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

Abstract

Continual Test-Time Adaptation (CTTA) is proposed to migrate a source pre-trained model to continually changing target distributions, addressing real-world dynamism. Existing CTTA methods mainly rely on entropy minimization or teacher-student pseudo-labeling schemes for knowledge extraction in unlabeled target domains. However, dynamic data distributions cause miscalibrated predictions and noisy pseudo-labels in existing self-supervised learning methods, hindering the effective mitigation of error accumulation and catastrophic forgetting problems during the continual adaptation process. To tackle these issues, we propose a continual self-supervised method, Adaptive Distribution Masked Autoencoders (ADMA), which enhances the extraction of target domain knowledge while mitigating the accumulation of distribution shifts. Specifically, we propose a Distribution-aware Masking (DaM) mechanism to adaptively sample masked positions, followed by establishing consistency constraints between the masked target samples and the original target samples. Additionally, for masked tokens, we utilize an efficient decoder to reconstruct a hand-crafted feature descriptor (e.g., Histograms of Oriented Gradients), leveraging its invariant properties to boost task-relevant representations. Through conducting extensive experiments on four widely recognized benchmarks, our proposed method attains state-of-the-art performance in both classification and segmentation CTTA tasks.

1. Introduction

Deep Neural Networks (DNNs) have demonstrated impressive performance across various computer vision tasks, including image-level classification [9, 18, 32], dense prediction [42, 53, 62], and multi-model task [29, 56, 57], when the test data distribution closely aligns with the training data. However, this assumption is frequently challenged in real-world scenarios due to dynamic environments, with deployed models exhibiting insufficient generalization capabilities and performance degradation [21, 45]. Therefore, the problem of continual test-time adaptation (CTTA) has been introduced [50], aiming to adapt a source pre-trained model to continually changing target distributions. Due to privacy and practical considerations, during the continual adaptation process, access to source domain data is not permitted,

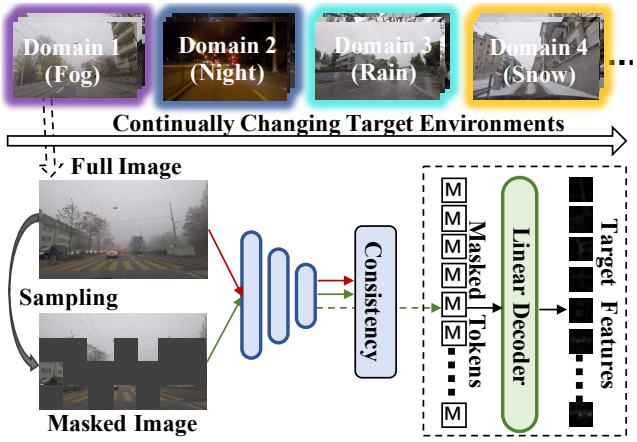

Figure 1. In continually changing environments, existing methods [46, 48] primarily focus on applying entropy minimization to update the normalization layer. However, these approaches are susceptible to miscalibrated predictions, resulting in uncontrollable error accumulation. Alternative mainstream approaches [11, 50] involve the teacher-student scheme for generating pseudo labels, but noisy pseudo labels limit the model's ability for continuous generalization. In this paper, we propose a novel approach to continual self-supervised learning known as Adaptive Distribution Masked Autoencoders (ADMA). ADMA introduces the mask reconstruction mechanism to enhance the extraction of target domain knowledge while mitigating the domain shift accumulation.

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

and each target data can only be accessed once. While the CTTA showcases promising potential applications, it also increases the difficulty of transfer learning, which introduces catastrophic forgetting and error accumulation problems.

Existing methods primarily focus on applying entropy minimization to update batch normalization layer [16, 35, 48] or a fraction of model parameters [46], which already leads to a performance improvement in target domains. Nonetheless, due to the continually changing environments, this self-training approach is susceptible to miscalibrated predictions, resulting in uncontrollable error accumulation. On the other hand, an alternative mainstream approach involves the teacher-student scheme for generating pseudo-labels in target domains. However, the traditional mean teacher method [47] yields noisy pseudo labels in dynamic environments, leading to the accumulation of distribution shifts. While [8, 31, 50] utilize the test-time augmentation method to enhance the accuracy of pseudo labels, it limits efficiency during the CTTA process.

In this paper, as shown in Figure 1, we introduce a novel approach to continual self-supervised learning called Adaptive Distribution Masked Autoencoders (ADMA). Classical masked autoencoders (MAE) [20] have the potential for various extensions and are becoming dominant in vision representation learning. The selection of reconstruction target and masked positions is particularly crucial during the pretraining process. Nonetheless, reconstructing low-level RGB signals in MAE is considered primitive and redundant, falling short of unlocking the potential of MAE in the context of downstream vision tasks [15, 22]. To create more effective reconstruction, several methods [2, 22, 51, 52] have explored the utilization of off-the-shelf vision foundation models [25, 37] as high-level supervisory signals and designed task-specific mask selection strategies. Different from previous MAE methods, we make the first attempt to introduce reconstruction techniques to address the continual adaptation problem. This innovation enhances the extraction of target domain knowledge while reducing the accumulation of distribution shifts.

