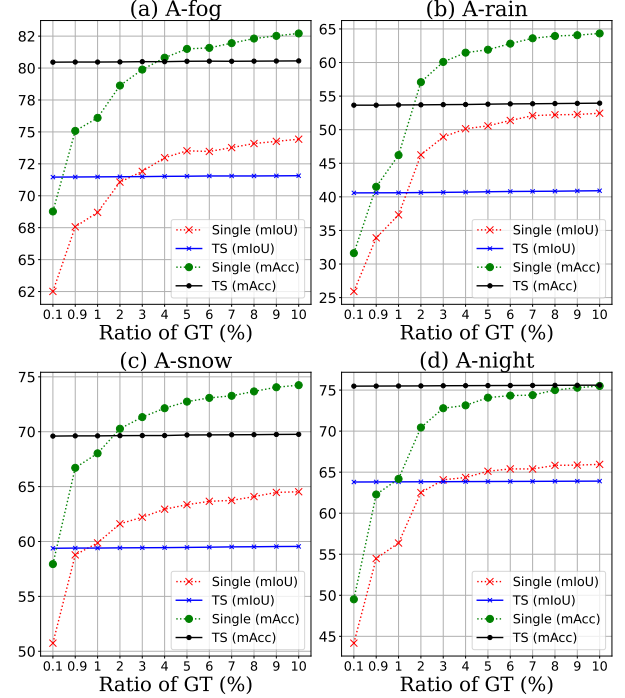
In our experiments, we deploy Segformer-B5 [12], a Transformer-based architecture, for segmentation TTA tasks. Compared to CNN-based architectures, the backbone of Transformer employs fewer BN layers. We apply DeepLabv3+ [1], a typical CNN-based architecture, on datasets ACDC [8], Cityscapes-foggy (CS-fog) [7] and Cityscapes-rainy (CS-rain) [3]. The results are depicted in Table 1, where we can observe an obvious drop compared to the results presented in Table 1 of the main manuscript. Thus, it is better to build segmentation TTA architectures based on Transformer instead of CNN. Based on the analysis of normalization updating in Section 3 of the main manuscript, it might be the attention mechanism of Transformer that contributes to its effectiveness in segmentation TTA.

## 2 MORE RESULTS REGARDING BATCH DEPENDENCY

Since online adaptation is one of the key characteristics of TTA, we also carry out experiments based on TENT [10] besides the single-model and TS scheme. The results are displayed in Table 6, further indicating that TENT is not sensitive to the temporal order of test samples. The reason might be that fewer parameters need to be updated in the deep architecture of TENT, compared to that in the single-model and TS scheme.

## 3 MORE RESULTS UNDER LONG-TAILED PHENOMENON

Although conventional wisdom may suggest that the performance of majority classes surpasses that of minority classes, we observe that this rule does not hold true in segmentation tasks. For example, in the third plot of Figure 4, class 19 attains an IoU of 0.59, whereas class 7 achieves an IoU of 0.52. However, it is worth noting that the count of class 7 is  $10^7$  while the count of class 19 is  $10^5$ , as illustrated in Figure 5. In summary, a segmentation task in TTA proves to be significantly more intricate than a classification task. The reason might be that the long-tailed (LT) phenomenon may



**Figure 1: Comparisons between the single-model and the teacher-student (TS) scheme under different degrees of ground-truth (GT) pseudo-labels (%) on ACDC. As the accuracy of pseudo-labels increases, the performance of the single-model experiences continual enhancement. However, the TS scheme’s performance remains stagnant since the strategy of test-time augmentation has not been introduced.**

cause error accumulation at pixel-level and negatively affect the training process. We provide more results on the night, rain and snow domains within the dataset ACDC, further indicating the complexity of LT problems in segmentation TTA. For instance, after adaptation, the Recall of class 7 increases from 0.27 to 0.68, while the Precision decreases from 0.78 to 0.73. An increase in Recall alongside a decrease in Precision implies a reduction in False Negative and an increase in False Positive. In summary, combining with region-level solution and introducing data augmentation might be a potential solution to address the LT phenomenon as discussed in Section 6.2 of the main manuscript.

