Supplementary Materials. Continual-MAE: Adaptive Distribution Masked Autoencoders for Continual Test-Time Adaptation

Jiaming Liu[1,2], Ran Xu[1,2]*, Senqiao Yang[1†], Renrui Zhang [3‡], Qizhe Zhang[1],
Zehui Chen [4], Yandong Guo[2], Shanghang Zhang[1] ✉
[1]National Key Laboratory for Multimedia Information Processing, School of Computer Science,
Peking University [2]AI[2]Robotics [3] MMLab, CUHK [4]University of Science and Technology of China
jiamingliu@stu.pku.edu.cn, xu_ran@bupt.edu.cn, shanghang@pku.edu.cn

The supplementary materials presented in this paper offer a comprehensive quantitative and qualitative analysis of the proposed method. In Appendix A.1, we present the fine-grained experiment results of various methods on the ImageNet-to-ImageNet-C CTTA task scenario. The additional ablation study is provided in Appendix A.2, encompassing experiments aimed at validating the effectiveness of individual components on the ImageNet-to-ImageNet-C task. Meanwhile, an exploratory experiment is conducted to assess the sensitivity of the masking ratio in our proposed method. We present additional qualitative visualization comparisons on the Cityscapes-to-ACDC CTTA task in Appendix B. In Appedix C, we furnish supplementary empirical observations and justifications supporting our motivation, including detailed computational procedures and quantitative analyses.

A. Additional Quantitative Analysis

A.1. Fine-Grained Performance

In this section, we provide a fine-grained performance of the classification results on the ImageNet-to-ImageNet-C task presented in our submissions. As shown in Table 1, our Adaptive Distribution Masked Autoencoders (ADMA) archive the lowest classification error rate 42.5%, and at the fine-grained level, demonstrate outstanding performance across 13 out of the 15 corruption types, validating the robustness of our method in the continual adaptation process.

A.2. Additional Ablation Study

Components Effectiveness on ImageNet-to-ImageNet-C. We conduct an additional experiment to evaluate each component of our proposed method on the ImageNet-to-ImageNet-C CTTA task. Consistent with our submission, we perform four sets of ablation studies. As shown in Table 2, the first set of experiments (Ex1) is applying random masking strategy to establish consistency constraints between the model outputs generated from the masked target samples and those from the original target samples. This obtains a 6.4% reduction in the error rate in contrast to the source method (Ex0). Secondly, with the implementation of the Distribution-aware Masking (DaM) mechanism, the error rate (Ex2) is further reduced to 47.9%, validating that DaM significantly enhances the model's ability to understand the target domain distribution. The remaining two sets of experiments (Ex3 and Ex4) are reconstructing the Histogram of Oriented Gradients (HOG) feature representations based on two masking strategies. Random masking strategy with HOG reconstruction scheme (Ex3) achieves a 1.7% reduction in the error rate compared to use random masking strategy individually (Ex1). Our method (Ex4) outperforms others, showcasing the best results, with a remarkable 12.2% reduction in error rate compared to the source method. These results confirm that incorporating HOG reconstruction into the continual adaptation process aids the model in acquiring task-relevant knowledge, particularly in the presence of domain shifts.

Masking Ratio Sensitivity. We conduct another set of ablation experiments to investigate the sensitivity of our DaM mechanism to the masking ratio. Given that the optimal results are obtained when DaM is coupled with the reconstructed HOG scheme, we specifically conduct these experiments directly on the CIFAR10-to-CIFAR10C CTTA task using our integrated method (DaM&HOG). The effect of the masking ratio is shown in Figure 1, a wide range of mask ratios from 30% to 80% produce different performances. Clearly, the optimal result showcased in the submission is achieved when the masking ratio is set at 50%. Remarkably, within the vicinity of a 50% masking ratio, the error rate exhibits minimal fluctuations, with the highest recorded error rate being 14.0% at a 70% masking ratio. However, setting the masking ratio to the extremes (i.e., 30% and 80%), yields

*Equal contribution, † Technical contribution, ‡ Project leader, ✉ Corresponding author. Web: https://sites.google.com/view/continual-mae/home

