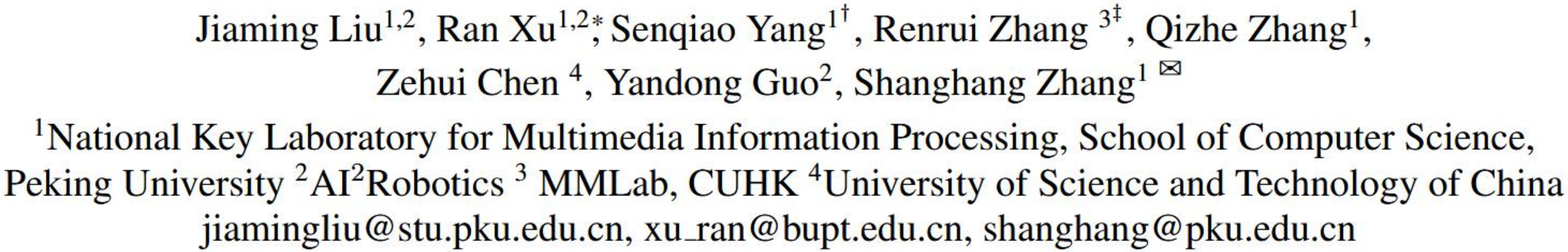
Continual-MAE: Adaptive Distribution Masked Autoencoders for Continual Test-Time Adaptation



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

CVPR 2024

Background

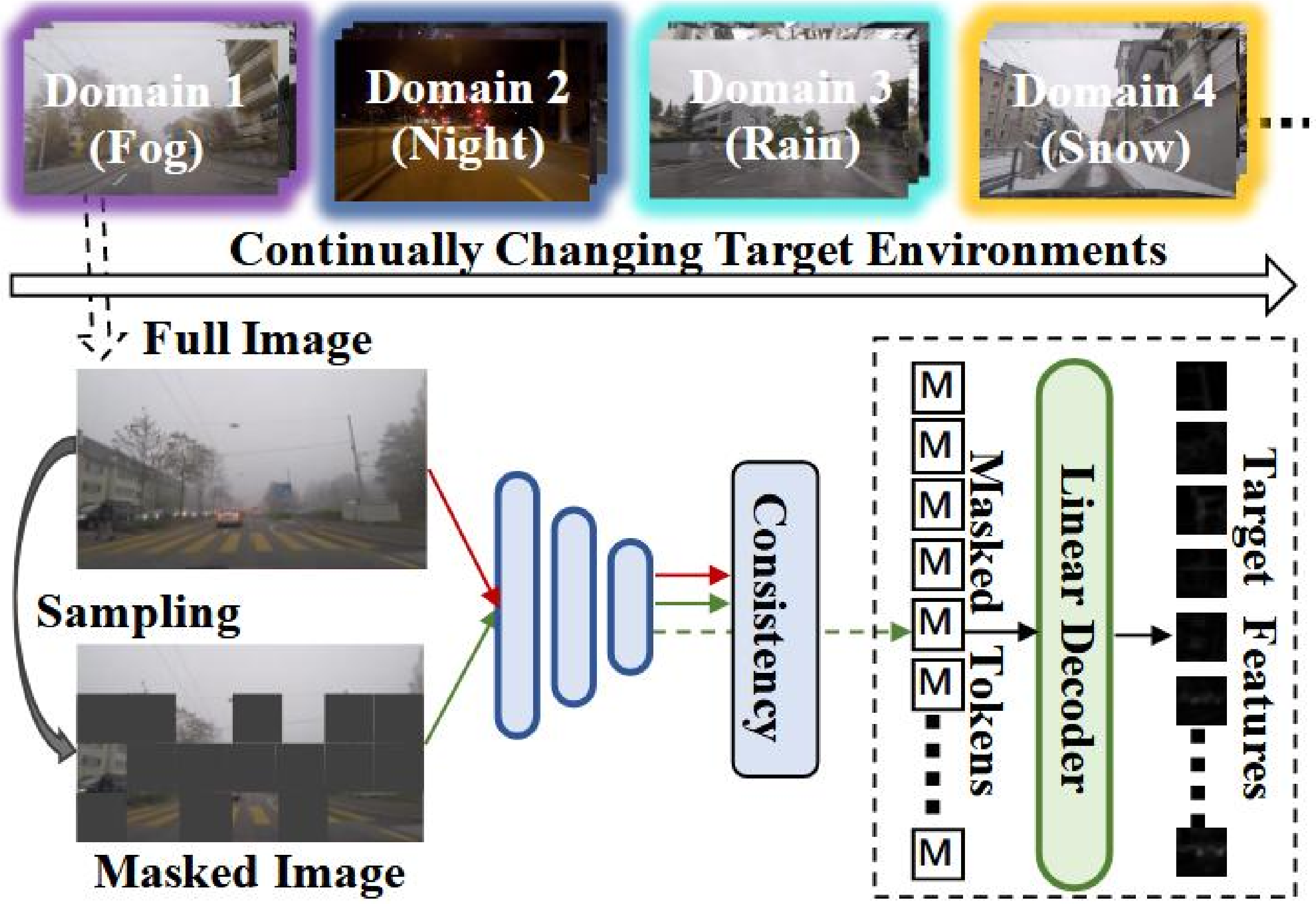
- Continual Test-Time Adaptation (CTTA) is proposed to migrate a source pre-trained model to continually changing target distributions, addressing real-world dynamism.

