Overcoming Domain Knowledge Forgetting in Continual Test-Time Adaptation via Siamese Networks
Supplemental Material

1. More Experimental Details

1.1. The Experimental Environment

We perform all experiments on PyTorch (2.2.1+cu118), using 8 × NVIDIA GeForce RTX 3090. The code is primarily implemented based on (Wang et al., 2022; Döbler et al., 2023; Marsden et al., 2024). We express our gratitude for their contributions to the open-source community.

1.2. Specific Adaptation Sequences on DomainNet126

DomainNet126 consists of four domains. We obtain a pretrained model for one domain at a time and then sequentially perform continual test-time adaptation on the remaining three domains. The specific adaptation sequences are shown in Tab. 1.

Table 1: The Adaption sequences on DomainNet126.

Source Domain	**Target Domain Sequence**		
real	clipart	painting	sketch
clipart	sketch	real	painting
painting	real	sketch	clipart
sketch	painting	clipart	real

2. Supplementary Experiments

2.1. Triple-Network Structure Ablation Study

This ablation study aims to validate the necessity of the bridge network in transferring and integrating knowledge across domains. We set up three groups: (1) using only "Tent-A"(Group A), (2) using a dual-network architecture of "Tent-A + Momentum Network"(Group B), and (3) using a triple-network architecture of "Tent-A + Bridge Network + Momentum Network" (i.e., the complete ATAN)(Group C). The experimental results on ImageNet-C are shown in Fig. 1.

Based on the experimental results, using the dual network architecture of "Tent-A + Momentum Network" resulted in even worse error rate. In contrast, employing ATAN significantly improved model performance. This demonstrates the necessity of the bridge network.

Figure 1: Error rates on ImageNet-C using different numbers of Siamese networks. Group A: using only "Tent-A"; Group B: using "Tent-A + Momentum Network"; Group C: using "Tent-A + Bridge Network + Momentum Network".

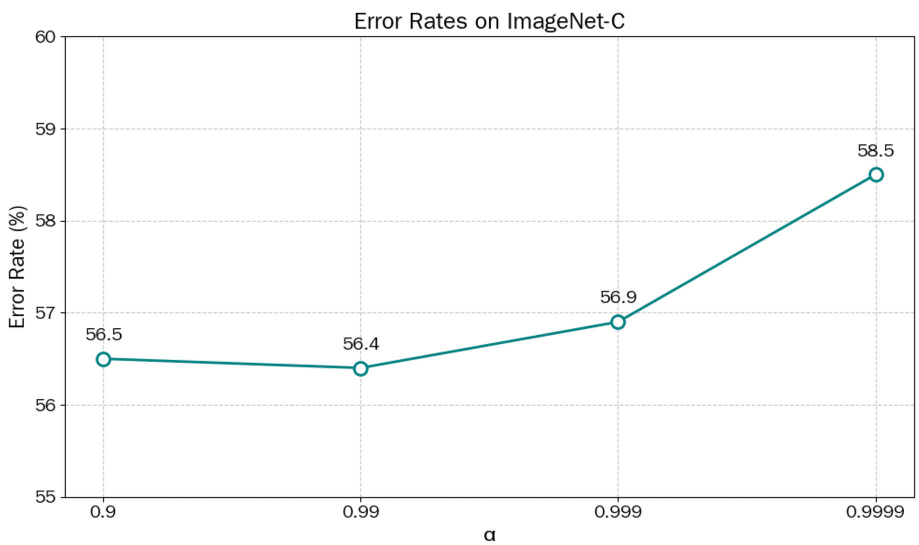

Figure 2: Error rates for ATAN (with Tent-A) using different momentum coefficients.

2.2. Sensitivity Analysis of Momentum Coefficient

The momentum coefficient α is a critical hyperparameter in the momentum network during exponential moving average updates. A larger value of α results in smaller parameter update increments for the momentum network. Figure 2 shows the variation in error rates of ATAN across different values of α.

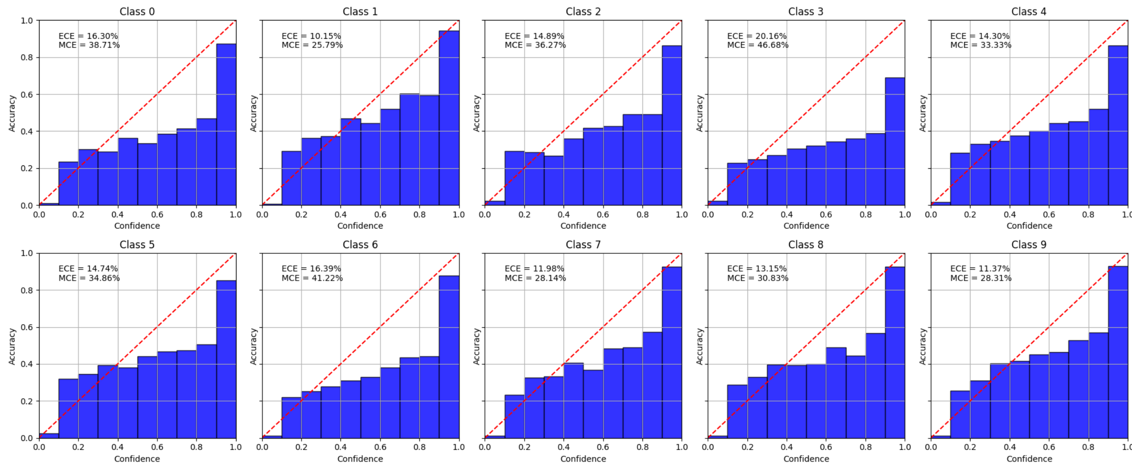

Figure 3: The Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) of Tent-A on CIFAR10-C. The closer the arrangement of the bar graph is to the diagonal line (red dashed line), the better. Lower values of ECE and MCE indicate better performance.