Method	REF	Gaussian	shot	impulse	defocus	glass	motion	zoom	snow	frost	fog	brightness	contrast	elastic_trans	pixelate	jpeg	Mean↓	Gain
Source [3]	ICLR2021	53.0	51.8	52.1	68.5	78.8	58.5	63.3	49.9	54.2	57.7	26.4	91.4	57.5	38.0	36.2	55.8	0.0
Pseudo-label [7]	ICML2013	45.2	40.4	41.6	51.3	**53.9**	45.6	47.7	40.4	45.7	93.8	98.5	99.9	99.9	98.9	99.6	61.2	-5.4
TENT-continual [10]	ICLR2021	52.2	48.9	49.2	65.8	73	54.5	58.4	44.0	47.7	50.3	23.9	72.8	55.7	34.4	33.9	51.0	+4.8
CoTTA [11]	CVPR2022	52.9	51.6	51.4	68.3	78.1	57.1	62.0	48.2	52.7	55.3	25.9	90.0	56.4	36.4	35.2	54.8	+1.0
VDP [4]	AAAI2023	52.7	51.6	50.1	58.1	70.2	56.1	58.1	42.1	46.1	**45.8**	23.6	70.4	54.9	34.5	36.1	50.0	+5.8
Ours	**Proposed**	**46.3**	**41.9**	**42.5**	**51.4**	54.9	**43.3**	**40.7**	**34.2**	**35.8**	64.3	**23.4**	**60.3**	**37.5**	**29.2**	**31.4**	**42.5**	**+13.3**

Table 1. A fine-grained Classification error rate(%) for standard ImageNet-to-ImageNet-C online CTTA task. Mean(%) denotes the average error rate across 15 target domains. Gain(%) represents the percentage of improvement in model accuracy compared with the source method.

	Random	DaM	HOG	Mean↓	Gain
Ex0	-	-	-	55.8	/
Ex1	✓	-	-	49.4	+6.4
Ex2	-	✓	-	47.9	+7.9
Ex3	✓	-	✓	47.7	+8.1
Ex4	-	✓	✓	**43.6**	**+12.2**

Table 2. Average error rate(%) for ImageNet-to-ImageNet-C online CTTA task. Random, DaM, and HOG represent the random masking strategy, our proposed Distribution-Aware Masking mechanism, and our introduced HOG reconstruction, respectively.

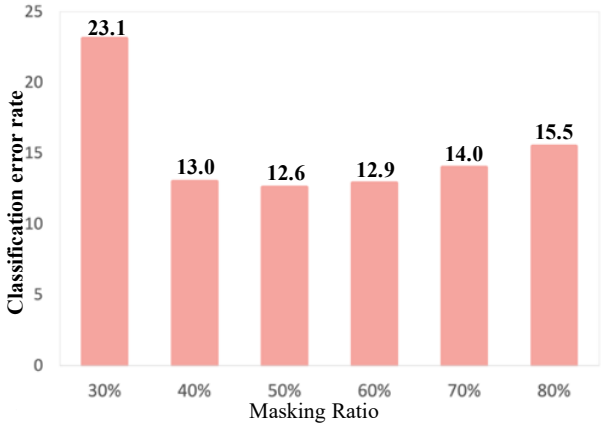

Figure 1. Average error rate(%) for CIFAR10-to-CIFAR10C CTTA task when applying different masking ratio(%).

intriguing results. The large masking ratio of 80% results in a 2.9% deterioration in the error rate compared to the best result, while the small masking ratio of 30% leads to a more pronounced degradation of 10.5%. Hence, the results suggest that DAM is not highly sensitive to the masking ratio when it exceeds 30%, consistently yielding a relatively robust adaptation process. At the same time, when the masking ratio is too large, such as 70% or 80%, the classification error rate also experiences some increase. The reason is that the mask covers a large portion of information, resulting in insufficient semantic information expression. Finally, we chose a masking ratio of 50% for our classification and segmentation CTTA experiments.