Specifically, we propose a Distribution-aware Masking (DaM) mechanism to distinguish image patches that are target domain-specific from the less significant background patches. The objective is to enhance the quality of the target domain representation, preventing error accumulation and enhancing the efficiency of continuous adaptation. DaM dynamically selects masked positions based on token-wise uncertainty estimation and places learnable masks on token embeddings with substantial domain shifts. Subsequently, it establishes consistency constraints between the network outputs generated from the masked target samples and those from the original target samples. Furthermore, for the masked tokens, we employ an efficient decoder to reconstruct hand-crafted feature descriptors, such as Histograms of Oriented Gradients (HOG). In contrast to pixel colors and high-level feature reconstruction, HOG excels at capturing local shapes and appearances, exhibiting partial invariance to geometric and distribution changes [6, 51]. Consequently, we harness its invariant properties to acquire task-relevant representations in target domains, mitigating the impact of domain shifts during continual adaptation and preventing catastrophic forgetting problems. In summary, our contributions are as follows:

- We make the first attempt to introduce reconstruction techniques to address the CTTA problem. Our approach, Adaptive Distribution Masked Autoencoders (ADMA), is a novel method for continual self-supervised learning that enhances the extraction of target domain knowledge while mitigating the accumulation of distribution shift.
- In ADMA, we propose a Distribution-aware Masking (DaM) mechanism to adaptively place learnable masks on token embeddings with significant distribution shifts, promoting the quality of target domain representation and improving continual adaptation efficiency.
- For masked tokens, we utilize an efficient decoder to reconstruct histograms of oriented gradients, leveraging its invariant properties to boost task-relevant representations and prevent the catastrophic forgetting problem.
- Our proposed approach surpasses previous state-of-the-art methods, as demonstrated in experiments across four benchmark datasets, covering both classification and segmentation CTTA. Notably, our method attains a promising 87.4% accuracy in CIFAR10-to-CIFAR10C and 61.8% mIoU in Cityscapes-to-ACDC scenarios.

2. Related Work

2.1. Continual Test-Time Adaptation

Test-time adaptation (TTA), also known as source-free domain adaptation [4, 27, 55, 60], is the process of adapting a source model to a target domain distribution without relying on any source domain data. Recent works have delved into techniques such as self-training and entropy regularization to fine-tune the source model [5, 30]. Tent [49] achieves this by updating the training parameters in batch normalization layers through entropy minimization. This approach has served as inspiration for subsequent research efforts in recent works [36, 59], which continue to investigate the robustness of normalization layers.And [13] first attempts to reconstruct RGB images in the TTA task. **Continual Test-Time Adaptation (CTTA)** denotes a scenario where the target domain is dynamically changing, introducing additional challenges for conventional TTA methods. The initial approach is presented in [50], which employs a teacher-student framework to integrate bi-average pseudo labels and stochastic weight reset. Drawing from the insight that mean teacher predictions are often more robust than standard models [47], a series

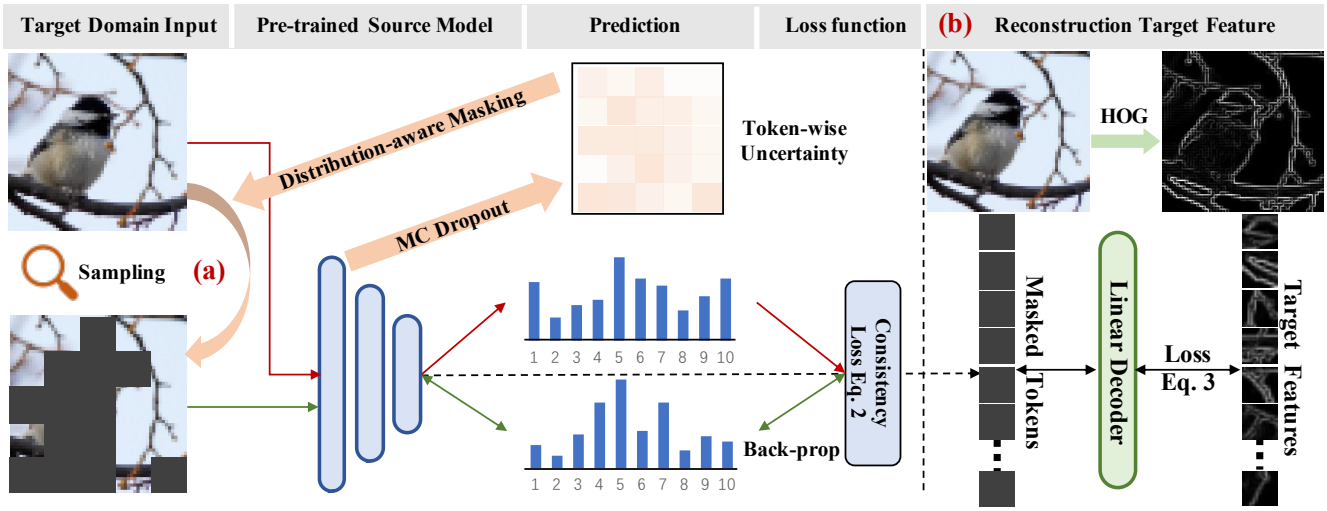

Figure 2. **The framework of Adaptive Distribution Masked Autoencoders (ADMA).** (a) We initiate the process by feeding the original target image into the model to generate features of the complete image. Simultaneously, this step facilitates the estimation of token-wise uncertainty, reflecting the token-wise 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.