- Existing methods primarily focus on applying entropy minimization to update batch normalization layer or a fraction of model parameters , which already leads to a performance improvement in target domains.

- On the other hand, an alternative mainstream approach involves the teacher-student scheme for generating pseudolabels in target domains.

Motivation

a novel method for continual self-supervised learning that enhances the extraction of target domain knowledge while mitigating the accumulation of distribution shift.



Methods

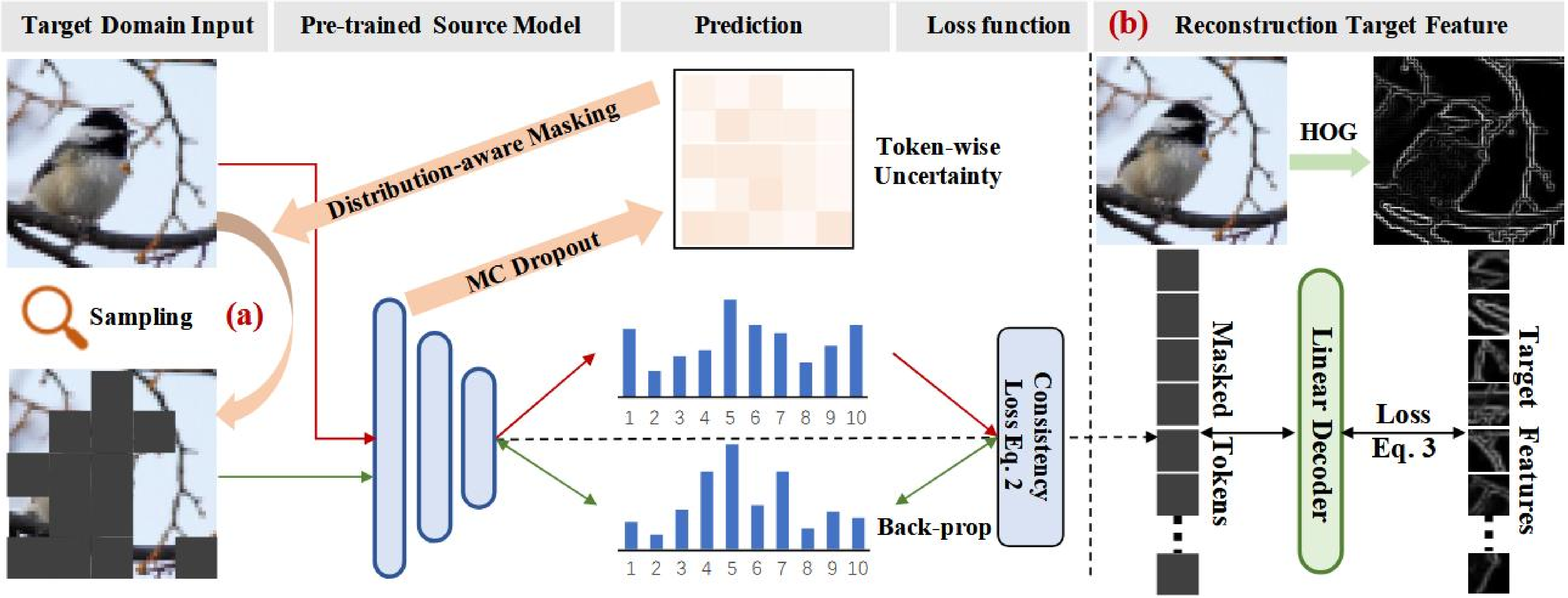


Figure 2. **The framework of Adaptive Distribution Masked Autoencoders (ADMA).** (a) We initiate the process by feeding the original target image into the model to generate features of the complete image. Simultaneously, this step facilitates the estimation of token-wise uncertainty, reflecting the token-wise distribution shift of each target sample, a process detailed in Sec. 3.2. Guided by the uncertainty values, we adaptively mask P% of the image tokens characterized by significant domain shifts, subsequently reintroducing the masked image into the model. In the classification task, the encoder's output embeddings are then fed into the classification heads, constructing a consistency loss (Eq. 2) between the two predictions. (b) For the masked tokens, we feed the masked token features into the linear decoder to compute the reconstruction loss (Eq. 3). We choose Histograms of Oriented Gradients (HOG) as the reconstruction target due to their invariant properties. Both losses are jointly optimized to address the CTTA problem.