B. Additional Qualitative Analysis

To intuitively assess the effectiveness of our approach, we conducted an additional set of qualitative experiments. Specifically, we performed four sets of comparative experiments in the Cityscapes-to-ACDC CTTA scenario. In the first set of experiments, we tested the Segformer-B5 model [12], pre-trained on the source domain Cityscapes dataset, directly on the four shifted domains of the ACDC dataset. Next, we adapted the model, initially pre-trained on the source domain, to the target domains using the leading CTTA methods TENT [10] and CoTTA [11] from recent years. The final set of experiments involved applying our proposed method for continual adaptation to the four target domains. The results of the visualization of the segmentation maps for all the methods are shown in Figure 2. The model applying our method has the best segmentation results for all the target domains compared to the original source model, the model applying the TENT method and the model applying the CoTTA method. Notably. Our method archives consistent improvements for most categories and the benefits in categories like $sidewalk$, $terrain$, and $traffic\ sign$ are very significant (shown in white box). Adapting to these challenging categories is inherently difficult. Therefore, our DaM mechanism strategically masks samples from these categories that are more susceptible to domain shifts during the testing process. Simultaneously, our model continues to process inputs from the original images. By leveraging this contextual knowledge for consistency constraints, coupled with the HOG reconstruction scheme that enhances task-relevant feature representation, the model achieves more accurate segmentation of intricate regions.

C. Additional Discussion and Justification

In this section, our goal is to furnish detailed implementation insights that substantiate our intuition. In Section C.1, we

