Uniformity First: Uniformity-aware Test-time Adaptation of Vision-language Models against Sensor Degradation

Anonymous authors
Paper under double-blind review

Abstract

Pre-trained vision-language models, such as contrastive language-image pre-training (CLIP), have demonstrated a remarkable generalizability, enabling a wide range of applications, including zero-shot classification. However, vision-language models still struggle to handle *distribution shifts*, where input samples have large gaps from training ones. We found that CLIP is especially vulnerable to sensor degradation, a type of realistic distribution shift caused by sensor conditions such as weather, light, or noise. Collecting a new dataset from a test distribution for fine-tuning is highly costly since sensor degradation occurs unexpectedly and has a wide variety of types. Thus, we investigate *test-time adaptation (TTA)* of zero-shot classification, which enables on-the-fly adaptation to the test distribution with unlabeled test data. Existing TTA methods for CLIP mainly focus on modifying image and text embeddings or predictions to address distribution shifts. Although these methods can adapt to domain shifts, such as out-of-distribution or different renditions in input images, they fail to adapt to distribution shifts beyond domain shifts, e.g., sensor degradation. We found that *uniformity* of image embeddings, which is related to the amount of information, is a key factor that differentiates domain shifts and other distribution shifts. To enable adaptation on distribution shifts including sensor degradation, we propose a novel method called *un*iformity-aware *info*rmation-balanced TTA (UnInfo). To address distribution shifts, we introduce uniformity-aware confidence maximization, information-aware loss balancing, and knowledge distillation from the exponential moving average (EMA) teacher. Through experiments, we demonstrate that our UnInfo improves accuracy under sensor degradation by retaining information in terms of uniformity.

1 Introduction

Vision-language models (VLMs) pre-trained on large-scale datasets, such as contrastive language-image pre-training (CLIP) (Radford et al., 2021), have demonstrated remarkable generalizability and rich feature representations. Specifically, pre-trained VLMs have enabled various applications such as zero-shot transfer (Radford et al., 2021; Ge et al., 2023; Wang et al., 2023), image/video retrieval (Baldrati et al., 2022; Fang et al., 2021), and image generation (Patashnik et al., 2021; Ramesh et al., 2022). VLMs owe their success to their rich feature representations that unify vision and language modalities and public availability of the pre-trained weights, such as OpenCLIP (Ilharco et al., 2021; Cherti et al., 2023; Schuhmann et al., 2022). However, despite their generalizability, VLMs still face a challenge in adapting to distribution shifts, i.e., making predictions on test datasets with large gaps from the training dataset (Zhang et al., 2022b; Huang et al., 2024; Chen et al., 2023; Shu et al., 2022; Zhou et al., 2024; Karmanov et al., 2024; Zhang et al., 2024; Zanella & Ben Ayed, 2024b; Wang et al., 2024b; Qian & Hu, 2024).

A naive way of adapting VLMs is to collect a dataset from the test distribution and fine-tune the entire model parameters or the head classifier. However, labeled data from the test distribution may not be available because the distribution shift is unknown before deployment (Wang et al., 2021; Adachi et al., 2023).

To adapt to distribution shifts instantly after being deployed in the test distribution, *test-time adaptation (TTA)* (Liang et al., 2024; Dong et al., 2025), a paradigm aiming to adapt models during testing using

only unlabeled test data, has attracted attention. In the context of VLMs, recent works (Shu et al., 2022; Zhou et al., 2024; Karmanov et al., 2024; Zhang et al., 2024; Wang et al., 2024b; Qian & Hu, 2024; Zanella & Ben Ayed, 2024b) have intensively studied the TTA for zero-shot classification, which is one of the most common applications of VLMs. When the domain changes, typical text prompts can be suboptimal. For example, in an art or illustration domain, text prompts such as "`a photo of a [class name]`" is not appropriate (Shu et al., 2022; Zhou et al., 2022b;a). In other words, there is a modality gap between text prompts and images in the embedding space, which is crucial for VLMs' generalization (Liang et al., 2022; Khattak et al., 2023; Qian et al., 2024; Yamaguchi et al., 2025; Eslami & de Melo, 2025), under domain shifts. Thus, existing TTA methods for VLMs are mainly designed for domain shifts (also called natural distribution shifts), such as changes in rendition or out-of-distribution (OOD) (Hendrycks et al., 2021a; Recht et al., 2019; Hendrycks et al., 2021b; Wang et al., 2019) by modifying image and/or text embeddings during testing, which can be viewed as addressing the modality gap.

While existing TTA methods successfully adapt to the domain shifts, we found that they are prone to overfitting to domain shifts and degrade the performance on other types of distribution shifts, e.g., *sensor degradation* (Hendrycks & Dietterich, 2019; Sójka et al., 2023). When an image recognition system is deployed in the real world, the model faces various perturbations even in the same domain. This is because of changes in weather, light conditions, noise, cameras, etc., which are crucial in a wide range of applications, such as autonomous driving or surveillance cameras (Dai & Gool, 2018; Volk et al., 2019; Eastwood et al., 2022; Adachi et al., 2023; 2024; Enomoto et al., 2024). Such perturbations occur unexpectedly even within a single domain. In the existing literature on ordinary classification models, sensor degradation deteriorates the model's accuracy (Hendrycks & Dietterich, 2019; Qin et al., 2022). However, the TTA of VLMs against distribution shifts outside domain shifts has not been explored or evaluated in existing works.

To examine the robustness of VLMs against various types of distribution shifts including sensor degradation, we evaluated existing TTA methods for VLMs on image corruption (Hendrycks & Dietterich, 2019; Mintun et al., 2021) in terms of zero-shot classification performance using CLIP-family models. Through the experiment, we found that they significantly degrade the performance on corrupted images and that existing TTA methods can fail to improve performance. Moreover, we analyzed image embeddings and CLIP's knowledge about sensor degradation to reveal the difference between domain shift and sensor degradation. As a result, we experimentally found that sensor degradation also causes the modality gap between image and text embeddings, but the mechanism differs from domain shifts. Under sensor degradation, the modality gap occurs by image embeddings being "corrupted" in terms of *uniformity*, a measure related to the amount of input information retained in the embedding space (Wang & Isola, 2020). In other words, the amount of input information retained in the image embeddings becomes small by sensor degradation. Nevertheless, existing CLIP TTA methods for domain shifts address the modality gap by updating embedding vectors in a post-hoc manner without updating encoders, which implicitly assumes that embedding vectors are discriminative under domain shifts. However, under sensor degradation, image embeddings retain less information, i.e., they are less discriminative. This is the primary reason why existing CLIP TTA methods fail to maintain their performance on sensor degradation. Furthermore, we found that CLIP models cannot sufficiently encode words related to image quality; thus, simple prompting techniques, such as ensembles or incorporating corruption names into prompts, cannot recover the performance degradation (Sec. 3). From these observations, existing TTA methods or simple prompting techniques targeting domain shifts suffer from sensor degradation, highlighting the necessity of a novel TTA method suitable for a wider range of distribution shifts.

To enable the TTA of CLIP under various distribution shifts including sensor degradation, we propose a novel method called ***un****iformity-aware* ***info****rmation-balanced test-time adaptation (UnInfo).* UnInfo addresses the fundamental challenge of image embeddings' corruption by updating the image encoder with a low-rank adapter (LoRA) (Hu et al., 2022). To realize effective adaptation, UnInfo consists of three components: (i) uniformity-aware confidence maximization, (ii) information-aware loss balancing, and (iii) knowledge distillation from the exponential moving average (EMA) teacher. Uniformity-aware confidence maximization seeks to maximize prediction confidence in terms of entropy, while incorporating uniformity to prevent embeddings from losing input information. The information-aware loss balancing adaptively controls the balance between confidence maximization and uniformity enhancement on the basis of mutual information

so that uniformity is first leveraged and then confidence is addressed. This balancing plays a critical role, specifically when confidence is unreliable because of severe image embedding corruption. The knowledge distillation from the EMA teacher stabilizes the encoder update by tracking the EMA of LoRA parameters and regularizing the student's prediction to be close to the teacher's.