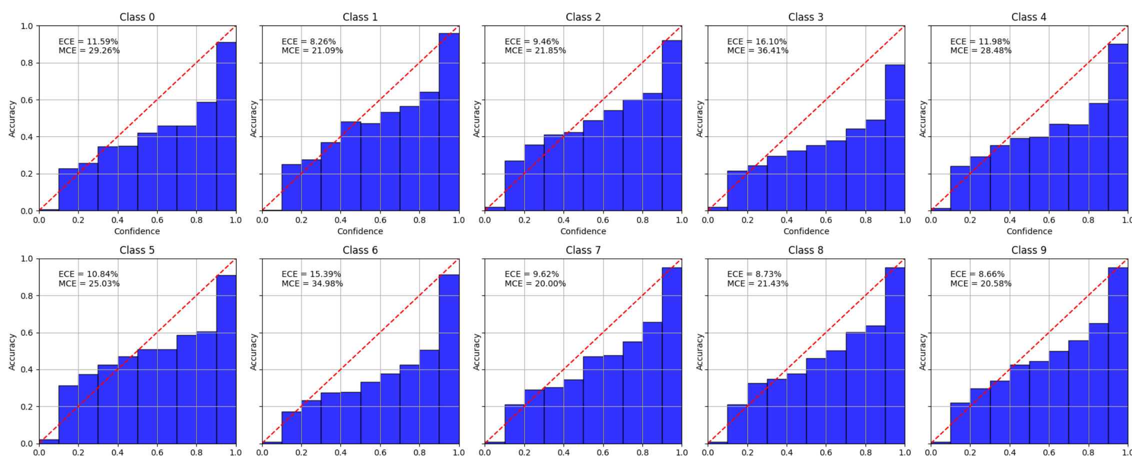

Figure 4: The Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) of ATAN (With Tent-A as the forerunner network) on CIFAR10-C. ATAN significantly reduces both the expected and the maximum calibration error.

It can be seen that when the momentum coefficient ranges between 0.9 and 0.999, the error rate fluctuates minimally; however, when the momentum coefficient reaches 0.999, the error rate increases significantly.

2.3. Model Calibration

Model calibration Guo et al. (2017) refers to the degree of alignment between the probabilities predicted by a model and the actual frequency of event occurrences. Good calibration aids in determining the extent to which one should trust a model's predictions, thereby enabling more reliable decision-making in practical applications. Conversely, poor calibration significantly diminishes a model's usability in real-world scenarios. Entropy minimization-based methods encourage models to produce overconfident predictions, thus reducing model calibration.

Calibration error measures the deviation between the predicted probability (confidence) and the actual accuracy. Expected Calibration Error (ECE) divides the predicted probabilities into several intervals (bins), calculates the difference between the average predicted probability and the actual accuracy within each bin, and then computes a weighted average of these differences.

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right| \tag{1}$$

where M is the number of bins, B_m is the m-th bin. $|B_m|$ is the number of samples in the m-th bin, n is the total number of samples, $\mathrm{acc}(B_m)$ is the actual accuracy in the m-th bin, and $\mathrm{conf}(B_m)$ is the average predicted probability in the m-th bin. Maximum Calibration Error (MCE) focuses on the maximum calibration error among all bins.

$$\mathrm{MCE} = \max_{m=1}^{M} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right| \tag{2}$$

where the symbols have the same meanings as in the ECE formula.

We set the number of the bins to 10. As shown in Fig. 3, Tent-A exhibits substantial calibration error. However, with Tent-A as the forerunner network, the momentum network in ATAN effectively reduces calibration error (Fig. 4), to some extent rectifying model calibration and enhancing the model's credibility.

3. More Related Work

Domain generalization (DG) addresses the challenge of out-of-distribution (OOD) generalization, where models trained on source domains must perform well on unseen target domains without access to target data (Blanchard et al., 2011). Recent advances propose diverse strategies, including domain alignment for invariant feature learning (Li et al., 2018b,c), meta-learning to simulate domain shift during training (Li et al., 2018a; Balaji et al., 2018), and data augmentation techniques to enrich training diversity (Zhou et al., 2020a,b). Ensemble methods (Zhou et al., 2021), self-supervised learning (Carlucci et al., 2019), and disentangled representations (Li et al., 2017) further expand the methodological landscape. DG has been applied to vision, speech, NLP, and medical imaging, highlighting its broad impact across machine learning domains (**?**Liu et al., 2020).

Continual learning focuses on catastrophic forgetting (Chen and Liu, 2018), closely related to CTTA. Existing approaches are commonly categorized into regularization-based methods, replay strategies, and dynamic architectural techniques (Wang et al., 2024). Regularization constrains parameter updates to preserve past knowledge (Kirkpatrick et al., 2017). Replay methods leverage memory or generative models to rehearse old experiences (Rebuffi et al., 2017). Architectural approaches allocate or expand parameters for new tasks (Rusu et al., 2016). Many CTTA works (Wang et al., 2022; Döbler et al., 2023) draw upon the concept of continual learning.

References

Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31, 2018.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011.

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2229–2238, 2019.

Zhiyuan Chen and Bing Liu. *Lifelong machine learning*. Morgan & Claypool Publishers, 2018.

Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7704–7714, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018a.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018b.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639, 2018c.

Quande Liu, Qi Dou, and Pheng-Ann Heng. Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains. In *International conference on medical image computing and computer-assisted intervention*, pages 475–485. Springer, 2020.

Robert A Marsden, Mario Döbler, and Bin Yang. Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2555–2565, 2024.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383, 2024.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, pages 561–578. Springer, 2020a.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13025–13032, 2020b.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018, 2021.