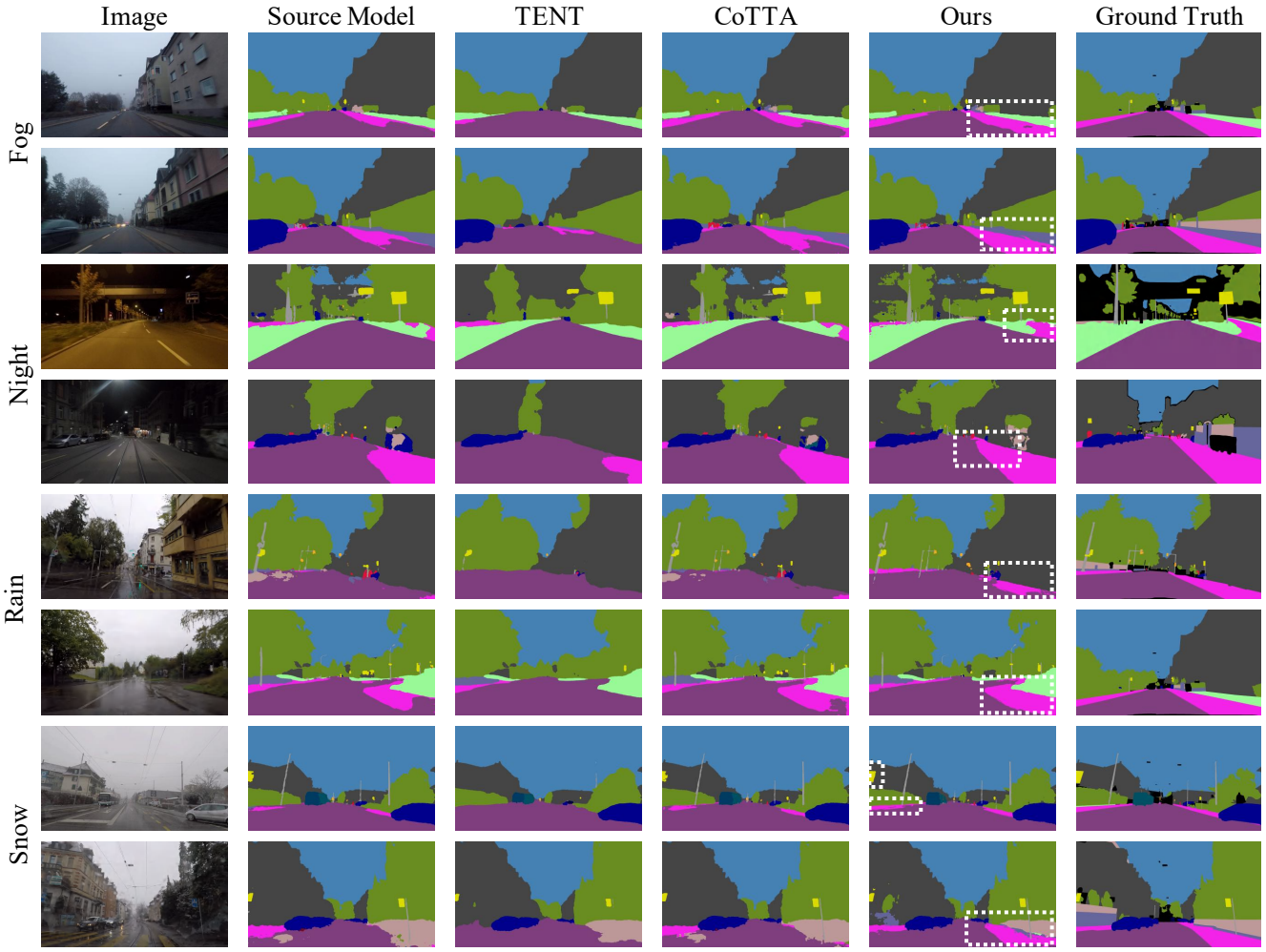

Figure 2. Qualitative comparison of our method with previous SOTA methods on the ACDC dataset. Our method could better segment different pixel-wise classes such as shown in the white box.

elucidate our choice of utilizing the Jensen–Shannon (JS) divergence for computing inter-domain divergence. Additionally, we illustrate the trends of inter-domain divergence in segmentation tasks. We extend the visualization of Class Activation Mapping (CAM) in Section C.2.

C.1. Inter-Domain Divergence.

To substantiate the rationale behind our proposed DaM and HOG reconstruction mechanism, we measure the distribution distances of feature representations across various target domains. Inspired by previous works [5], we utilize the domain distance definition proposed by Ben-David [2] and employ the $\mathcal{H}$-divergence metric to assess the domain representations of our proposed method. The $\mathcal{H}$-divergence between D_S and D_{T_i} can be calculated as:

$$d_{\mathcal{H}}(D_S, D_{T_i}) = 2 \sup_{\mathcal{D}\sim\mathcal{H}} | \Pr_{x\sim D_S}[\mathcal{D}(x)=1] - \Pr_{x\sim D_{T_i}}[\mathcal{D}(x)=1]| \quad (1)$$

, where $\mathcal{H}$ denotes hypothetical space and $\mathcal{D}$ denotes discriminator. Drawing inspiration from [1, 8, 9], we employ the Jensen–Shannon (JS) divergence between two adjacent domains as an approximation of $\mathcal{H}$-divergence, as it has proven effective in distinguishing domain representations. When the inter-domain divergence is relatively small, it indicates a consistent feature representation less affected by cross-domain shifts [6].

$$JS(P_{D_S}||P_{D_{T_i}}) = \frac{1}{2}KL(P_{D_S}||\frac{P_{D_S}+P_{D_{T_i}}}{2}) + \frac{1}{2}KL(P_{D_{T_i}}||\frac{P_{D_S}+P_{D_{T_i}}}{2}) \quad (2)$$

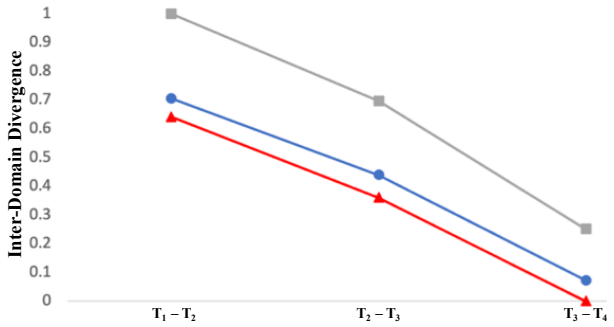

Figure 3. The inter-domain divergency. T_1 to T_4 represent the 4 target domains in ACDC, listed in sequential order.

Where $Kullback\text{-}Leibler\ (KL)\ divergence$ between two domain is

$$KL(P_1||P_2) = \sum_{i=0}^{n} P_1(x_i) log(\frac{P_1(x_i)}{P_2(x_i)}) \tag{3}$$

Utilizing P to represent the probability distribution of the model output features, we partition the output feature space into mutually disjoint intervals x_i, where n denotes the total number of samples in each target domain. As depicted in Figure 4 of the main paper, our proposed method exhibits a gradual reduction in inter-domain divergence.