Through extensive experiments, our UnInfo improved the test zero-shot accuracy on various distribution shifts including sensor degradation by incorporating uniformity and balancing priority between uniformity and entropy. Also under domain shifts, we observed improvements by UnInfo, which validates the generality of our approach.

2 Zero-shot Classification with CLIP

Given a pre-trained CLIP composed of a text encoder $f^{\text{txt}}_{\theta_{\text{txt}}} : \mathcal{T} \to \mathbb{R}^d$ and image encoder $f^{\text{img}}_{\theta_{\text{img}}} : \mathcal{X} \to \mathbb{R}^d$, we first encode the text prompts to obtain text embeddings, which are used for the prototype of each class in the embedding space $\mathbb{R}^d$, where $\mathcal{T}$ and $\mathcal{X}$ are text and image input spaces, and θ_{txt} and θ_{img} are pre-trained weights of the encoders. The text prompts typically consist of a template and class names, like "`a photo of a [class name]`," denoted by $\mathbf{p}_c$ for class c. We denote the corresponding text embeddings as $\{\mathbf{t}_c = f^{\text{txt}}_{\theta_{\text{txt}}}(\mathbf{p}_c)\}_{c=1}^{C}$, where C is the total number of classes. We assume the text embeddings are normalized, i.e., $\|\mathbf{t}_c\|_2 = 1$.

For a test image $\mathbf{x} \in \mathcal{X}$, we compute the image embedding $\mathbf{z} = f^{\text{img}}_{\theta_{\text{img}}}(\mathbf{x})$, where $\|\mathbf{z}\|_2 = 1$. The similarity between the image and text embeddings $\mathbf{z}^\top \mathbf{t}_c$ is regarded as the logit for class c. The zero-shot prediction probability is obtained by

$$\hat{p}_c = \text{softmax}(\mathbf{z}^\top [\mathbf{t}_1, \ldots, \mathbf{t}_C]/\tau)_c, \tag{1}$$

where $\tau > 0$ is the temperature parameter. The final prediction is made by taking the argmax of $\hat{p}_c$.

3 Preliminary Experiment

First, we empirically demonstrate the vulnerability of CLIP under sensor degradation in terms of zero-shot classification accuracy. Moreover, CLIPs cannot properly encode images under such distribution shifts, i.e., the performance degradation cannot be recovered by simple prompting techniques because CLIPs cannot sufficiently represent concepts related to image quality.

Setup. We evaluated a ViT-B/16 CLIP pre-trained on the LAION dataset (Schuhmann et al., 2022) downloaded via OpenCLIP (Ilharco et al., 2021). We used the ImageNet-C (Hendrycks & Dietterich, 2019) dataset, which includes 15 types of image corruption simulating the sensor degradation (see Sec. 5.1 for dataset details). We performed zero-shot classification with the text prompt "`a photo of a [class name]`" (Normal prompt). We also used the ensemble of 80 text prompts, e.g., "`a bad photo of a [class name]`," "`a photo of many [class name]`," and so on, which is widely adopted as one baseline (Radford et al., 2021) (Ensemble), to see the effectiveness of prompt engineering on sensor degradation. To check the expressiveness of CLIP representation, we also examined text prompts that included descriptions of corruption. Specifically, we used the prompts "`a photo of a [class name] corrupted by [corruption name]`." For each corruption, we generated ten synonyms of the `corruption name` with GPT-4o (Hurst et al., 2024) and ensembled the text prompts (Corruption prompt). Details of the prompt ensemble are provided in B. We tested on ImageNet (Deng et al., 2009), ImageNet-A/R (Hendrycks et al., 2021b;a) (domain shifts), and ImageNet-C (Hendrycks & Dietterich, 2019) to observe the effect of sensor degradation.

Evaluation metrics. We evaluated accuracy, entropy, uniformity loss, and the modality gap between text and image embeddings by the earth mover's distance (EMD). The entropy measures the uncertainty of predictions (a lower value indicates high confidence), and the uniformity loss measures how image embeddings are uniformly distributed on the unit hypersphere, which is related to the amount of input information preserved in the image embedding (Oord et al., 2018) (a lower value indicates more information is preserved). The modality gap measures distributional distance between image and text embeddings, which is related to

Table 1: Zero-shot classification metrics of OpenCLIP ViT-B/16 with simple prompting techniques on ImageNet family and ImageNet-C. "Clean" corresponds to the ImageNet.

Metric	ImageNet	ImageNet-A	ImageNet-R	Domain shift Mean	Defocus blur	Glass blur	Motion blur	Zoom blur	Contrast	Elastic transform	Jpeg compression	Pixelate	Gaussian noise	Impulse noise	Shot noise	Brightness	Fog	Frost	Snow	Corruption Mean
Accuracy (Normal prompt, ↑)	66.97	32.91	74.36	58.08	28.18	11.81	19.13	17.65	17.90	13.37	36.77	36.92	6.02	6.12	7.88	54.75	34.39	27.32	27.77	23.06
Accuracy (Ensemble, ↑)	67.59	33.15	76.51	59.08	29.09	12.56	20.56	19.12	18.55	14.34	37.71	37.89	6.36	6.54	8.27	55.78	35.99	28.37	28.34	23.97
Accuracy (Corruption prompt, ↑)	-	-	-	-	26.53	12.15	18.39	17.99	17.65	13.96	36.29	36.35	6.02	6.10	7.80	53.93	33.43	27.34	26.55	22.70
Entropy (↓)	0.748	1.220	0.595	0.854	2.011	2.703	2.317	2.245	3.247	2.297	1.632	1.615	3.773	3.714	3.671	1.061	1.681	1.905	2.048	2.395
Uniformity loss (↓)	0.513	0.538	0.500	0.517	0.682	0.735	0.722	0.715	0.744	0.706	0.630	0.641	0.855	0.853	0.839	0.601	0.665	0.655	0.686	0.715
Modality gap (EMD, ↓)	1.291	1.333	1.348	1.324	1.298	1.326	1.325	1.327	1.337	1.341	1.293	1.296	1.299	1.297	1.296	1.293	1.307	1.315	1.308	1.311

Table 2: Zero-shot corruption type classification accuracy (%).

Defocus blur	Glass blur	Motion blur	Zoom blur	Contrast	Elastic transform	Jpeg compression	Pixelate	Gaussian noise	Impulse noise	Shot noise	Brightness	Fog	Frost	Snow	Total
68.76	11.12	75.61	2.61	0.36	18.91	1.28	8.64	30.28	0.23	0.70	3.43	78.95	49.16	7.32	23.53

the generalizability of CLIP (Liang et al., 2022; Khattak et al., 2023; Qian et al., 2024; Yamaguchi et al., 2025; Eslami & de Melo, 2025). The entropy and uniformity loss are defined as follows:

$$\text{Entropy} := \sum_{c=1}^{C} -\hat{p}_c \log \hat{p}_c, \tag{2}$$

$$\text{Uniformity loss} := \mathbb{E}_{\mathbf{z}_1, \mathbf{z}_2} \left[\exp(-\|\mathbf{z}_1 - \mathbf{z}_2\|_2^2) \right], \tag{3}$$

$$\text{Modality gap} := \text{EMD}(\{\mathbf{z}_1, \ldots, \mathbf{z}_N\}, \{\mathbf{t}_1, \ldots, \mathbf{t}_C\}). \tag{4}$$

Details of EMD are provided in Sec. C of the appendix.