of mainstream methods [8, 12, 31] continue this approach to self-supervised learning in CTTA. Concurrently, existing methods also focus on applying entropy minimization to update normalization layers [16, 48] or a fraction of model parameters [34, 46]. However, due to continually changing environments, these self-training approaches are susceptible to miscalibrated predictions and noisy pseudo-labels, resulting in uncontrollable error accumulation.

2.2. Masked Image Modeling

Mask-reconstruction-based self-supervised learning has been successful in reducing the reliance on extensive labeled datasets in both Natural Language Processing (NLP) and Computer Vision (CV). The concept was first introduced by BERT [7] in NLP, where a portion of the input word tokens is randomly masked, and the model learns to reconstruct the vocabularies of these masked tokens. In the field of computer vision (CV), similar techniques have been applied in various works [3, 20, 54]. These methods involve randomly masking a significant percentage of input image patches. Specifically, BEiT [3] was the first to explore Masked Image Modeling (MIM) in vision transformers by reconstructing the vision dictionary derived from DALL-E [40, 41]. MAE [20] scaled up MIM to larger models and demonstrated that a simple pixel reconstruction loss can enhance the visual representations of pre-trained models. However, relying solely on low-level RGB signals in MAE is considered rudimentary and limited in unlocking the full potential of MAE in downstream vision tasks. Subsequently, approaches like MaskFeat [51], data2vec [2], MVP [52], and MILAN [22] have uncovered various high-level signals [17, 61], including pre-trained DINO features [37], hand-crafted features [6], momentum features [19], and multi-modality features [39]. Utilizing these high-level signals has been proven to be more effective in extracting contextual information.

3. Method

3.1. Overview

Preliminary. In Continual Test-Time Adaptation (CTTA), we first pre-train the model $q(y|x)$ using the source domain data $D_S = (Y_S, X_S)$. Subsequently, we adapt $q(y|x)$ to dynamic target domains, denoted as $D_{T_1}, D_{T_2}, ..., D_{T_n}$, where $D_{T_i} = (X_{T_i})_{i=1}^{n}$ and n represents the number of continual target datasets. The entire process is restricted from accessing source data and can only utilize each target sample once. Our goal is to continually adapt the pre-trained model to these target domains while mitigating issues such as error accumulation and catastrophic forgetting.

Adaptive Distribution Masked Autoencoders Prevalent CTTA approaches primarily focus on applying entropy minimization to update the batch normalization layer [16, 35, 48] or a fraction of model parameters [46]. However, due to the dynamic environments, this self-training approach is susceptible to miscalibrated predictions, resulting in uncontrollable error accumulation. On the other hand, alternative mainstream approaches utilize the teacher-student pseudo-labeling in continual target domains. Nevertheless, the con-

ventional mean teacher method [47] increases computational cost and produces noisy pseudo labels in dynamic environments, leading to the accumulation of distribution shifts. While methods such as [8, 31, 50] utilize Test-time augmentation methods to enhance the accuracy of pseudo labels, they may limit efficiency during the CTTA process. Different from prior continual self-supervised approaches, we lead the way in incorporating a masked autoencoder (MAE) to tackle the CTTA problem, abandoning the impact of miscalibrated predictions and the cumbersome teacher-student model. Our key insight lies in adopting the reconstruction scheme to effectively extract target domain knowledge while mitigating the accumulation of domain shift. The overall framework is illustrated in Figure 2. Specifically, We propose a Distribution-aware Masking (DaM) mechanism, detailed in Sec. 3.2, which adaptively places learnable masks on token embeddings with substantial distribution shifts. By establishing consistency constraints between the masked input and the original input, DaM significantly enhances the understanding of target domain knowledge and mitigates the challenge of error accumulation. For masked tokens, we adopt a linear decoder to reconstruct Histograms of Oriented Gradients, leveraging its invariant properties to acquire task-relevant representations while averting the introduction of target domain shifts, as elaborated in Sec. 3.3. This reconstruction method serves as a preventive measure, avoiding catastrophic forgetting in continual adaptation. We provide the intuitive explanation and justification in Sec. 5.