## 4 THE EFFECT OF ATTENTION

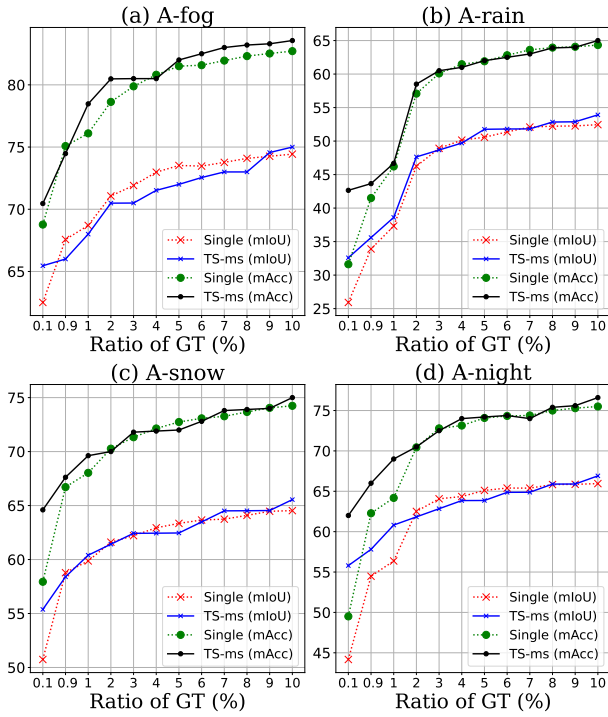
Our work demonstrates that the **attention** mechanism plays a pivotal role in a **Transformer-based model**, which is also shown in Table 3 of the main manuscript. We have shown that GN and LN do not perform well in pixel-level segmentation TTA, as displayed

**Table 2: Comparisons between the teacher-student scheme and the source-only manner on ACDC (%). “SO”/“TS” are short for source only/the teacher-student scheme, and “PL”/“Aug” are short for pseudo-labeling/test-time augmentation, respectively.**

Method	PL	Aug	A-fog	A-night	A-rain	A-snow	Avg.
SO			68.2	39.5	59.7	57.6	56.3
SO		✓	70.6 (+2.4)	40.0 (+0.5)	63.7 (+4.0)	59.2 (+1.6)	58.4 (+2.2)
TS	✓	✓	70.5 (+2.3)	39.7 (+0.2)	63.8 (+4.1)	59.2 (+1.6)	58.4 (+2.1)

**Table 3: Ablation studies on ACDC-fog of data augmentation (Aug) in terms of F1 Score and mIoU (%).**

Method	Aug	F1 Score				mIoU			
		head	mid	tail	Avg.	head	mid	tail	Avg.
Pseudo labeling	✓	89.8	82.4	82.7	85.6	82.8	71.1	69.9	74.5
		89.7 (-0.1)	82.7 (+0.3)	81.6 (-1.1)	84.7 (-0.9)	82.9 (+0.1)	73.5(+2.4)	74.3(+4.4)	76.9(+2.4)

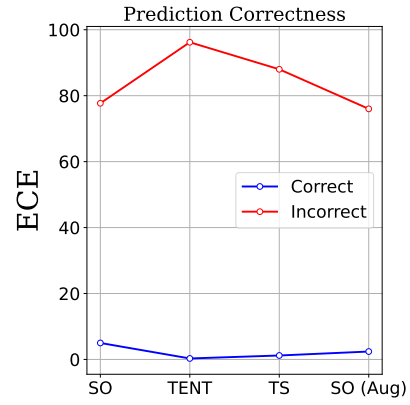


**Figure 2: Additional results based on the strategy of data augmentation in TS scheme (TS-MS). Due to this strategy, TS yields comparable results to those of the single-model.**

in Figure 1 of the main manuscript. The results demonstrate that updating Normalization layers is not very effective in segmentation TTA, while updating the attention mechanism is a promising and novel direction for transformer-based models as illustrated in that Table.