Methods

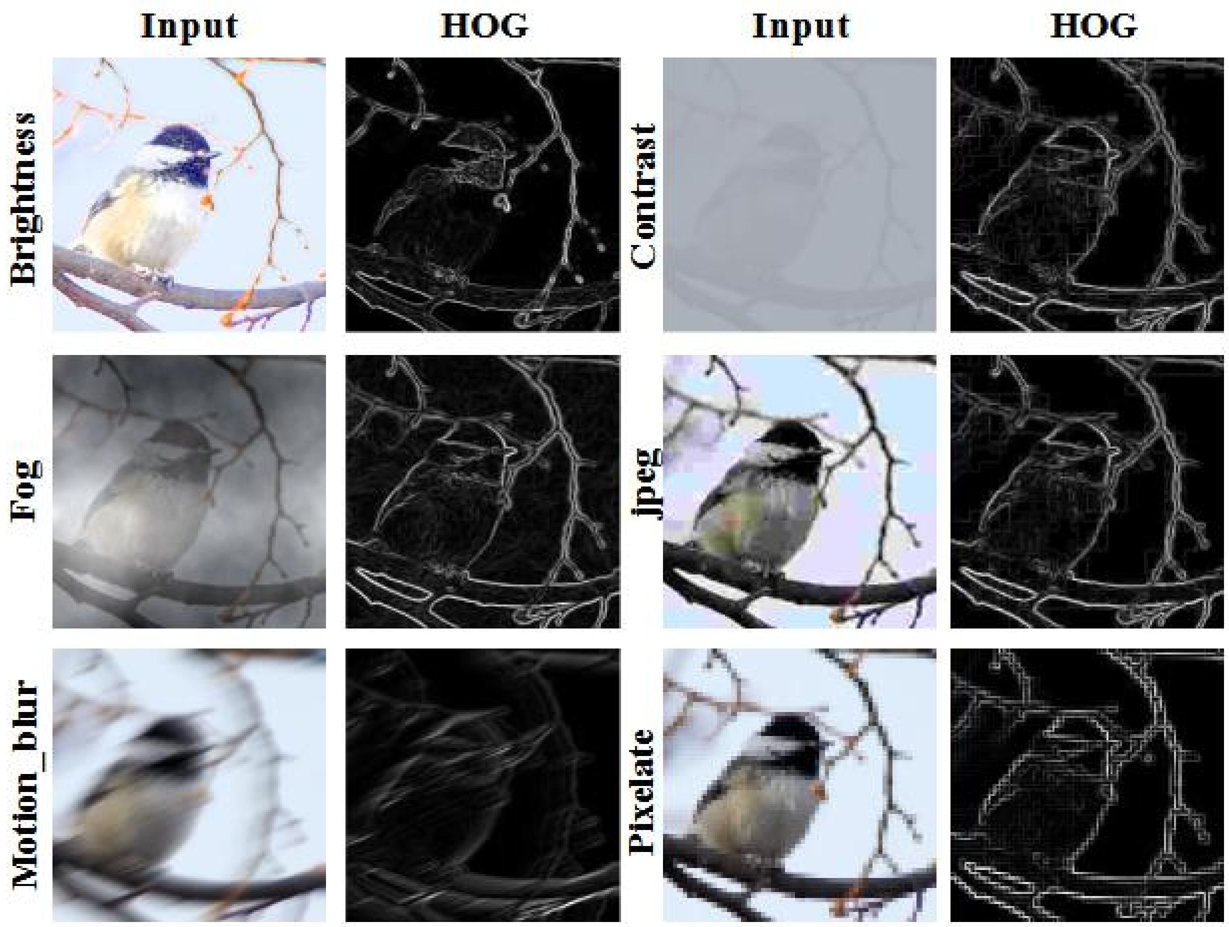


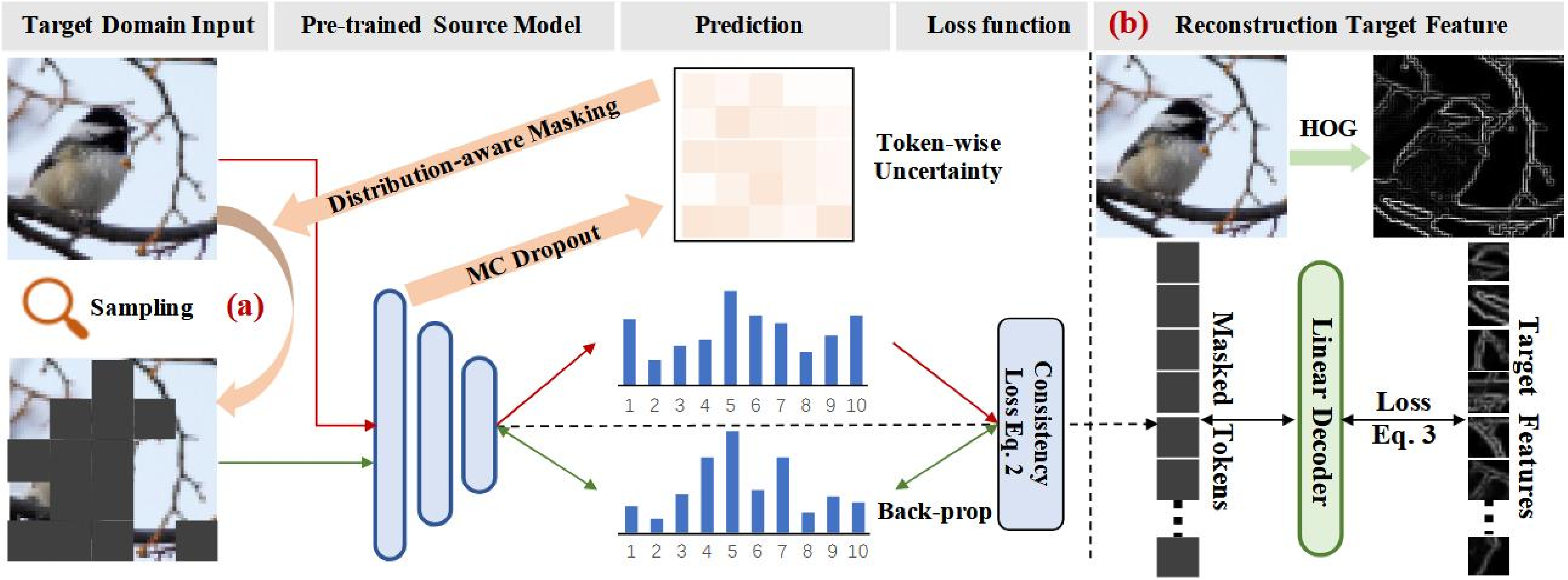
Figure 3. The visualization of HOG features in various target domain distributions (ImageNet-C [21]).