3.2. Distribution-aware Masking

To establish masked image reconstruction as a meaningful pretext task, previous studies have commonly applied an aggressive masking approach by randomly masking a substantial portion of input image patches [20, 51]. This strategy, however, introduces a potential drawback: the remaining visible patches may predominantly comprise background information, potentially lacking the crucial cues essential for reconstructing foreground details [22]. In our framework, the precise selection of masked positions is crucial due to significant distribution shifts in each target sample, impacting the representational capacity of visible patches. Furthermore, in the CTTA task, unlike traditional MAE pre-training, each sample is encountered only once, demanding high efficiency in the reconstruction process.

To this end, as shown in Figure 2 (a), we introduce a Distribution-aware Masking (DaM) strategy, enabling a thoughtful selection of tokens to mask in dynamic environments. The key concept involves choosing tokens with substantial domain shifts for masking, ensuring that the preserved visible tokens exhibit relatively fewer domain shifts while providing reliable semantic knowledge through the model encoder. To quantify token-wise distribution shifts, we draw inspiration from [38, 43] and introduce a method for token-wise uncertainty estimation. Specifically, we employ the MC Dropout [10], enabling multiple forward propagations to obtain m (e.g., $m = 10$) sets of features for each token. Subsequently, we calculate the uncertainty value $\mathcal{U}(x)$ for a given token x_j, as formulated below:

$$\mathcal{U}(x_j) = \left(\frac{1}{m} \sum_{i=1}^{m} \| f_i(x_j) - \mu \|^2 \right)^{\frac{1}{2}} \tag{1}$$

Where $f_i(x_j)$ is the feature value of the token x_j in the i^{th} forward propagation, and μ is the average value of the token feature over m forward propagations. To calculate the feature value within a token, we utilize average pooling, reducing the token's dimension from 1×768 to 1×1. Note that we only apply MC Dropout to the linear layer within the FFN layer in the first Transformer block. We conduct m forward propagations in the local FFN layer, calculating the uncertainty value does not significantly increase computational cost. In this manner, we obtain the uncertainty value for each token. After sorting, we select the top P% (e.g., 50%) of tokens with the highest uncertainty for masking.

We input the masked image into the model, leveraging the remaining contextual clues to reconstruct the class label. Subsequently, we establish consistency constraints between the network outputs $(\hat{y}(c), y(c))$ generated from the masked target samples and those from the original target samples. The formulation of the consistency loss is as follows:

$$\mathcal{L}_{con}(x) = -\frac{1}{C} \sum_{c}^{C} y(c) \log \hat{y}(c) \tag{2}$$

Where $\hat{y}$ is the output from the masked image, C means the number of categories. Through DaM and consistency constraints, we can mask a substantial amount of distribution shift, learning the target domain contextual knowledge while avoiding domain shift accumulation.

3.3. Reconstruction Target Feature

The reconstruction target plays an important role in masked image modeling, exerting a direct influence on the learning of feature representations. Previous MAE methods typically opt for either low-level RGB information [20] or high-level semantic information [15, 22] as the reconstruction target. However, in the CTTA tasks, reconstructing the RGB signal of the target domain introduces inherent domain shifts in the reconstruction process. Similarly, when reconstructing the semantic features of the target domain, such as features from CLIP [39], the absence of any operation to domain transfer during the feature extraction process means that this approach still fails to alleviate the domain shift. These statements are demonstrated in Sec. 4.4. Therefore, in the face of continual distribution shifts, the choice of the reconstruction target gains greater importance.

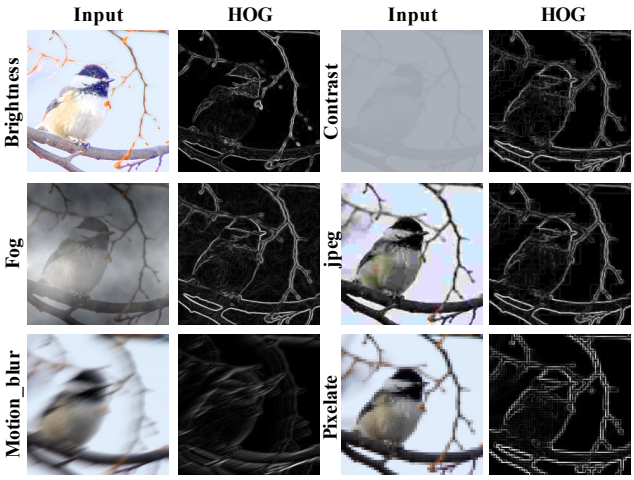

Figure 3. The visualization of HOG features in various target domain distributions (ImageNet-C [21]).

Inspired by [51] in model pretraining, we introduce Histograms of Oriented Gradients (HOG) reconstruction in CTTA tasks, showcasing notable benefits in the continual adaptation process. HOG is a feature descriptor that delineates the distribution of gradient orientations or edge directions within a localized subregion [6]. Using HOG as the reconstruction target in CTTA offers two advantages: 1) its inherent ability to capture local shapes and appearances ensures invariance to geometric changes, and 2) the absorption of brightness through image gradients and local contrast normalization provides invariance to varying environments and weather conditions. As illustrated in Figure 3, the visualization of HOG under various target domain distributions presents similar feature representations, clearly indicating their invariant properties.