Results. Tab. 1 shows the results. On the domain shifts, the prompt ensemble had 1%pt accuracy improvement on average compared to the normal prompt. On the other hand, ImageNet-C significantly degraded accuracy, and the ensemble did not result in significant improvement. Moreover, including corruption information in the prompt (Corruption prompt) resulted in accuracy degradation. Notably, the entropy and uniformity loss significantly increased compared to that of the domain shifts on all corruption types, while there is no significant difference in the modality gap. In other words, under sensor degradation, less information is preserved in the image embeddings, which makes the zero-shot prediction uncertain. Thus, sensor degradation, such as image corruption, can distort the semantical alignment between image and text embeddings, but the characteristic of the modality gap differs from that of the domain shift.

Next, to check whether the CLIP recognizes image corruption types, we performed zero-shot classification of the 15 corruption types of input images instead of object categories using the text prompts "`a photo corrupted by [corruption name]`." Tab. 2 shows the results. Most corruption types had poor accuracy, which suggests that CLIP cannot recognize image corruption types. Even in the cases of corruption types with high accuracies (such as defocus blur, motion blur, and fog), the contribution of including the corruption information for object classification is limited or, even worse, as shown in Tab. 1.

From these observations, sensor degradation causes the modality gap as well as domain shifts, but has fundamentally different properties from domain shifts in terms of the uniformity and entropy, which suggests that the CLIP cannot properly encode images corrupted by the sensor degradation.

4 Uniformity-aware Information-balanced Test-time Adaptation

We introduce our proposed method, *Uniformity-aware Information-balanced Test-time Adaptation (UnInfo)*. Fig. 1 illustrates the overview of UnInfo. The goal of TTA is to perform zero-shot classification on each incoming batch of test images $\{\mathbf{x}_i\}_{i=1}^{B}$ with updating the model parameters to make accurate predictions,

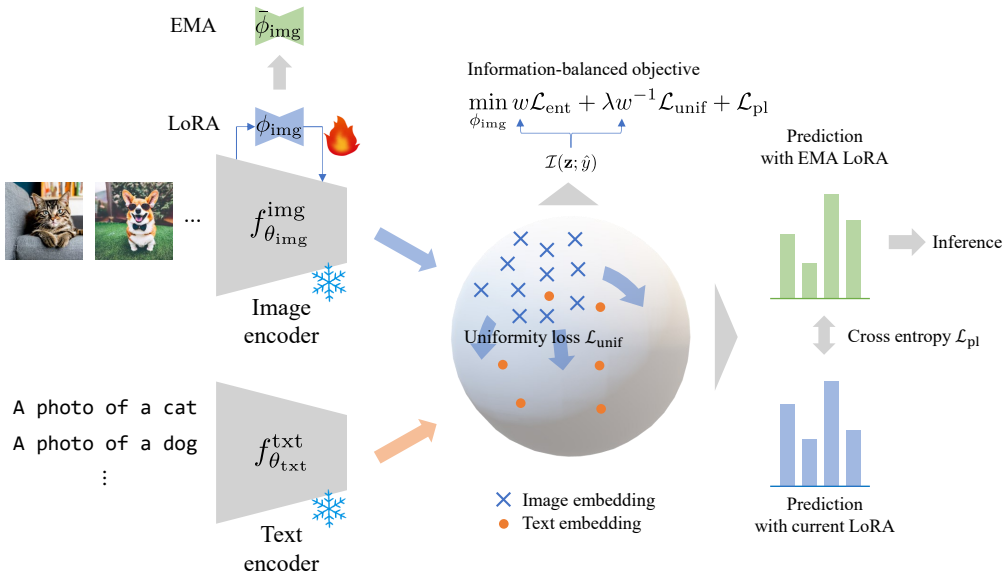


Figure 1: Overview of Uniformity-aware Information-balanced Test-time Adaptation (UnInfo).

where B is the batch size. Since distribution shifts occur on input images and text embeddings are fixed in zero-shot classification, we specifically update θ_{img} in our method.

4.1 Uniformity-aware Confidence Maximization

Our method's basic approach is to enhance the confidence of predictions via minimizing entropy, which is widely adopted in existing TTA methods for general classification models (Wang et al., 2021; Zhou & Levine, 2021; Niu et al., 2022; Adachi et al., 2023; Enomoto et al., 2024; Zhang et al., 2022a; Adachi et al., 2024). In zero-shot classification with CLIP, the entropy loss is computed by using the prediction probability $\hat{p}_c$ on the basis of similarity, defined in Eq. (1):

$$\mathcal{L}_{\text{ent}} = \frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{C} -\hat{p}_{i,c}\log\hat{p}_{i,c}. \tag{5}$$

Minimizing entropy is expected to improve the model performance because entropy is a promising proxy for classification accuracy when the images are appropriately embedded and input information is preserved, e.g., domain shifts. However, the amount of input information preserved in the image embeddings decreases under sensor degradation as the entropy and uniformity loss increase in Tab. 1, as discussed in Sec. 3. In such cases, the improvement by solely minimizing the entropy is limited since it attempts to leverage less information. In other words, images are less distinguishable in the embedding space. For enhancing information retained in image embeddings, we aim to improve the distinguishability of image embeddings from each other by minimizing the uniformity loss (Oord et al., 2018):

$$\mathcal{L}_{\text{unif}} = \log\frac{1}{B^2}\sum_{i=1}^{B}\sum_{j=1}^{B}\exp\left(-\|\mathbf{z}_i - \mathbf{z}_j\|_2^2\right). \tag{6}$$

We minimize the entropy and uniformity losses simultaneously for refining prediction confidence and making image embeddings distinguishable (i.e., enhancing retained information):

$$\mathcal{L} = \mathcal{L}_{\text{ent}} + \lambda\mathcal{L}_{\text{unif}}, \tag{7}$$

where $\lambda > 0$ is a hyperparameter, and $\mathbf{z}_i, \mathbf{z}_j$ are image embeddings of a batch of input images.

4.2 Information-aware Loss Balancing

As described in Secs. 3 and 4.1, minimizing the uniformity loss helps retain input information of image embeddings. However, we found that the importance of uniformity and entropy can dynamically change during TTA. For example, entropy should be prioritized when zero-shot classification works well, e.g., when sensor degradation is not severe. In fact, uniformity differs depends on corruption types, which is demonstrated in the preliminary experiment (Tab. 1). On the other hand, uniformity loss should be leveraged first, and then entropy should be addressed when zero-shot classification goes to a degenerated solution, e.g., classifying all images into a single class under severe sensor degradation.

To recognize the current regime and adaptively assign weights to the entropy and uniformity loss, we propose *information-aware loss balancing*. We employ the mutual information between the image embedding $\mathbf{z}$ and prediction $\hat{y}$, denoted by $\mathcal{I}(\mathbf{z}; \hat{y})$, to detect whether classification works without supervision. We assign the weights to the two losses as follows:

$$\mathcal{L} = w\mathcal{L}_{\text{ent}} + \lambda w^{-1}\mathcal{L}_{\text{unif}}, \quad w = \exp(\mathcal{I}(\mathbf{z}; \hat{y}) - \mathcal{I}_0), \tag{8}$$

where $\mathcal{I}_0$ is a hyperparameter to determine the threshold between the two regimes. The mutual information $\mathcal{I}(\mathbf{z}; \hat{y})$ is widely used in representation learning for measuring the quality of features and is computed as follows (Bridle et al., 1991; Krause et al., 2010; Shi & Sha, 2012; Hu et al., 2017):

$$\mathcal{I}(\mathbf{z}; \hat{y}) = \mathcal{H}(\hat{y}) - \mathcal{H}(\hat{y}|\mathbf{z}) = \sum_{c=1}^{C} -\bar{p}_c \log \bar{p}_c - \frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{C} -\hat{p}_{i,c} \log \hat{p}_{i,c}, \tag{9}$$

where $\mathcal{H}(\cdot)$ is entropy and $\bar{p}_c = (1/B)\sum_{i=1}^{B} \hat{p}_{i,c}$. Intuitively, $\mathcal{I}(\mathbf{z}; \hat{y})$ takes small values when the image embedding $\mathbf{z}$ is less diverse and predictions are not confident. In such a case, a larger weight is assigned to $\mathcal{L}_{\text{unif}}$ to make the image embeddings diverse and retain more information. In the opposite case, $\mathcal{L}_{\text{ent}}$ is leveraged to make predictions more confident. Note that gradients are not propagated to the weight w since the balance is not an objective to be optimized.