## 5 STATE-OF-THE-ART AND RECENT SEGMENTATION METHODS

We use OneFormer [4], a typical state-of-the-art and recent segmentation method, as the pre-trained model instead of SegFormer [12]. As shown in Table 4, although OneFormer shows better performance, it still deteriorates when updating BN layers. We also adopt



**Figure 3: Independently calculating the ECE for both correct and incorrect predictions (ACDC-fog, %).**

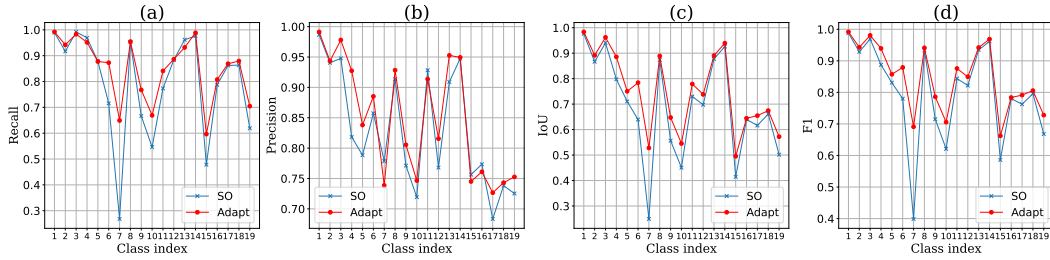
SAM [5] and find that it encounters the same problem. These results indicate that our previous analysis is reasonable and solid.

**Table 4: Results on two state-of-the-art and recent segmentation approaches, i.e., OneFormer [4] and SAM [5].**

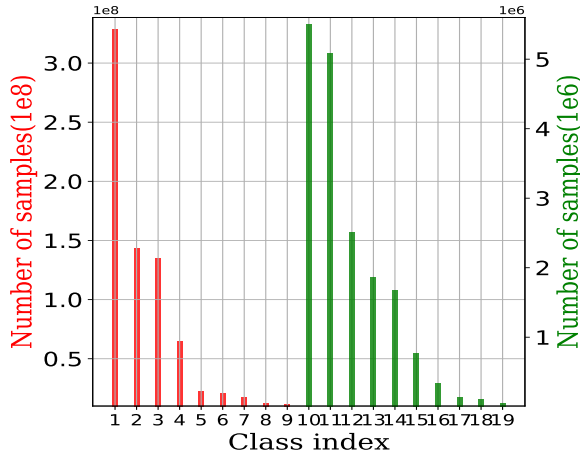
Method	A-fog	A-night	A-rain	A-snow
OneFormer + SO	70.5	48.7	62.3	61.8
OneFormer + TENT	69.1	46.5	61.2	59.8
OneFormer + SAM + SO	74.9	50.8	64.4	65.1
OneFormer + SAM + TENT	73.8	49.6	64.5	64.1

**Table 5: Comparisons between our prompt-based solution (Ours) and other methods related to prompt. It is clear that the performance of Ours is the best.**

Method	SO	DePT	DVPT	UniPT	SVDP	Ours
CS (GTA)	68.6	65.1	66.3	60.2	69.1	<b>71.1</b>
CS (Syn)	51.1	48.2	48.6	43.3	52.2	<b>56.1</b>



**Figure 4: Quantitative metrics analysis (ACDC-fog).** After adaptation, the IoU and F1 Score for the majority classes experience improvement. Specifically, there is an increase in the Recall for numerous classes, while the Precision for a limited number of classes actually witnesses a decline.



**Figure 5: The class distribution in ACDC-fog is highly imbalanced,** where the order of magnitude for **classes 1 to 9** exceeds  $10^8$  while that for **classes 10 to 19** just exceeds  $10^6$ .

## 6 MORE RESULTS OF MODEL CALIBRATION: REFLECTING THE COMPLEXITY OF SEGMENTATION TTA

In the real world, a decision-making system, such as an autonomous car, should not only improve the decision accuracy but also understand when they are potentially unreliable [2, 9]. Attaining an optimal solution in practice proves elusive. Thus, we conduct experiments to delve into model interpretability, aiming to unearth the primary challenges associated with the uncertainty of segmentation TTA where there lacks a comprehensive study on model calibration.