To obtain HOG features, we employ a two-channel convolution that generates gradients along the x and y axes, followed by histogramming and normalization. Following [51], we configure the orientation bins to be 9, spatial cell size to be 8×8, and channels to be 3. Consequently, the HOG feature $F_{HOG} \in \mathbb{R}^{3\times9\times H/8\times W/8}$, where H and W represent the height and width of the input image. After obtaining the HOG features, we employ a linear layer to project learnable masked tokens to the same dimension as F_{HOG}, minimizing the L2 distance between the HOG prediction P_{HOG} and HOG label F_{HOG} of the masked token positions. The reconstruction loss is formulated as:

$$\mathcal{L}_{rec} = \|P_{HOG} - F_{HOG}\|_2^2 \tag{3}$$

Through HOG reconstruction, we harness its invariant properties to extract task-relevant knowledge in the CTTA problem. This enables the model to concentrate more on the task at hand and mitigates the impact of domain shift, consequently reducing the risk of catastrophic forgetting.

3.4. Optimization Objective

In the ongoing adaptation process, we update the model by incorporating the total loss formulated as Eq. 4, where $\lambda = 0.5$ serves as a balancing factor for the loss values.

$$\mathcal{L}_{all} = \mathcal{L}_{con} + \lambda \times \mathcal{L}_{rec} \tag{4}$$

4. Experiments

In Sec. 4.1, we present experiments and implementation details. Sec. 4.2 and Sec. 4.3 provide a comparative analysis of our approach against previous methods in classification and semantic segmentation CTTA tasks. Furthermore, we conduct a comprehensive ablation study in Sec. 4.4. Due to page constraints, additional quantitative and qualitative analyses are available in Appendices A and B, respectively.

4.1. Experiments Details

Datasets. Our method undergoes evaluation on three classification CTTA benchmarks, which encompass CIFAR10-to-CIFAR10C, CIFAR100-to-CIFAR100C [26], and ImageNet-to-ImageNet-C [21]. In the segmentation CTTA, following the definition by [50, 58], we conduct assessments on the Cityscapes-to-ACDC, using the Cityscapes [58] as the source domain and the ACDC [45] as the target domain.

Task setting. Following the task setting outlined in [31, 50], in the classification CTTA tasks, we utilize a sequential adaptation process. The pre-trained source model adapts to each of the fifteen target domains characterized by the largest corruption severity (level 5). Online prediction results are immediately assessed after processing the input. For the segmentation CTTA task, we use the ACDC [45] dataset to represent the target domain, which includes images captured under four distinct unobserved visual conditions: Fog, Night, Rain, and Snow. To simulate continuous environmental changes resembling real-world scenarios, we cyclically iterate through the same sequence of target domains (Fog $\rightarrow$ Night $\rightarrow$ Rain $\rightarrow$ Snow) multiple times.

Implementation Details. In our CTTA experiments, to ensure consistency and fair comparisons, we follow the implementation details as proposed in the prior CTTA works [31, 50]. For the classification CTTA tasks, we employ the ViT-base [9] as our backbone model. We resize the input images to 384×384 for CIFAR10C, CIFAR100C benchmark, and 224×224 resolution for ImageNet-C benchmark. In the case of segmentation CTTA task, we employ the Segformer-B5 [53] pre-trained on the Cityscapes dataset as our source model. We down-sample the original input images resolution from 1920×1080 to 960×540. We use Adam [24] with $(\beta_1, \beta_2) = (0.9, 0.999)$ as the optimizer. Different learning rates are assigned for different CTTA tasks and backbone models, such as we use 1e-5 for ViT on CIFAR10C and CIFAR100C, 1e-3 for ViT on ImageNetC, and 3e-4 for Segformer on ACDC. In the masking process, following the

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.

approach in MIC [23], we use [MASK] tokens initialized with all zeros to replace 50% of tokens that are embedded as patches. These [MASK] tokens are shared learnable embeddings that indicate masked patches. The reconstruction decoder is a randomly initialized linear layer used to project the output tokens associated with masked patches to HOG features. The parameters of mask tokens and projection layer are optimized in parallel with other parameters. To achieve better results, we inject ViDA, as proposed by previous SOTA work [31], into the model and utilize our proposed self-supervised method to update.

4.2. Classification CTTA Tasks

CIFAR10-to-CIFAR10C & CIFAR100-to-CIFAR100C. The source model is trained through supervised learning on the CIFAR10 or CIFAR100 dataset. During testing, we apply CTTA to the CIFAR10C or CIFAR100C dataset, which contains fifteen corruption types continuously fed into the model in a specific order. In the CIFAR10-to-CIFAR10C scenario, as shown in Table 1, the average classification error of the source model reaches 28.2% when directly testing on CIFAR10C. However, our method significantly reduces the error to 12.6%. Compared to other continual self-supervised methods, our approach outperforms them by 10.9% and 12.0% compared to the previous entropy minimization method (TENT [49]) and the teacher-student method (CoTTA [50]), demonstrating significant potential in addressing the CTTA problem. To be mentioned, our Adaptive Distribution Masked Autoencoders (ADMA) demonstrate outstanding performance across 12 out of the 15 corruption types, validating the robustness of our method in the continual adaptation process. Note that, for TENT, we implement entropy minimization to update the Layer Normalization layers in the transformer instead of BN.