Furthermore, we apply the same approach to calculate inter-domain divergence in the segmentation CTTA task, conducted on the Cityscapes-to-ACDC scenario. As depicted in Figure 3, the DaM mechanism yields smaller inter-domain divergence compared to the source model across all adjacent domains. Our method further reduces the divergence on all adjacent domains. The results demonstrate that DaM excels in extracting target domain knowledge, while HOG reconstruction increases stability in cross-domain learning and mitigates the impact of domain shift erosion.

C.2. Class Activation Mapping (CAM)

To empirically validate our intuition, we extend the use of CAM visualization to a larger set of samples within the ImageNet-C dataset. As illustrated in Figure 4, the results demonstrate a consistent trend with those presented in the submission. Specifically, when employing only the source model, the attention of the features appears scattered. This dispersion is a consequence of the domain shift influence, which hinders the model's ability to focus on foreground samples. In contrast, with the DaM mechanism, there is a noticeable concentration of attention on foreground samples, indicating that DaM assists the model in better understanding the target domain knowledge. Our approach leverages the domain-invariant property of HOG features. Through HOG reconstruction, we further enhance the model's task-relevant feature representations, allowing the output features

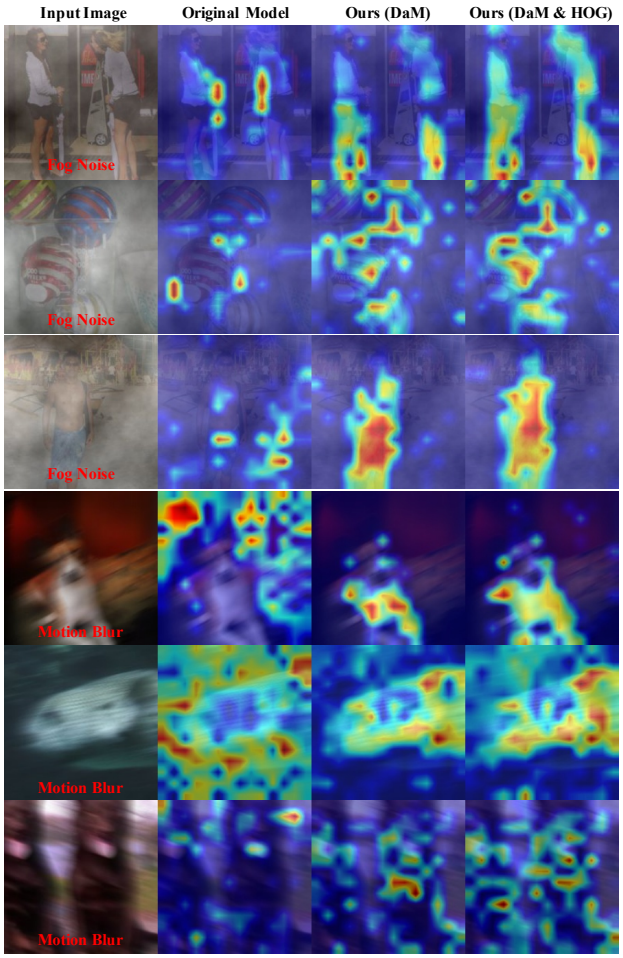

Figure 4. The CAM visualizations.

to disregard background domain shift and achieve higher response values on the foreground samples.

References

[1] Emily Allaway, Malavika Srikanth, and Kathleen McKeown. Adversarial learning for zero-shot stance detection on social media. *arXiv preprint arXiv:2105.06603*, 2021. 3

[2] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006. 3

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[4] Yulu Gan, Yan Bai, Yihang Lou, Xianzheng Ma, Renrui Zhang, Nian Shi, and Lin Luo. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In

Proceedings of the AAAI Conference on Artificial Intelligence, pages 7595–7603, 2023. 2

[5] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 3

[6] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 3

[7] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. 2013. 2

[8] Jiaming Liu, Senqiao Yang, Peidong Jia, Ming Lu, Yandong Guo, Wei Xue, and Shanghang Zhang. Vida: Homeostatic visual domain adapter for continual test time adaptation. *arXiv preprint arXiv:2306.04344*, 2023. 3

[9] Sebastian Ruder and Barbara Plank. Learning to select data for transfer learning with bayesian optimization. *arXiv preprint arXiv:1707.05246*, 2017. 3

[10] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 2

[11] 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. 2

[12] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090, 2021. 2