Miscalibration arises from a misalignment between predictive confidence and accuracy, as defined by the expected calibration error (ECE) formalism, i.e.,  $ECE = \sum_{i=1}^m \frac{|B_i|}{N} |\text{acc}(B_i) - \text{conf}(B_i)|$ , where  $m$  is the number of bins,  $B_i$  denotes a set of samples falling into the bin, and  $\text{acc}(B_i)$  and  $\text{conf}(B_i)$  are actual accuracy and confidence averaged over the samples in the bin, respectively. As displayed in Figure 3, the ECE arising from incorrect predictions markedly outweighs that from correct predictions for both methods. This disparity underscores the predominant role of mispredictions in leading to miscalibration, and it also reinforces the argument that over-confidence remains a paramount concern in segmentation TTA [9].

**Table 6: Comparisons under different temporal orders of images from datasets ACDC, Cityscapes-fog and Cityscapes-rain (% TEnt).** Different random seeds (i.e., 0/9/99/999/9999) represent different time orders. For each row of the table, the results under different random seeds are relatively stable, representing that this approach is not sensitive to the order of test samples.

Domain	0	9	99	999	9999
A-fog	65.8	65.6	65.6	65.6	65.5
A-night	40.5	41.0	41.1	40.9	41.0
A-rain	62.0	62.2	62.3	62.2	62.0
A-snow	57.8	57.9	57.7	57.9	57.8
CS-fog	73.8	73.8	73.7	73.8	73.8
CS-rain	66.8	66.8	66.8	66.7	66.8

## 7 VISUALIZATION OF SEGMENTATION TTA RESULTS

In this Section, we will visualize the results of different segmentation TTA approaches applied on the dataset ACDC. Some of the results are displayed in Figure 10, where it is clear that TEnt [10] is hard to differentiate between the road and the sky (marked in black boxes). Moreover, thanks to the TS scheme and the data augmentation strategy, CoTTA [11] produces a more refined segmentation map (shown in white boxes).

The presence of noisy pseudo-labels tends to aggravate error accumulation and catastrophic forgetting in TTA [6, 11, 13]. However, we find the experimental results of CoTTA [11] and “SO + aug” are extremely similar, confusing the actual impact of error accumulation and catastrophic forgetting on segmentation TTA. To elucidate this, we conduct a more refined visual analysis, focusing on two strategies proposed by CoTTA [11], i.e., weight-averaged and stochastic restore. As depicted in Figure 11, we can find that these strategies can not guarantee results improvement. For example, in ACDC-fog (shown in the white box), “TS” correctly identifies pixels labeled as sidewalk, although accompanied by numerous misclassifications (the upper part in the box). Utilizing the weight-averaged strategy eliminates these misclassifications, but compromises sidewalk predictions. The subsequent application of the stochastic restore strategy yields prediction in more complex sidewalk areas (the left area in the box), but reintroduces prior noise. A similar pattern is discernible across the remaining domains. In summary, these strategies are not thoroughly effective in genuinely

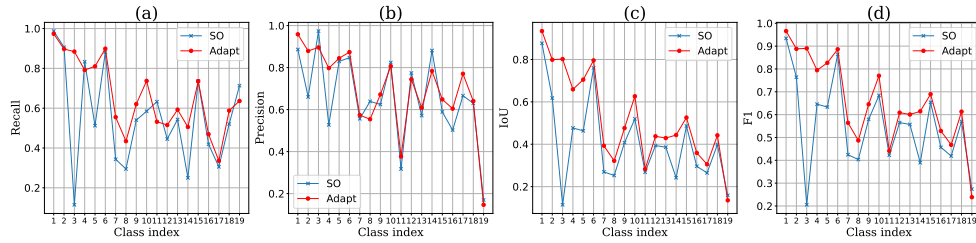


Figure 6: Quantitative metrics analysis on ACDC-night.