Target	Method	Source	Tent	CoTTA	VDP	**Ours**
ImageNet-C	Mean↓	55.8	51.0	54.8	50.0	**42.5**
	Gain	0.0	+4.8	+1.0	+5.8	**+13.3**

Table 3. Average error rate (%) for the ImageNet-to-ImageNet-C CTTA. The fine-grained performances are shown in Appendix A.

We expand our evaluation to the CIFAR100-to-CIFAR100C scenario, as depicted in Table 2, encompassing a more extensive range of categories within each domain. Our approach outperforms the previous entropy minimization and teacher-student pseudo-labeling methods by 5.7% and 8.4%, respectively, exhibiting a consistent trend with the aforementioned CTTA experiments. Therefore, the results validate the universality of our method, unaffected by the number of categories, and it can efficiently mitigate error accumulation and catastrophic forgetting problems.

ImageNet-to-ImageNet-C. In order to comprehensively evaluate the effectiveness of our method, we conduct experiments on the ImageNet-to-ImageNet-C scenario. The source model is pre-trained on the ImageNet. As indicated in Table 3, due to the large number of categories in the ImageNet-C, previous entropy minimization (TENT [49]) and the teacher-student approach (CoTTA [50]) only achieve error rates of 51.0% and 54.8%, respectively. In contrast, our proposed method achieves the best performance, showcasing a significant reduction in classification error rates to 42.5%. This outcome further demonstrates the effectiveness of our method, enhancing the feature representation of the target

Time		t ⟶																
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.

	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.

Target	RGB	CLIP	Mean-Teacher	HOG
Error rate	49.2	48.5	48.1	**43.6**
Target	SIFT	Sobel	Laplacian	ORB
Error rate	50.4	48.2	48.7	50.0

Table 6. Average error rate(%) for ImageNet-to-ImageNet-C. RGB, CLIP, Mean-Teacher, SIFT [33], ORB, edge detectors(Sobel and Laplacian), and HOG are different reconstruction target.

domain without succumbing to domain shift accumulation.

4.3. Semantic Segmentation CTTA Task

Cityscapes-to-ACDC. We validate the effectiveness of our approach in the more challenging segmentation CTTA task by adapting the pre-trained Segformer model from the Cityscapes dataset to the ACDC [45] dataset, as depicted in Table 4. To ensure the reliability of the model's pixel-level outputs, we adopt the update strategy from previous works [31, 50], such as utilizing multi-scale augmentation. Our proposed method exhibits a notable improvement, achieving a 5.1% increase in mIoU over the source model, thereby confirming its efficacy in dense prediction CTTA tasks. Moreover, our method outperforms the previous entropy minimization method (TENT [49]) and the teacher-student method (CoTTA [50]) by 6.1% and 3.2%, respectively. It is worth highlighting the stability of our method in comparison to others. While TENT [49] reduces 1.7% performance from the first to the third round of experiments, CoTTA [50] maintains consistent results between these rounds. In contrast, our method demonstrates a 0.2% increase in mIoU during the third round. This observation underscores the effectiveness of our approach in extracting target domain knowledge through efficient mask modeling. Additionally, our reconstruction scheme ensures task-relevant feature representation, mitigating catastrophic forgetting.

4.4. Ablation Study

Effectiveness of each component. We initially conduct a series of ablation experiments on CIFAR10-to-CIFAR10C to assess the contributions of different components in our approach. As shown in Table 5, the first set of experiments (Ex1) involve randomly masking a portion of patches from the input image and establishing consistency constraints between the model outputs generated from the masked target samples and those from the original target samples. This directly led to an 11.1% reduction in the error rate compared to the source method (Ex0), indicating that the masking strategy enhances target domain knowledge extraction in CTTA. In the second set of experiments (Ex2), we replace the random masking strategy with our proposed Distribution-aware Masking (DaM) mechanism. Ex2 further reduced the error rate by 2.7%, validating that DaM helps the model more efficiently understand the target domain distribution. In the subsequent experiments (Ex3 and Ex4), we introduced the Histogram of Oriented Gradients (HOG) reconstruction scheme. For Ex4, HOG reconstruction contributed to a 1.8% accuracy improvement than Ex2. When integrating HOG reconstruction into our proposed DaM, a final classification error rate of 12.6% is achieved, leading to an overall performance improvement of 15.6%. These results confirm that leveraging HOG reconstruction during the continual adaptation process assists the model in learning task-relevant knowledge under the presence of domain shifts.