4.3 Update with Low-rank Adapters

Although we aim to update the image encoder $f^{\text{img}}_{\theta_{\text{img}}}$ to minimize the proposed loss in Eq. (8), updating the whole θ_{img} naively leads to catastrophic model forgetting (Lai et al., 2023; Vesdapunt et al., 2024). To avoid this, we fix the original pre-trained parameter θ_{img} and add the LoRA (Hu et al., 2022) to the linear weights in the attention layers, inspired by Zanella & Ben Ayed (2024a). Specifically, given an attention layer at the l-th layer

$$\mathbf{h}^{l+1} = \text{softmax}\left((W^l_{\text{Q}}\mathbf{h}^l)(W^l_{\text{K}}\mathbf{h}^l)^\top/\sqrt{d^l}\right) W^l_{\text{V}}\mathbf{h}^l, \tag{10}$$

we attach low-rank matrices to $W^l_{\cdot}$:

$$W^l_{\cdot} \to W^l_{\cdot} + A^l_{\cdot}B^l_{\cdot}, \tag{11}$$

where $A^l_{\cdot} \in \mathbb{R}^{d^l \times d_{\text{LoRA}}}$ and $B^l_{\cdot} \in \mathbb{R}^{d_{\text{LoRA}} \times d^l}$ ($d_{\text{LoRA}} < d^l$). We denote the LoRA parameters as ϕ_{img}.

4.4 Knowledge Distillation from EMA Teacher

For stabilizing the adaptation, we track the EMA of ϕ_{img} following previous works (Wang et al., 2022; Gao et al., 2022; Döbler et al., 2023; Wang et al., 2024a). That is, we update $\bar{\phi}_{\text{img}}$ in every iteration:

$$\bar{\phi}_{\text{img}} \leftarrow m\phi_{\text{img}} + (1-m)\bar{\phi}_{\text{img}}, \tag{12}$$

where $m \in (0, 1)$ is the momentum parameter. We adopt predictions made with $\bar{\phi}_{\text{img}}$ for inference. Using $\bar{\phi}_{\text{img}}$ as the teacher, we penalize the current LoRA parameters ϕ_{img} (student) to make predictions close to those of the teacher. Specifically, we take the cross entropy between the teacher and student outputs (Wang et al., 2022; Gao et al., 2022):

$$\mathcal{L}_{\text{pl}} = \frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{C} -\hat{q}_{i,c} \log \hat{p}_{i,c}, \tag{13}$$

where $\hat{p}_{i,c}$ is the student's output defined in Eq. (1), and $\hat{q}_{i,c}$ is the teacher's output computed in the same way with $\hat{p}_{i,c}$ but using the EMA parameter $\bar{\phi}_{\text{img}}$.

To sum up Secs. 4.1 to 4.4, our objective is as follows:

$$\min_{\phi_{\text{img}}} w\mathcal{L}_{\text{ent}} + \lambda w^{-1}\mathcal{L}_{\text{unif}} + \mathcal{L}_{\text{pl}}. \tag{14}$$

5 Experiment

We conducted experiments on TTA under datasets that include sensor degradation.

5.1 Datasets

ImageNet-C (Hendrycks & Dietterich, 2019): This dataset is constructed to evaluate the robustness of vision models. It consists of the corrupted version of the validation set of ImageNet (Deng et al., 2009). ImageNet-C includes 15 types of corruption, such as blur or digital noise. Each corruption type has five severity levels. We used the images corrupted at the highest severity level.

ImageNet-C-bar (Mintun et al., 2021): This dataset is constructed to evaluate the robustness of the vision models on a broader range of corruption types. Like ImageNet-C, it also consists of the corrupted version of the ImageNet validation set. ImageNet-C-bar includes 10 corruption types that are algorithmically selected to be dissimilar from ImageNet-C. We used the images corrupted at the highest severity level, as in the ImageNet-C case.

5.2 Implementation

We used the ViT-B/16 CLIP pre-trained on the DataComp-1B dataset (Gadre et al., 2024). We downloaded the pre-trained weights via the OpenCLIP official repository (Cherti et al., 2023)[1]. We set the temperature of the zero-shot prediction $\tau = 0.01$. For our UnInfo, we used the AdamW optimizer (Loshchilov & Hutter, 2019) with learning rate = 0.001, weight decay = 0.01, and batch size $B = 64$. We set the weight of the uniformity loss $\lambda = 1$ and the threshold of the information-aware loss balancing weight $\mathcal{I}_0 = 3$. We chose these hyperparameters by using a few corruption types in ImageNet-C and used the selected hyperparameters for the others as default. $\mathcal{I}_0$ was chosen on the basis of the mutual information computed on clean data (ImageNet). For the LoRA in UnInfo, we used the implementation provided by Zanella & Ben Ayed (2024a)[2]. We set the LoRA hyperparameters, such as α and the rank d_{LoRA}, to the default values and the momentum of EMA $m = 0.001$. The LoRA matrices $A^l_{\cdot}$ and $B^l_{\cdot}$ are initialized by the Kaiming uniform initialization (He et al., 2015) and zero, respectively, so that image embeddings are unaffected by random ϕ_{img} at the initial state.

5.3 Baseline

We compared our UnInfo with existing TTA methods for CLIP zero-shot classification and few-shot adaptation methods. For few-shot methods, we split $n \in \{1, 5, 10\}$ test samples per class, used them for adaptation, and then tested the model on the rest test samples.
No-adapt: Just performs zero-shot classification without adaptation.
LP (Linear probing): Trains a linear classifier head with few-shot labeled samples from the test set.
Tip-adapter (Zhang et al., 2022b): Modifies predictions by using cached features and logits of few-shot labeled samples from the test set.
TPT (Test-time prompt tuning) (Shu et al., 2022): Updates the text token embeddings corresponding to the words of a text template, e.g., "`a photo of a`," to minimize the marginal entropy over augmented views of an input image.
TDA (Training-free dynamic adapter) (Karmanov et al., 2024): Constructs positive and negative caches on

[1] `https://github.com/mlfoundations/open_clip`
[2] `https://github.com/MaxZanella/CLIP-LoRA`

Table 3: Test accuracy (%) on ImageNet-C. The numbers (1), (5), and (10) presented with the method names are the shot numbers per class n used for the few-shot adaptation methods.