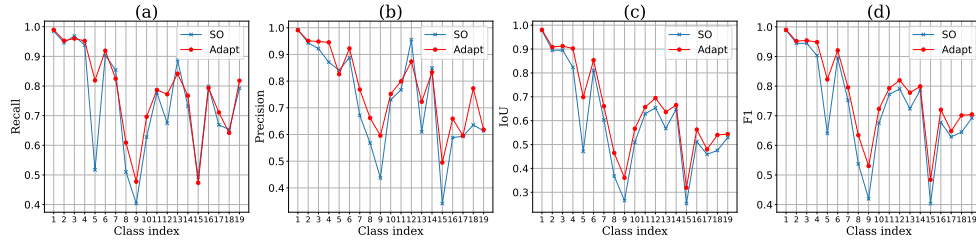


Figure 7: Quantitative metrics analysis on ACDC-rain.

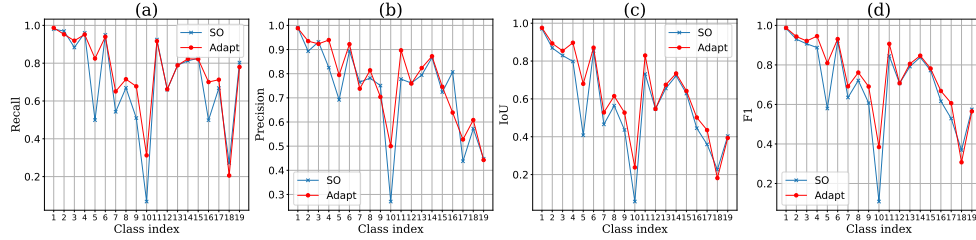


Figure 8: Quantitative metrics analysis on ACDC-snow.