HOG is a feature descriptor that delineates the distribution of gradient orientations or edge directions within a localized subregion.

two advantages:
1) its inherent ability to capture local shapes and appearances ensures invariance to geometric changes.

2) the absorption of brightness through image gradients and local contrast normalization provides invariance to varying environments and weather conditions.

Methods



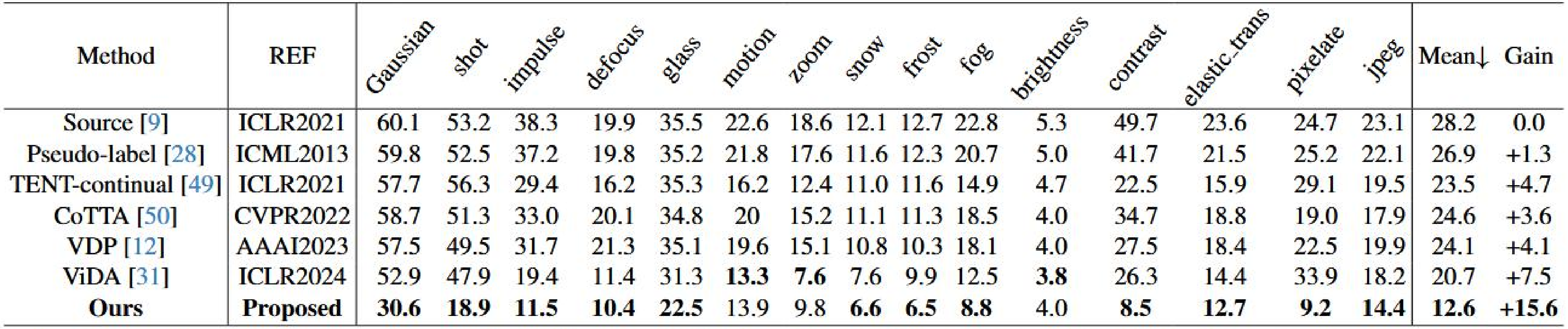
$$\mathcal{L}_{con}(x) = -\frac{1}{C}\sum_{c}^{C} y(c)\log \hat{y}(c)$$