Method	Defocus blur	Glass blur	Motion blur	Zoom blur	Contrast	Elastic transform	Jpeg compression	Pixelate	Gaussian noise	Impulse noise	Shot noise	Brightness	Fog	Frost	Snow	Mean
No-adapt	28.31	11.89	19.16	17.61	17.87	13.22	36.79	37.01	6.08	6.17	7.85	54.89	34.55	27.35	27.67	23.09
Linear probing (1)	11.53	6.15	8.79	9.02	5.87	8.17	15.57	16.36	2.79	2.76	3.41	27.05	16.43	9.88	11.87	10.38
Linear probing (5)	19.55	10.53	15.16	15.25	10.04	14.47	25.68	27.02	4.76	4.95	5.62	43.28	26.10	17.31	19.48	17.28
Linear probing (10)	23.84	12.72	17.97	18.57	12.45	17.69	30.50	32.32	5.78	6.04	6.83	49.90	31.15	21.49	23.63	20.73
Tip-adapter (Zhang et al., 2022b) (1)	19.00	9.09	13.92	14.04	9.53	12.60	26.98	27.19	4.06	4.02	5.15	44.92	25.99	18.10	19.28	16.92
Tip-adapter (5)	23.43	12.27	17.61	17.76	11.56	16.82	30.87	32.09	4.88	4.86	6.03	49.88	30.59	21.43	22.82	20.19
Tip-adapter (10)	26.11	13.99	19.85	20.23	12.99	19.26	33.03	34.60	5.52	5.82	6.82	52.40	33.15	24.01	24.83	22.17
TPT (Shu et al., 2022)	29.66	12.87	21.11	20.54	20.11	15.21	39.27	41.14	6.48	6.74	8.50	57.35	37.05	29.81	30.23	25.07
ZERO (Farina et al., 2024)	26.85	8.86	18.11	19.89	16.46	12.38	35.09	37.44	3.69	5.33	4.43	53.35	33.50	26.56	27.42	21.96
MTA (Zanella & Ben Ayed, 2024b)	27.79	11.29	19.25	18.88	21.18	13.92	37.23	38.95	2.41	2.87	2.96	53.56	34.32	28.02	28.66	22.75
TDA (Karmanov et al., 2024)	30.13	14.59	22.10	**21.09**	19.59	**17.15**	38.58	39.53	7.23	7.45	9.34	56.99	38.09	30.24	31.02	25.54
UnInfo (ours)	**31.51**	**16.76**	**23.47**	20.40	**22.81**	16.59	**42.03**	**42.38**	**7.56**	**10.60**	**11.36**	**57.75**	**39.16**	**31.65**	**32.40**	**27.10**

Table 4: Test accuracy (%) on ImageNet-C-bar. The numbers (1), (5), and (10) presented with the method names are the shot numbers per class n used for the few-shot adaptation methods.

Method	Blue noise sample	Brownish noise	Caustic refraction	Checkerboard cutout	Concentric sine waves	Inverse sparkles	Perlin noise	Plasma noise	Single frequency greyscale	Sparkles	Mean
No-adapt	21.87	46.66	39.55	45.51	10.05	19.57	51.21	20.69	16.10	50.00	32.12
Linear probing (1)	9.05	24.22	17.77	21.99	3.74	8.59	24.87	10.04	4.23	26.19	15.07
Linear probing (5)	15.45	38.05	29.12	35.58	6.89	14.28	39.24	15.43	6.82	40.55	24.14
Linear probing (10)	18.58	44.03	34.96	41.54	8.83	17.36	45.82	18.41	9.05	47.20	28.58
Tip-adapter (Zhang et al., 2022b) (1)	14.95	37.35	29.99	35.58	7.12	14.82	40.06	15.46	10.06	41.29	24.67
Tip-adapter (5)	18.41	43.08	34.86	40.79	8.37	17.58	45.58	18.15	10.24	46.44	28.35
Tip-adapter (10)	20.22	45.93	37.59	43.65	9.77	19.62	48.52	20.07	11.51	49.26	30.61
TPT (Shu et al., 2022)	24.63	49.86	42.53	46.36	10.96	21.93	54.31	23.16	17.43	52.20	34.34
ZERO (Farina et al., 2024)	25.65	45.60	41.14	45.30	10.59	22.79	49.76	20.42	**19.81**	45.81	32.69
MTA (Zanella & Ben Ayed, 2024b)	23.75	45.13	40.28	45.05	9.93	20.53	50.50	20.04	19.32	45.89	32.04
TDA (Karmanov et al., 2024)	24.53	50.30	42.90	49.85	**12.69**	22.56	53.89	24.75	17.77	**55.00**	35.42
UnInfo (ours)	**26.78**	**51.21**	**43.73**	**50.03**	12.49	**23.58**	**55.22**	**24.88**	19.67	53.74	**36.13**

the basis of prediction confidence on incoming test images and modifies predictions of subsequent inputs.
ZERO (Farina et al., 2024): Performs voting within predictions of augmented views of an input image.
MTA (MeanShift for test-time augmentation) (Zanella & Ben Ayed, 2024b): Selects reliable image embeddings among augmented views and modifies the image embedding.

5.4 Results

5.4.1 Adaptation Performance

We evaluated each corruption type's test classification accuracy on ImageNet-C and C-bar. We ran TTA three times with different random seeds for each method and corruption type, and report the mean score. Tabs. 3 and 4 show the results. Our UnInfo consistently surpasses the zero-shot baselines. Intriguingly, the baselines, even the few-shot methods that use labeled test samples for adaptation, sometimes underperformed No-adapt. This is because the baseline methods aim to refine text and/or image embeddings in a post-hoc manner during testing. While corrupted images are not appropriately encoded, as discussed in Sec. 3, the encoders themselves remain fixed. In contrast, UnInfo successfully adapts to image corruption by updating the image encoder with LoRA. Specifically, UnInfo significantly improved the accuracy on difficult corruption types in terms of uniformity listed in Tab. 1, e.g., blur and noise, where the amount of information retained in image embeddings is smaller than the other corruption types.

Table 5: Ablation of UnInfo on ImageNet-C.

Method	Defocus blur	Glass blur	Motion blur	Zoom blur	Contrast	Elastic transform	Jpeg compression	Pixelate	Gaussian noise	Impulse noise	Shot noise	Brightness	Fog	Frost	Snow	Mean
$\mathcal{L}_{\text{ent}}$	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
$\mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{pl}}$	30.63	15.56	22.61	19.54	21.83	15.99	41.03	40.82	3.31	6.68	8.06	57.17	37.78	30.84	31.29	25.54
$\mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{unif}} + \mathcal{L}_{\text{pl}}$	31.08	16.48	23.40	20.32	22.49	**16.67**	41.26	41.26	3.90	7.87	7.17	57.24	38.74	31.24	32.01	26.07
$\mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{unif}} + \mathcal{L}_{\text{pl}}$+Balancing (UnInfo)	**31.51**	**16.76**	**23.47**	**20.40**	**22.81**	16.59	**42.03**	**42.38**	**7.56**	**10.60**	**11.36**	**57.75**	**39.16**	**31.65**	**32.40**	**27.10**

Table 6: Ablation of UnInfo on ImageNet-C-bar.

Method	Blue noise sample	Brownish noise	Caustic refraction	Checkerboard cutout	Concentric sine waves	Inverse sparkles	Perlin noise	Plasma noise	Single frequency greyscale	Sparkles	Mean
$\mathcal{L}_{\text{ent}}$	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
$\mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{pl}}$	26.20	49.67	42.81	48.87	11.67	22.78	53.94	23.85	19.48	49.63	34.89
$\mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{unif}} + \mathcal{L}_{\text{pl}}$	26.75	50.28	43.14	49.59	12.08	23.36	54.47	24.27	**19.69**	**54.27**	35.79
$\mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{unif}} + \mathcal{L}_{\text{pl}}$+Balancing (UnInfo)	**26.78**	**51.21**	**43.73**	**50.03**	**12.49**	**23.58**	**55.22**	**24.88**	19.67	53.74	**36.13**

5.4.2 Ablation Study

Here, we examined the effect of each component in UnInfo: uniformity-aware confidence maximization, information-aware loss balancing, and knowledge distillation from the EMA teacher. Tabs. 5 and 6 show the results. Solely minimizing the entropy loss $\mathcal{L}_{\text{ent}}$ resulted in catastrophically poor accuracy in all cases. This is because image corruption affects uniformity, as observed in Sec. 3; entropy can assign unreasonably high confidence to wrong classes, resulting in overfitting quickly. In contrast, incorporating the knowledge distillation loss $\mathcal{L}_{\text{pl}}$ drastically improved the stability. The uniformity loss $\mathcal{L}_{\text{unif}}$ further improved accuracy compared to $\mathcal{L}_{\text{ent}} + \mathcal{L}_{\text{pl}}$. However, its improvement is sometimes marginal because it overlooks the balance between entropy and uniformity, as Sec. 4.2 describes. Thus, adding the balancing further improved accuracy by properly enhancing entropy or uniformity. Specifically, we observed significant improvements in difficult corruption types that produce high uniformity loss, such as noise corruption. In such cases, uniformity should be recovered first before minimizing entropy loss. UnInfo successfully controls the priority of the losses.