Reconstruction target selection. Another set of ablation experiments aims to assess the impact of the reconstruction target. This includes the original RGB pixel, high-level CLIP features, Mean-Teacher features, as well as SIFT, ORB, Sobel, Laplacian, and HOG features. Since the CIFAR datasets have an input size of 32×32 with limited RGB pixel information, these ablation experiments are conducted in the context of the ImageNet-to-ImageNet-C. As shown in Table 6, our introduced HOG reconstruction outperforms other reconstruction targets by a significant margin. Specifically, compared with classical image pixel reconstruction, our method achieves a 5.6% improvement, demonstrating that reconstructing the RGB signal of the target domain intro-

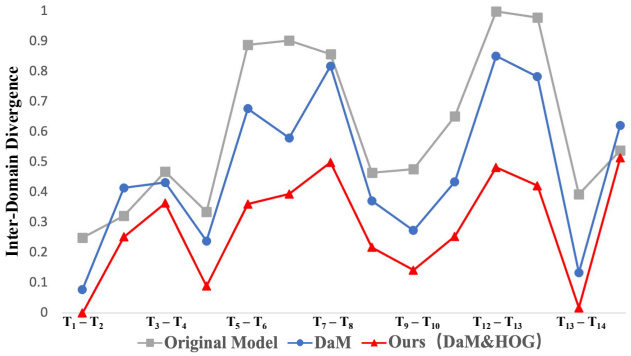

Figure 4. The inter-domain divergency. T_1 to T_{15} represent the 15 target domains in CIFAR-10C, listed in sequential order.

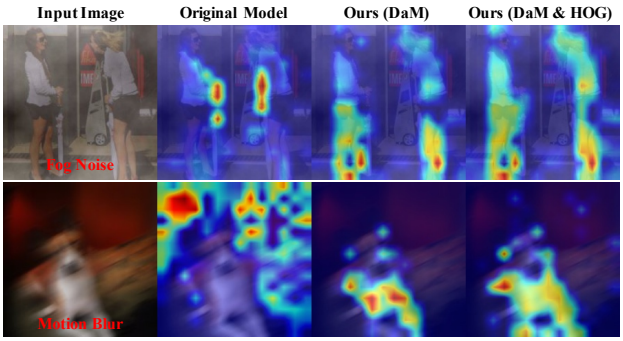

Figure 5. The CAM visualizations.

duces inherent domain shifts. In contrast to reconstructing the CLIP feature, our method outperforms it by 4.9%, while also significantly reducing computational costs as it does not require an additional model. The results validate that features directly extracted by the large-scale model still introduce domain shifts. Lastly, while reconstructing features of the Mean-Teacher model is slightly better (by 0.4%) than reconstructing CLIP features, it is still 4.5% lower than our method. In conclusion, reconstructing HOG and leveraging its invariant properties can boost task-relevant representations and avoid domain shift accumulation in CTTA.

5. Discussion and Justification

In this section, we offer an intuitive justification for our proposed method, seeking to demonstrate its efficiency in extracting target domain knowledge while avoiding domain-shift accumulation. Additional details on these justifications can be found in Appendix C.

Inter-Domain Divergence. To provide clearer evidence for the intuition of our proposed DaM and HOG reconstruction mechanism, we calculate the distribution distances of the feature representations across different target domains. Following prior works [1, 31, 44], we compute the Jensen–Shannon (JS) divergence between two adjacent domains to represent the inter-domain divergence. If the inter-domain divergence is small, it indicates that the feature representation remains consistent and is less susceptible to cross-domain shifts [14]. For comparative experiments, we compute the inter-domain JS divergence of the source model, using the DaM mechanism and our proposed method on the CIFAR10-to-CIFAR10C. As illustrated in Figure 4, when using the DaM mechanism, the inter-domain divergence is smaller than the source model on the vast majority of adjacent domains. Our method achieves the minimum inter-domain divergence on all fourteen adjacent domains. The observed pattern in inter-domain divergence suggests that DaM excels in extracting target domain knowledge but may exhibit limitations in robustness during continual adaptation. Meanwhile, after the HOG reconstruction of masked tokens, the model attains increased stability in cross-domain learning, mitigating the impact of domain shift erosion.

Class Activation Mapping (CAM). To directly validate our intuition, we employ CAM visualization on the ImageNet-C dataset. As shown in Figure 5, when utilizing only the source model, the attention of the features appears scattered. This dispersion is a consequence of the domain shift influence, causing the model to struggle in focusing 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 explores the domain-invariant property of HOG features. Through HOG reconstruction, we further enhance the model's task-relevant feature representations, enabling the output features to disregard background domain shift and attain higher response values on the foreground samples.

6. Conclusion

In this paper, we pioneer the integration of reconstruction techniques to tackle the Continual Test-Time Adaptation (CTTA) problem. Our contribution, the Adaptive Distribution Masked Autoencoders (ADMA) method, represents a novel approach to continual self-supervised learning. ADMA is designed to enhance the extraction of target domain knowledge while mitigating the accumulation of distribution shift, thereby addressing the issues of error accumulation and catastrophic forgetting. The proposed Distribution-aware Masking (DaM) mechanism plays a pivotal role in promoting the extraction of target domain knowledge, improving the efficiency of continual adaptation. Simultaneously, our introduced HOG reconstruction strategy elevates the task-relevant representations of the model and acts as a preventive measure against distribution shift accumulation.