$$\mathcal{L}_{rec} = \|P_{HOG} - F_{HOG}\|_2^2$$

Experiments

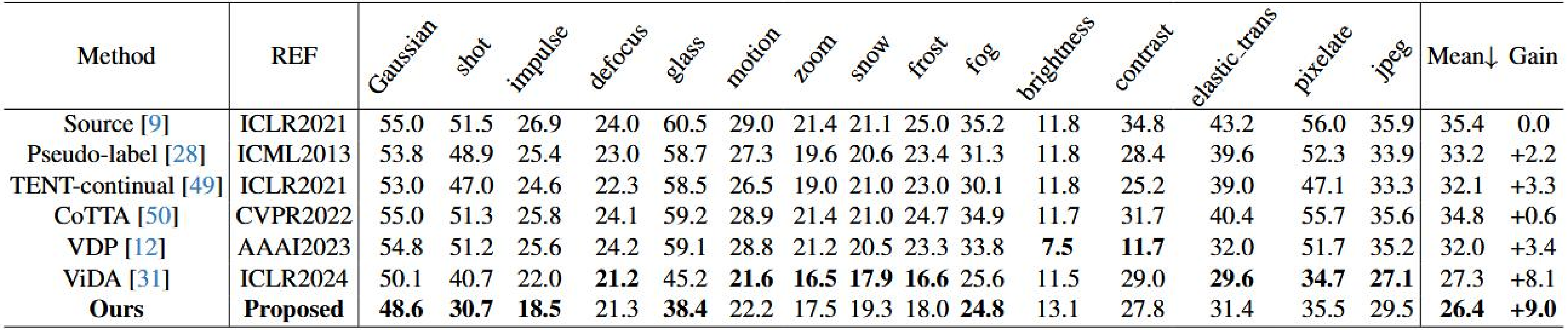
Task:classification CTTA task

Backbone Model:ViT-base



Method	REF	Gaussian	shot	impulse	defocus	glass	motion	zoom	snow	frost	fog	brightness	contrast	elastic_trans	pixelate	jpeg	Mean↓	Gain
Source [9]	ICLR2021	60.1	53.2	38.3	19.9	35.5	22.6	18.6	12.1	12.7	22.8	5.3	49.7	23.6	24.7	23.1	28.2	0.0
Pseudo-label [28]	ICML2013	59.8	52.5	37.2	19.8	35.2	21.8	17.6	11.6	12.3	20.7	5.0	41.7	21.5	25.2	22.1	26.9	+1.3
TENT-continual [49]	ICLR2021	57.7	56.3	29.4	16.2	35.3	16.2	12.4	11.0	11.6	14.9	4.7	22.5	15.9	29.1	19.5	23.5	+4.7
CoTTA [50]	CVPR2022	58.7	51.3	33.0	20.1	34.8	20	15.2	11.1	11.3	18.5	4.0	34.7	18.8	19.0	17.9	24.6	+3.6
VDP [12]	AAAI2023	57.5	49.5	31.7	21.3	35.1	19.6	15.1	10.8	10.3	18.1	4.0	27.5	18.4	22.5	19.9	24.1	+4.1
ViDA [31]	ICLR2024	52.9	47.9	19.4	11.4	31.3	**13.3**	**7.6**	7.6	9.9	12.5	**3.8**	26.3	14.4	33.9	18.2	20.7	+7.5
Ours	**Proposed**	**30.6**	**18.9**	**11.5**	**10.4**	**22.5**	13.9	9.8	**6.6**	**6.5**	**8.8**	4.0	**8.5**	**12.7**	**9.2**	**14.4**	**12.6**	**+15.6**

Table 1. Classification error rate(%) for CIFAR10-to-CIAFAR10C 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.