5.4.3 Sensitivity Analysis

Fig. 2 plots the sensitivity analysis of the hyperparameters of UnInfo. We changed the weight of the uniformity loss λ in Eq. (8) and the threshold of the mutual information $\mathcal{I}_0$ for the loss balancing. We ran UnInfo on the ImageNet-C corruptions and reported the average test accuracy. Increasing λ produces better accuracies, mainly in $0.0 \leq \lambda \leq 0.75$, suggesting the efficacy of the uniformity loss, and further increasing λ beyond 1.0 results in slightly higher accuracy. On the other hand, varying $\mathcal{I}_0$ hits the best accuracy when $\mathcal{I}_0 = 2.75$ but slightly affects the accuracy within $2.0 \leq \mathcal{I}_0 \leq 3.25$. However, increasing

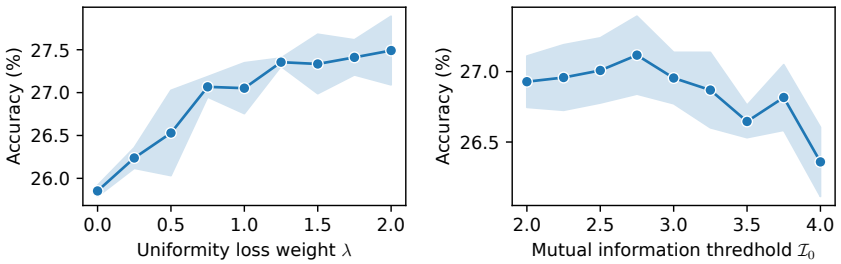


Figure 2: Sensitivity analysis on the uniformity loss weight λ (left), and the mutual information threshold $\mathcal{I}_0$ used in the loss balancing (right). The mean and standard deviation of the test accuracy calculated over the ImageNet-C corruptions are plotted.

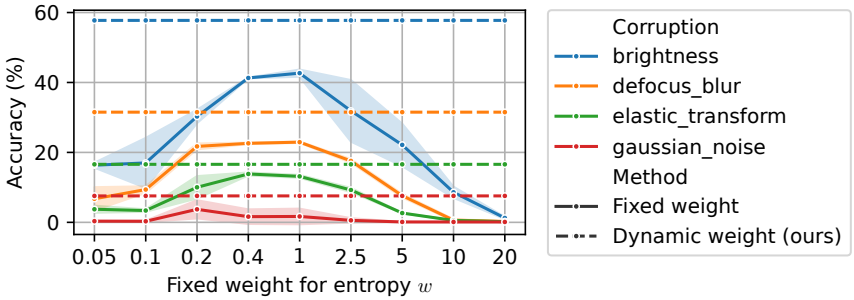
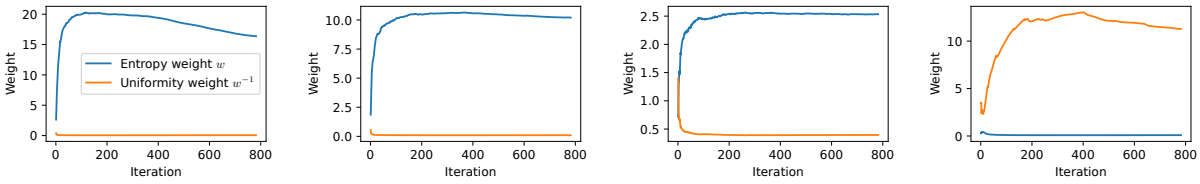


Figure 3: Sensitivity analysis on the balancing weight w with being fixed. The dashed lines represent accuracies when w is dynamically updated with our balancing mechanism described in Sec. 4.2.



(a) Brightness (b) Defocus blur (c) Elastic transform (d) Gaussian noise

Figure 4: Evolution of the information-aware loss balancing weights. The weights are adaptively assigned to the entropy and uniformity losses by the difficulty of distribution shifts. A larger weight is assigned to the entropy for easy distribution shifts such as brightness in (a). In contrast, the uniformity loss is prioritized for difficult distribution shifts such as Gaussian noise in (d).

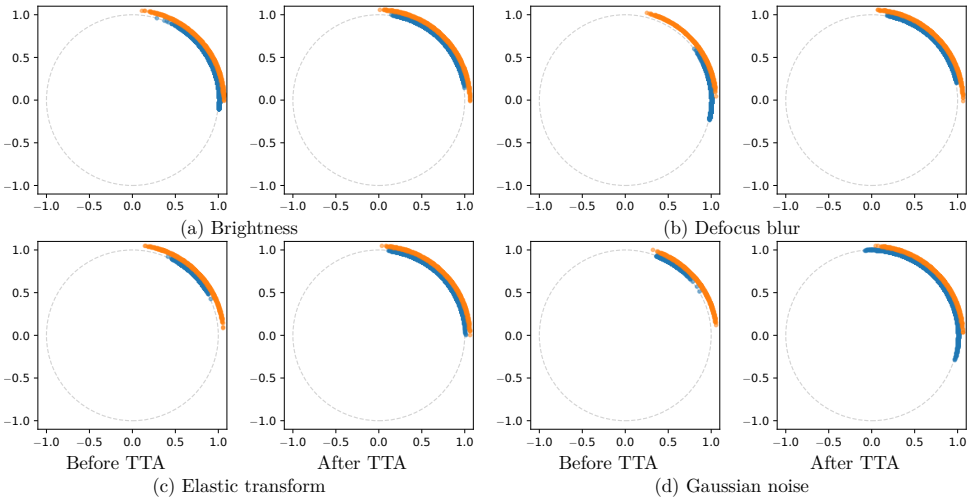
$\mathcal{I}_0$ too much deteriorates the accuracy because it controls the bias of the balance between entropy and uniformity. When $\mathcal{I}_0$ is too high, the uniformity loss is constantly overweighted, and the entropy is no longer optimized.

For checking the effectiveness of the dynamic weight by information-aware loss balancing, we fixed the weight w in Eq. (8). Fig. 3 shows the accuracy with fixed value of w being changed. Although leveraging both entropy and uniformity ($0.2 \leq w \leq 1$) produces higher accuracy regardless of corruption types, dynamic w had significant improvements compared to the best accuracies of fixed w.

5.4.4 Qualitative Analysis

Fig. 4 plots the evolution of the dynamic weights w and w^{-1} in Eq. (8). For easy image corruption types on which No-adapt produced relatively high accuracy, such as brightness in (a), the weight for the entropy w quickly increased, and the weight for uniformity loss w^{-1} was suppressed. This is because image embeddings under the brightness corruption retain information for classification; solely addressing entropy can improve accuracy. The defocus blur in (b) and elastic transform in (c) also showed similar evolutions in which the entropy quickly increases. However, the maximum value of the weight for entropy differs, and the weight for the uniformity loss is retained. This suggests the uniformity loss needs to be minimized along with entropy to retain information for these corruptions. On the other hand, the Gaussian noise in (d) showed a different evolution: the weight is larger for uniformity loss than for the entropy in the initial phase, indicating that retaining information of image embeddings is prioritized.

Next, we visualized embeddings to observe how the uniformity is improved. As the CLIP's embeddings are normalized and distributed on a unit hypersphere, we used the spherical PCA (Liu et al., 2019), which projects data points on a high-dimensional unit hypersphere onto a low-dimensional hypersphere (a 2D circle here). Fig. 5 visualizes the image embeddings before and after TTA with UnInfo on several corruptions of ImageNet-C, along with the text embeddings. After TTA, the image embeddings are distributed in a broader range of the circle than those before TTA on all corruptions, which suggests that the uniformity is



(a) Brightness (b) Defocus blur (c) Elastic transform (d) Gaussian noise

Figure 5: Spherical PCA (Liu et al., 2019) visualization of image (blue dots) and text (orange dots) embeddings before and after TTA with UnInfo on ImageNet-C. The image embeddings are distributed in a broader range of the circle after TTA, which suggests that the uniformity is improved. Moreover, the distribution of the image embeddings is aligned with the text embeddings, i.e., the image embeddings are more classification-friendly.