road	0.41	0.0039	5.5e-06	0.00047	0.00063	3.9e-05	1.3e-06	2e-05	4.5e-05	0.0012	0	8.8e-06	1.2e-06	0.00054	9.6e-05	4.6e-05	1.7e-05	6.3e-07	8.7e-06
sidewalk	0.0061	0.056	0.00018	0.00066	0.00014	8.5e-05	1e-07	4.1e-06	9.7e-05	0.00064	0	4.3e-05	2e-06	8.6e-05	2e-06	1.1e-06	4.8e-06	3e-06	1.5e-05
building	1.1e-05	0.00019	0.019	0.0004	0.00024	0.00049	6.7e-05	0.00018	0.0004	0.0001	0.00043	5.2e-05	9e-06	0.00037	0.00017	1.9e-05	0.00023	2e-06	2.6e-05
wall	0.00024	0.0009	0.00068	0.0027	0.003	8.8e-05	1.7e-07	1.7e-06	0.0021	0.00057	1.4e-05	2.7e-05	2.3e-06	6.9e-05	3.1e-05	1.1e-07	0	1.3e-06	1.4e-05
fence	0.0003	0.00059	0.0038	0.0025	0.036	0.0004	4e-06	4.7e-05	0.0065	0.0032	0.00073	2.4e-05	3.1e-07	0.00021	5e-05	7.8e-07	8.4e-06	2.2e-06	1.2e-05
pole	8.8e-05	0.00018	0.0014	6.9e-05	0.00068	0.011	0.00019	0.00031	0.0017	0.00023	0.0011	1.9e-05	3.3e-06	0.00011	5.9e-05	8.9e-06	2.3e-05	7.2e-06	4.7e-06
traffic light	7.1e-08	0	0.00015	0	7.4e-06	0.00011	0.0014	0.00057	0.00011	1.9e-07	5.6e-05	7.7e-08	0	5.2e-06	6.6e-06	0	2.8e-06	0	0
traffic sign	3.2e-05	2.7e-06	0.0002	6e-06	4.7e-05	9.1e-05	0.00012	0.00068	0.00023	5.1e-06	8.8e-05	7.9e-06	2.5e-07	4.3e-05	9.2e-06	7.7e-06	7.3e-06	0	1.9e-06
vegetation	0.00018	0.00045	0.00071	0.00018	0.00041	0.00098	9.5e-05	0.00016	0.01	0.0009	0.00021	2.9e-05	1.1e-05	0.00038	0.00011	3.7e-05	2e-05	5e-06	2.8e-06
terrain	0.0013	0.0013	7.6e-05	0.00045	0.0013	0.00013	8.8e-06	1.8e-05	0.0038	0.06	4.7e-05	3.2e-06	1e-06	6.9e-05	5.6e-06	9.9e-07	0	1.6e-06	1.3e-06
sky	1.2e-05	0	0.00082	2.3e-05	0.00047	0.001	5.8e-05	0.0001	0.0053	3.8e-05	1	8.3e-08	0	2.9e-05	8.2e-05	4e-06	2.5e-05	0	0
person	8e-07	1e-05	3.4e-05	6.5e-06	2.9e-05	1.1e-05	4.8e-07	6.7e-08	3.1e-05	7.2e-07	3.1e-08	0.00083	3.3e-05	3e-05	4.8e-07	1.9e-06	6.1e-07	2.1e-06	1.4e-05
rider	3.3e-07	4e-07	4.9e-07	3.1e-06	1.8e-08	2.1e-08	7.7e-08	0	3.7e-06	0	0	1.2e-06	0.00029	5.5e-06	0	6.7e-07	4.1e-06	1.9e-05	0
car	0.00039	2.8e-05	0.0018	6.3e-05	0.00019	3.1e-05	1.6e-05	7.9e-06	0.00017	3.8e-05	8.3e-06	3.2e-05	9.1e-06	0.036	0.00065	7.8e-06	6.8e-06	2.1e-06	9.1e-06
truck	0.0019	1e-05	0.00024	0.00018	0.00038	6.7e-05	3.3e-05	0.00018	0.00022	3.9e-05	4e-05	5.6e-06	5.7e-07	0.0002	0.00011	1.9e-05	1.5e-06	1.5e-06	1.2e-06
bus	1.8e-05	1.6e-06	9.1e-06	2.4e-07	0	1.9e-06	5.8e-07	2.2e-07	3.8e-06	3.9e-07	1.6e-06	4.8e-07	1.2e-06	1.5e-05	6.8e-06	0.00011	2.1e-06	0	0
train	3.1e-05	4.3e-06	5.5e-05	3.3e-06	9.9e-05	2e-05	1.1e-05	3.1e-06	3.9e-05	1.7e-07	9.5e-06	3.4e-06	0	1.9e-05	6.2e-07	1.3e-05	0.00054	0	0
motorcycle	3.4e-07	3.7e-07	3.7e-06	4e-08	4.9e-08	5e-07	0	3.9e-06	3.7e-08	0	3e-06	3.9e-06	1.5e-05	0	0	0	0.00011	9.7e-06	0
bicycle	2.1e-06	4e-06	2.3e-06	5.7e-06	3.1e-08	1.8e-06	3.7e-08	0	8.3e-07	1.3e-07	0	5e-06	2.7e-05	4.5e-06	0	0	4.6e-07	0	0.00036
	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle

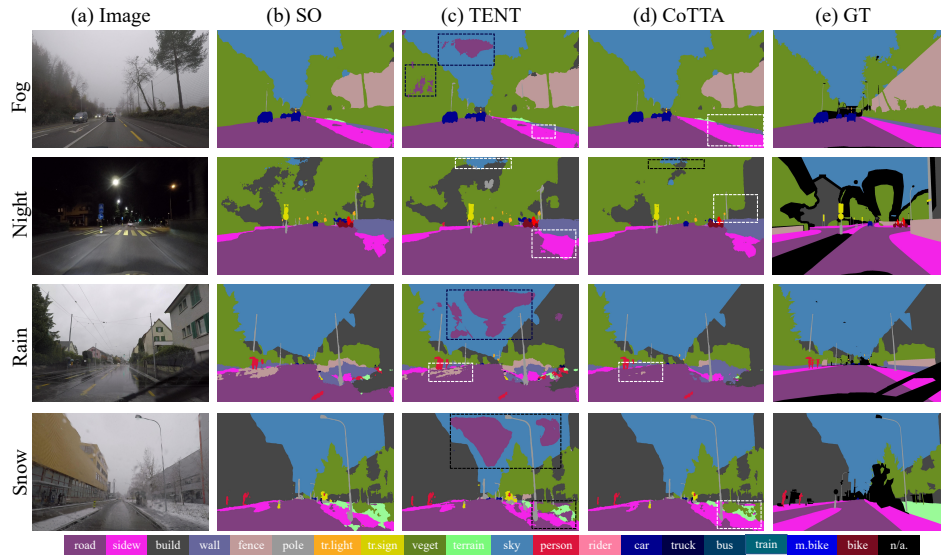
Figure 9: Confusion matrix of ACDC-fog. Here, x-axis indicates the predicted labels, while y-axis represents the ground-truth labels. Moreover, the data has been normalized to Min-Max Normalization. We can observe a substantial disparity in performance between the majority and minority classes, underscoring the challenges inherent in segmentation TTA.