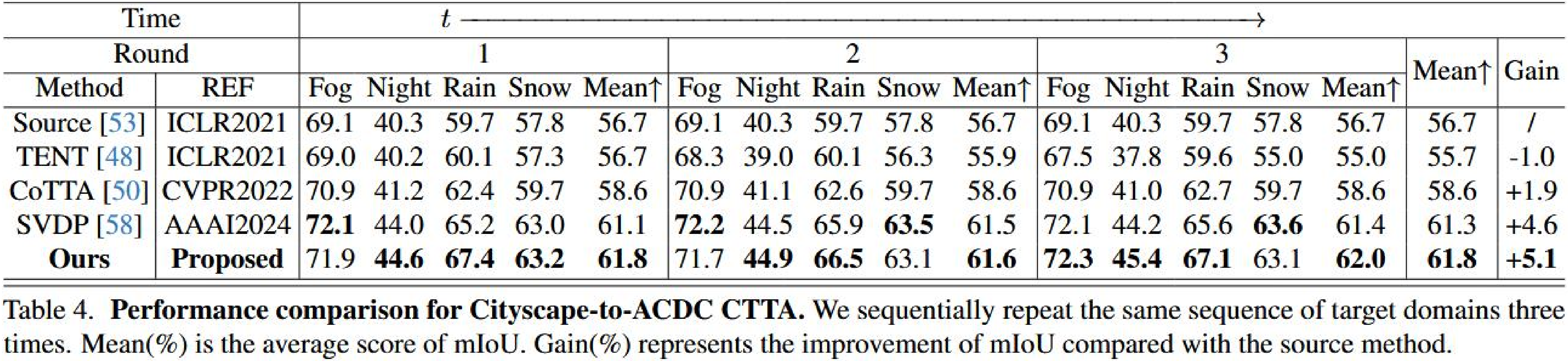
Method	REF	Gaussian	shot	impulse	defocus	glass	motion	zoom	snow	frost	fog	brightness	contrast	elastic_trans	pixelate	jpeg	Mean↓	Gain
Source [9]	ICLR2021	55.0	51.5	26.9	24.0	60.5	29.0	21.4	21.1	25.0	35.2	11.8	34.8	43.2	56.0	35.9	35.4	0.0
Pseudo-label [28]	ICML2013	53.8	48.9	25.4	23.0	58.7	27.3	19.6	20.6	23.4	31.3	11.8	28.4	39.6	52.3	33.9	33.2	+2.2
TENT-continual [49]	ICLR2021	53.0	47.0	24.6	22.3	58.5	26.5	19.0	21.0	23.0	30.1	11.8	25.2	39.0	47.1	33.3	32.1	+3.3
CoTTA [50]	CVPR2022	55.0	51.3	25.8	24.1	59.2	28.9	21.4	21.0	24.7	34.9	11.7	31.7	40.4	55.7	35.6	34.8	+0.6
VDP [12]	AAAI2023	54.8	51.2	25.6	24.2	59.1	28.8	21.2	20.5	23.3	33.8	**7.5**	**11.7**	32.0	51.7	35.2	32.0	+3.4
ViDA [31]	ICLR2024	50.1	40.7	22.0	**21.2**	45.2	**21.6**	**16.5**	**17.9**	**16.6**	25.6	11.5	29.0	**29.6**	**34.7**	**27.1**	27.3	+8.1
Ours	**Proposed**	**48.6**	**30.7**	**18.5**	21.3	**38.4**	22.2	17.5	19.3	18.0	**24.8**	13.1	27.8	31.4	35.5	29.5	**26.4**	**+9.0**

Table 2. Classification error rate(%) for CIFAR100-to-CIAFAR100C online CTTA task.