Acknowledgements. Shanghang Zhang is supported by the National Science and Technology Major Project of China (No. 2022ZD0117801).

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. 8

[2] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022. 2, 3

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3

[4] Malik Boudiaf, Tom Denton, Bart van Merriënboer, Vincent Dumoulin, and Eleni Triantafillou. In search for a generalizable method for source free domain adaptation. 2023. 2

[5] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. *ArXiv*, abs/2204.10377, 2022. 2

[6] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pages 886–893. Ieee, 2005. 2, 3, 5

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[8] 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. 2, 3, 4

[9] 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. 1, 5, 6

[10] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 4

[11] Yulu Gan, Xianzheng Ma, Yihang Lou, Yan Bai, Renrui Zhang, Nian Shi, and Lin Luo. Decorate the newcomers: Visual domain prompt for continual test time adaptation. *arXiv preprint arXiv:2212.04145*, 2022. 1

[12] 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. 3, 6

[13] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems*, 35:29374–29385, 2022. 2

[14] 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. 8

[15] Peng Gao, Renrui Zhang, Rongyao Fang, Ziyi Lin, Hongyang Li, Hongsheng Li, and Qiao Yu. Mimic before reconstruct: Enhancing masked autoencoders with feature mimicking. *arXiv preprint arXiv:2303.05475*, 2023. 2, 4

[16] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Note: Robust continual test-time adaptation against temporal correlation. *Advances in Neural Information Processing Systems*, 35:27253–27266, 2022. 2, 3

[17] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzhi Li, and Pheng Ann Heng. Joint-mae: 2d-3d joint masked autoencoders for 3d point cloud pre-training. *IJCAI 2023*, 2023. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 3, 4

[21] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1, 5

[22] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and Sun-Yuan Kung. Milan: Masked image pretraining on language assisted representation. *arXiv preprint arXiv:2208.06049*, 2022. 2, 3, 4

[23] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. *arXiv preprint arXiv:2212.01322*, 2022. 6

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[27] Jogendra Nath Kundu, Naveen Venkat, Rahul M, and R. Venkatesh Babu. Universal source-free domain adaptation. 2020. 2

[28] Dong-Hyun Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. 2013. 6

[29] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao

Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. *arXiv preprint arXiv:2312.16217*, 2023. 1

[30] Jian Liang, D. Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020. 2

[31] 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. 2, 3, 4, 5, 6, 7, 8

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[33] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, pages 1150–1157. Ieee, 1999. 7

[34] Jiayi Ni, Senqiao Yang, Jiaming Liu, Xiaoqi Li, Wenyu Jiao, Ran Xu, Zehui Chen, Yi Liu, and Shanghang Zhang. Distribution-aware continual test time adaptation for semantic segmentation. *arXiv preprint arXiv:2309.13604*, 2023. 3

[35] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR, 2022. 2, 3

[36] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023. 2

[37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3

[38] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019. 4

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 3, 4

[40] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[43] Subhankar Roy, Martin Trapp, Andrea Pilzer, Juho Kannala, Nicu Sebe, Elisa Ricci, and Arno Solin. Uncertainty-guided source-free domain adaptation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*, pages 537–555. Springer, 2022. 4

[44] Sebastian Ruder and Barbara Plank. Learning to select data for transfer learning with bayesian optimization. *arXiv preprint arXiv:1707.05246*, 2017. 8

[45] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021. 1, 5, 7

[46] Junha Song, Jungsoo Lee, In So Kweon, and Sungha Choi. Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11920–11929, 2023. 1, 2, 3

[47] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 2, 4

[48] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 1, 2, 3, 7

[49] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 2, 6, 7

[50] 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. 1, 2, 4, 5, 6, 7

[51] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 2, 3, 4, 5

[52] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. In *European Conference on Computer Vision*, pages 337–353. Springer, 2022. 2, 3

[53] 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. 1, 5, 7

[54] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 3

[55] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. *international conference on computer vision*, 2021. 2

[56] Senqiao Yang, Jiaming Liu, Ray Zhang, Mingjie Pan, Zoey Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Yandong Guo, and Shanghang Zhang. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. *arXiv preprint arXiv:2312.14074*, 2023. 1

[57] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023. 1

[58] Senqiao Yang, Jiarui Wu, Jiaming Liu, Xiaoqi Li, Qizhe Zhang, Mingjie Pan, and Shanghang Zhang. Exploring sparse visual prompt for cross-domain semantic segmentation. *arXiv preprint arXiv:2303.09792*, 2023. 5, 7

[59] Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15922–15932, 2023. 2

[60] Zelin Zang, Lei Shang, Senqiao Yang, Fei Wang, Baigui Sun, Xuansong Xie, and Stan Z Li. Boosting novel category discovery over domains with soft contrastive learning and all in one classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11858–11867, 2023. 2

[61] Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. *CVPR 2023*, 2023. 3

[62] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1