resolving the issues of error accumulation and catastrophic forgetting. Thus, further improvement of segmentation TTA approaches is necessary.

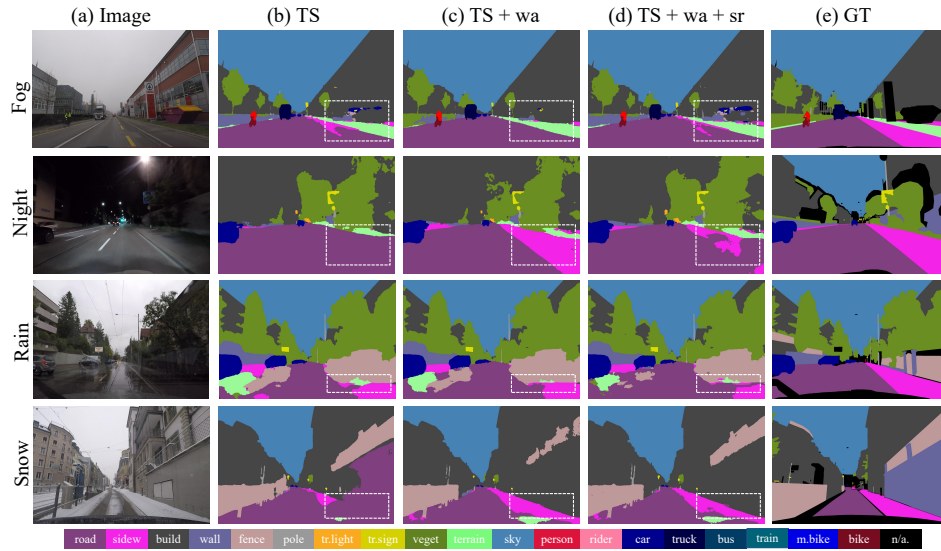
## REFERENCES

- [1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*. 801–818.
- [2] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*. 1321–1330.
- [3] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. 2019. Depth-attentional features for single-image rain removal. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 8022–8031.
- [4] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. 2023. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2989–2998.
- [5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.
- [6] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. 2022. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*. 16888–16905.
- [7] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2018. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision* 126 (2018), 973–992.





**Figure 10: Qualitative comparisons of segmentation results on dataset ACDC. Compared to SO (Source Only), the black box indicates inferior results while the white box signifies improved outcomes.**



**Figure 11: Segmentation results of different strategies in CoTTA [11] applied on dataset ACDC. “TS”/“wa”/“sr” are short for teacher-student scheme/ weight-averaged strategy/stochastic restore, respectively.**

- [8] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2021. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10765–10775.
- [9] Dongdong Wang, Boqing Gong, and Liqiang Wang. 2023. On Calibrating Semantic Segmentation Models: Analyses and an Algorithm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23652–23662.
- [10] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. Tent: Fully test-time adaptation by entropy minimization. In *International Conference On Learning Representations*. 1–12.
- [11] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. 2022. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7201–7211.
- [12] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems* 34 (2021), 12077–12090.
- [13] Longhui Yuan, Binhui Xie, and Shuang Li. 2023. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15922–15932.