Experiments

Task:segmentation CTTA task

Backbone Model:Segformer-B5

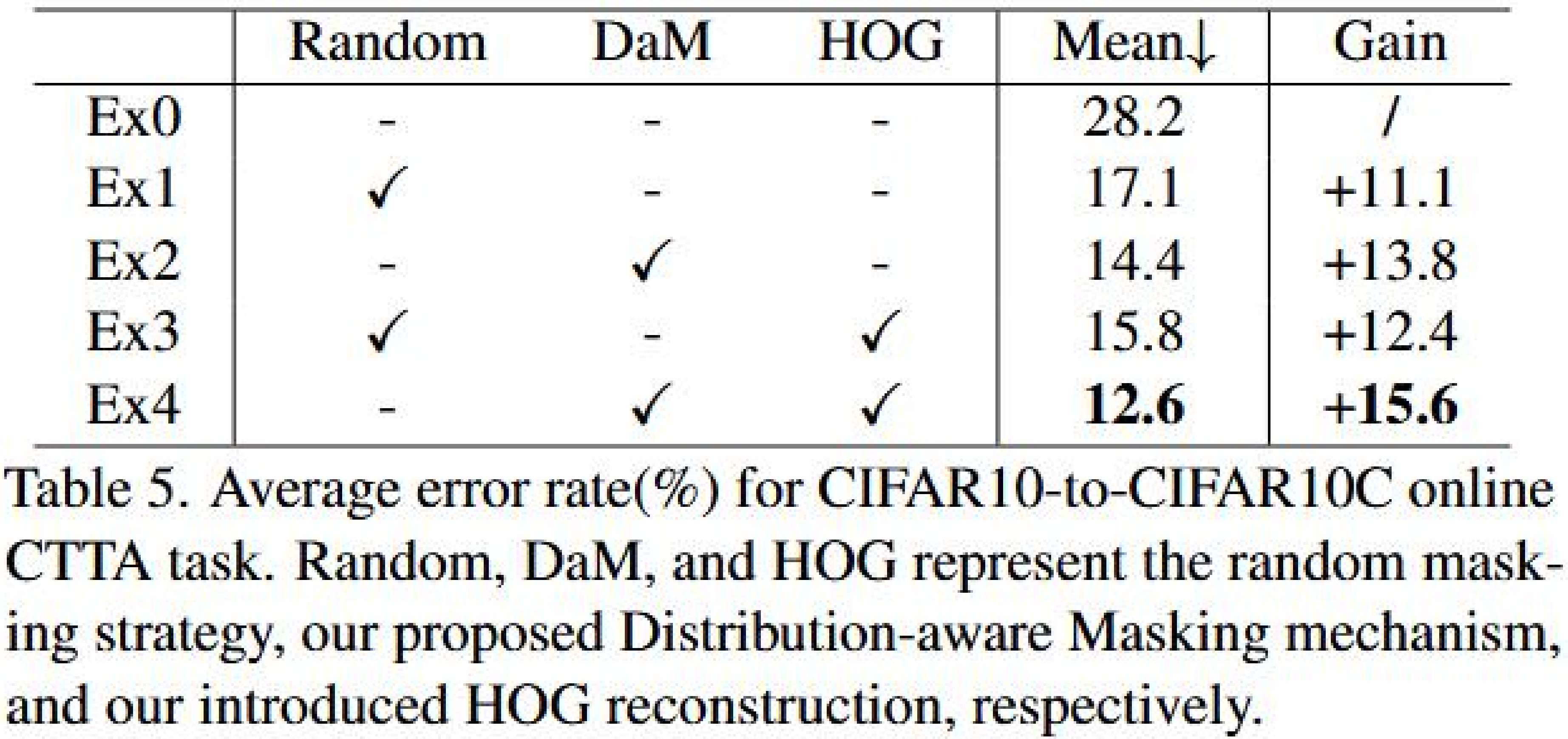


Time		$t \longrightarrow$																
Round		1					2					3					Mean↑	Gain
Method	REF	Fog	Night	Rain	Snow	Mean↑	Fog	Night	Rain	Snow	Mean↑	Fog	Night	Rain	Snow	Mean↑		
Source [53]	ICLR2021	69.1	40.3	59.7	57.8	56.7	69.1	40.3	59.7	57.8	56.7	69.1	40.3	59.7	57.8	56.7	56.7	/
TENT [48]	ICLR2021	69.0	40.2	60.1	57.3	56.7	68.3	39.0	60.1	56.3	55.9	67.5	37.8	59.6	55.0	55.0	55.7	-1.0
CoTTA [50]	CVPR2022	70.9	41.2	62.4	59.7	58.6	70.9	41.1	62.6	59.7	58.6	70.9	41.0	62.7	59.7	58.6	58.6	+1.9
SVDP [58]	AAAI2024	**72.1**	44.0	65.2	63.0	61.1	**72.2**	44.5	65.9	**63.5**	61.5	72.1	44.2	65.6	**63.6**	61.4	61.3	+4.6
Ours	**Proposed**	71.9	**44.6**	**67.4**	**63.2**	**61.8**	71.7	**44.9**	**66.5**	63.1	**61.6**	**72.3**	**45.4**	**67.1**	63.1	**62.0**	**61.8**	**+5.1**

Table 4. **Performance comparison for Cityscape-to-ACDC CTTA.** We sequentially repeat the same sequence of target domains three times. Mean(%) is the average score of mIoU. Gain(%) represents the improvement of mIoU compared with the source method.

Experiments

Ablation Study



	Random	DaM	HOG	Mean↓	Gain
Ex0	-	-	-	28.2	/
Ex1	✓	-	-	17.1	+11.1
Ex2	-	✓	-	14.4	+13.8
Ex3	✓	-	✓	15.8	+12.4
Ex4	-	✓	✓	**12.6**	**+15.6**

Table 5. Average error rate(%) for CIFAR10-to-CIFAR10C online CTTA task. Random, DaM, and HOG represent the random masking strategy, our proposed Distribution-aware Masking mechanism, and our introduced HOG reconstruction, respectively.

Thanks