Table 7: Mean adaptation throughput and GPU memory usage.

Method	Throughput (images/sec.)	GPU memory (MiB)
No-adapt	$299.6_{\pm 1.0}$	1348
TPT (Shu et al., 2022)	$1.8_{\pm 0.0}$	14409
ZERO (Farina et al., 2024)	$4.2_{\pm 0.0}$	2319
MTA (Zanella & Ben Ayed, 2024b)	$1.3_{\pm 0.0}$	2331
TDA (Karmanov et al., 2024)	$44.6_{\pm 2.4}$	**1414**
UnInfo (ours)	$\mathbf{73.7}_{\pm \mathbf{0.5}}$	11736

improved. Moreover, the distributions of image and text embeddings are aligned after TTA, which suggests that the image embeddings become more classification-friendly. Uniformity improved most significantly in the Gaussian noise corruption in (d) because uniformity loss was prioritized by the information-aware loss balancing, as shown in Fig. 4 (d), as we intended. On the other hand, uniformity improved less significantly in the brightness corruption in (a) than in the other corruptions since the entropy is highly prioritized for the brightness corruption in Fig. 4 (a). In other words, uniformity is less important for this corruption because the image embeddings are already spread in a broader range than for the other corruptions, and their distribution is already aligned with that of the text embeddings before TTA.

5.4.5 Computational Efficiency

Tab. 7 shows each method's throughput (images per second) and GPU memory usage (MiB). The baseline methods based on marginal confidence over augmented views of an input (TPT (Shu et al., 2022), ZERO (Farina et al., 2024), and MTA (Zanella & Ben Ayed, 2024b)) had very low throughput because they needed to run forward passes for 64 augmented views per image. TPT and MTA had the lowest throughput because they require a further backward pass and solve an optimization problem for each image, respectively. Especially, TPT had the highest GPU memory usage because of backward pass for prompt optimization. In contrast, UnInfo had the highest throughput since it does not require data augmentation and only requires one forward and backward pass per image, while it had the second-highest GPU memory usage. Moreover, UnInfo can further speed up inference and save GPU memory as much as No-adapt because the knowledge of the test distribution is accumulated in the LoRA adapters; one may stop adaptation and merge LoRA to the stem model when the test distribution is stable. In contrast, TPT, ZERO, and MTA are episodic

methods, i.e., they do not update the model or accumulate any information. Thus, they always have to perform adaptation and inference together, unlike UnInfo.

6 Related Work

6.1 Contrastive Language-image Pre-training

CLIP (Radford et al., 2021) is a multimodal foundation model training paradigm, especially between image and text modalities. In CLIP, two encoders (an image encoder and a text encoder), are trained to map image and text inputs into a unified embedding space so that semantically corresponding inputs are mapped to close embeddings and vice versa. Although the training strategy is simple, CLIP has demonstrated remarkable generalization on downstream tasks by being trained on a huge dataset (e.g., hundreds of millions of image-text pairs). However, CLIP degrades downstream performance when faced with datasets with a large gap from the training dataset (Zhang et al., 2022b; Huang et al., 2024; Chen et al., 2023; Shu et al., 2022; Zhou et al., 2024; Karmanov et al., 2024; Zhang et al., 2024; Zanella & Ben Ayed, 2024b; Wang et al., 2024b; Qian & Hu, 2024). Re-training is often infeasible for adapting CLIP to a new dataset because it incurs a substantial computational cost, as described above. To address this challenge, TTA of CLIP has been actively studied.

6.2 Test-time Adaptation of Vision-language Models

For instantly adapting to test distributions without incurring heavy computational costs, TTA of zero-shot classification with CLIP has been studied. TTA aims to adapt a zero-shot CLIP classifier to the test distribution with only unlabeled test data. The representative approach of CLIP TTA is to update the image and/or text embeddings. Test-time prompt tuning methods (Shu et al., 2022; Yoon et al., 2024; Wang et al., 2025) updates the text token embeddings (e.g., four embedding vectors corresponding to words of a prompt template "`a photo of a`") during testing to minimize the prediction entropy marginalized with augmented views of an input image. The text embedding corresponding to each class is updated to be more appropriate to the current domain by updating the text token embeddings. Existing method also directly update the embeddings or logits computed from the similarity of image and text embeddings TDA (Karmanov et al., 2024) and DMN (Zhang et al., 2024) accumulate test inputs and construct the image embedding caches to modify subsequent inputs' predictions. MTA (Zanella & Ben Ayed, 2024b) selects reliable image embeddings and updates the feature centroid. OnZeta (Qian & Hu, 2024) and ZERO (Farina et al., 2024) dynamically correct the prediction probabilities for each test input.

These existing methods can adapt well to domain shifts, such as changes in environments, rendition, or out-of-distribution (Hendrycks et al., 2021a; Recht et al., 2019; Hendrycks et al., 2021b; Wang et al., 2019), using fixed image and text encoders. This is because a pre-trained CLIP generalizes to a wide range of domains enough to encode the current domain's semantics properly. However, these studies do not examine the ability to adapt to another type of distribution shift, sensor degradation. Moreover, we experimentally found that CLIP is vulnerable to this type of distribution shift, and existing methods fail to recover the performance.

7 Conclusion

We proposed UnInfo, a novel test-time adaptation (TTA) method for zero-shot classification with vision-language models under the sensor degradation. Unlike existing methods, UnInfo updates the image encoder to address the specific challenge of the sensor degradation, where loss input information is retained in the image embeddings unlike other natural distribution shifts. In the experiments, UnInfo achieved higher classification performance than baselines by refining the uniformity along with the entropy, and the information-aware loss balancing further improved the performance. One limitation of UnInfo is that it requires test data to be mini-batched. Our future work is to extend UnInfo to a fully online setting and broader types of distribution shifts.

References

Kazuki Adachi, Shin'ya Yamaguchi, and Atsutoshi Kumagai. Covariance-Aware Feature Alignment with Pre-Computed Source Statistics for Test-Time Adaptation to Multiple Image Corruptions. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 800–804, 2023. doi: 10.1109/ICIP49359.2023.10222901.

Kazuki Adachi, Shohei Enomoto, Taku Sasaki, and Shin'ya Yamaguchi. Test-time similarity modification for person re-identification toward temporal distribution shift. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2024. doi: 10.1109/IJCNN60899.2024.10650113.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21466–21474, 2022.

John Bridle, Anthony Heading, and David MacKay. Unsupervised classifiers, mutual information and 'phantom targets. *Advances in Neural Information Processing Systems (NIPS)*, 4:1096–1101, 1991.

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. PLOT: Prompt learning with optimal transport for vision-language models. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023.

Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 3819–3824, 2018. doi: 10.1109/ITSC.2018.8569387.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

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 (CVPR)*, pp. 7704–7714, June 2023.

Hao Dong, Lijun Sheng, Jian Liang, Ran He, Eleni Chatzi, and Olga Fink. Adapting vision-language models without labels: A comprehensive survey. *arXiv preprint arXiv:2508.05547*, 2025.

Cian Eastwood, Ian Mason, Chris Williams, and Bernhard Schölkopf. Source-Free Adaptation to Measurement Shift via Bottom-Up Feature Restoration. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*, 2022.

Shohei Enomoto, Naoya Hasegawa, Kazuki Adachi, Taku Sasaki, Shin'ya Yamaguchi, Satoshi Suzuki, and Takeharu Eda. Test-time adaptation meets image enhancement: Improving accuracy via uncertainty-aware logit switching. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2024. doi: 10.1109/IJCNN60899.2024.10650964.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in CLIP. In *Proceedings of The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021.

Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:129062–129093, 2024.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research (JMLR)*, 22(78):1–8, 2021.

Rémi Flamary, Cédric Vincent-Cuaz, Nicolas Courty, Alexandre Gramfort, Oleksii Kachaiev, Huy Quang Tran, Laurène David, Clément Bonet, Nathan Cassereau, Théo Gnassounou, Eloi Tanguy, Julie Delon, Antoine Collas, Sonia Mazelet, Laetitia Chapel, Tanguy Kerdoncuff, Xizheng Yu, Matthew Feickert, Paul Krzakala, Tianlin Liu, and Eduardo Fernandes Montesuma. Pot python optimal transport (version 0.9.5), 2024. URL `https://github.com/PythonOT/POT`.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

Yunhe Gao, Xingjian Shi, Yi Zhu, Hao Wang, Zhiqiang Tang, Xiong Zhou, Mu Li, and Dimitris N Metaxas. Visual prompt tuning for test-time domain adaptation. *arXiv preprint arXiv:2210.04831*, 2022.

Yunhao Ge, Jie Ren, Andrew Gallagher, Yuxiao Wang, Ming-Hsuan Yang, Hartwig Adam, Laurent Itti, Balaji Lakshminarayanan, and Jiaping Zhao. Improving zero-shot generalization and robustness of multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11093–11101, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034, December 2015.

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *Proceedings of the Seventh International Conference on Learning Representations (ICLR)*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349, October 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural Adversarial Examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15262–15271, June 2021b.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*, 2022.

Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *Proceedings of the Thirty-Fourth International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1558–1567. PMLR, 06–11 Aug 2017.

Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. Lp++: A surprisingly strong linear probe for few-shot clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23773–23782, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, July 2021.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient Test-Time Adaptation of Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14162–14171, 2024.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19113–19122, 2023.

Andreas Krause, Pietro Perona, and Ryan Gomes. Discriminative clustering by regularized information maximization. *Advances in Neural Information Processing Systems (NIPS)*, 23:775–783, 2010.

Zhengfeng Lai, Noranart Vesdapunt, Ning Zhou, Jun Wu, Cong Phuoc Huynh, Xuelu Li, Kah Kuen Fu, and Chen-Nee Chuah. Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16155–16165, October 2023.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision (IJCV)*, pp. 1–34, 2024.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:17612–17625, 2022.

Kai Liu, Qiuwei Li, Hua Wang, and Gongguo Tang. Spherical principal component analysis. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pp. 387–395. SIAM, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the Seventh International Conference on Learning Representations (ICLR)*, 2019.

Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3571–3583, 2021.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the Thirty-Ninth International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905. PMLR, 17–23 Jul 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2085–2094, 2021.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Qi Qian and Juhua Hu. Online zero-shot classification with clip. In *Proceedings of the Eighteenth European Conference on Computer Vision (ECCV)*, pp. 462–477, 2024.

Qi Qian, Yuanhong Xu, and Juhua Hu. Intra-modal proxy learning for zero-shot visual categorization with clip. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:25461–25474, 2024.

Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, and Xuezhi Wang. Understanding and improving robustness of vision transformers through patch-based negative augmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:16276–16289, 2022.

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. In *Proceedings of the Thirty-Eighth International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the Thirty-Sixth International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 09–15 Jun 2019.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, pp. 25278–25294, 2022.

Yuan Shi and Fei Sha. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In *Proceedings of the Twenty-Ninth International Conference on Machine Learning (ICML)*, pp. 1275–1282, 2012.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:14274–14289, 2022.

Damian Sójka, Sebastian Cygert, Bartłomiej Twardowski, and Tomasz Trzciński. Ar-tta: A simple method for real-world continual test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3491–3495, 2023.

Noranart Vesdapunt, Kah Kuen Fu, Yue Wu, Xu Zhang, and Pradeep Natarajan. HVCLIP: High-dimensional vector in CLIP for unsupervised domain adaptation. In *Proceedings of the Eighteenth European Conference on Computer Vision (ECCV)*, pp. 36–54, 2024.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

Georg Volk, Stefan Müller, Alexander von Bernuth, Dennis Hospach, and Oliver Bringmann. Towards robust cnn-based object detection through augmentation with synthetic rain variations. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 285–292, 2019. doi: 10.1109/ITSC.2019.8917269.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *Proceedings of the Ninth International Conference on Learning Representations (ICLR)*, 2021.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning Robust Global Representations by Penalizing Local Predictive Power. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 10506–10518. 2019.

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 (CVPR)*, pp. 7201–7211, June 2022.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the Thirty-Seventh International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 13–18 Jul 2020.

Xin Wang, Kai Chen, Jiaming Zhang, Jingjing Chen, and Xingjun Ma. Tapt: Test-time adversarial prompt tuning for robust inference in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19910–19920, June 2025.

Yanshuo Wang, Jie Hong, Ali Cheraghian, Shafin Rahman, David Ahmedt-Aristizabal, Lars Petersson, and Mehrtash Harandi. Continual test-time domain adaptation via dynamic sample selection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1701–1710, January 2024a.

Zhengbo Wang, Jian Liang, Ran He, Nan Xu, Zilei Wang, and Tieniu Tan. Improving zero-shot generalization for clip with synthesized prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3032–3042, 2023.

Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free CLIP-based adaptation. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024b.

Shin'ya Yamaguchi, Dewei Feng, Sekitoshi Kanai, Kazuki Adachi, and Daiki Chijiwa. Post-pre-training for Modality Alignment in Vision-Language Foundation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4256–4266, June 2025.

Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D Yoo. C-TPT: Calibrated Test-Time Prompt Tuning for Vision-Language Models via Text Feature Dispersion. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1593–1603, June 2024a.

Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23783–23793, 2024b.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:38629–38642, 2022a.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free Adaption of CLIP for Few-shot Classification. In *Proceedings of the Seventeenth European Conference on Computer Vision (ECCV)*, pp. 493–510, 2022b.

Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28718–28728, 2024.

Aurick Zhou and Sergey Levine. Bayesian Adaptation for Covariate Shift. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:914–927, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16816–16825, June 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 130(9):2337–2348, 2022b.

Yifei Zhou, Juntao Ren, Fengyu Li, Ramin Zabih, and Ser Nam Lim. Test-time distribution normalization for contrastively learned visual-language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

Algorithm 1 Procedure of UnInfo.

Input: Pre-trained image encoder $f^{\text{img}}_{\theta_{\text{img}}}$, class text embeddings $\{\mathbf{t}_c\}_{c=1}^{C}$, initialized LoRA ϕ_{img}, target dataset $\mathcal{D}$

Output: Adapted LoRA parameter (EMA) $\bar{\phi}_{\text{img}}$

for mini-batch $\{\mathbf{x}_i\}_{i=1}^{B}$ in $\mathcal{D}$ **do**
  Compute image embeddings with the current LoRA $\{\mathbf{z}_i = f^{\text{img}}_{\theta_{\text{img}},\phi_{\text{img}}}(\mathbf{x}_i)\}_{i=1}^{B}$.
  Compute zero-shot prediction probabilities $\{\hat{p}_{i,c}\}_{i=1}^{B}$ in accordance with Eq. (1).
  Compute teacher zero-shot prediction probabilities $\{\hat{q}_{i,c}\}_{i=1}^{B}$ using the EMA LoRA $\bar{\phi}_{\text{img}}$.
  Compute the loss in accordance with Eq. (14).
  Update ϕ_{img}.
  Update EMA teacher $\bar{\phi}_{\text{img}}$ in accordance with Eq. (12).
end for

Table 8: Text prompt templates used for ensemble in the preliminary experiment (Sec. 3). These templates are proposed by Radford et al. (2021)[4].

a bad photo of a {}.	a photo of many {}.	a sculpture of a {}.	a photo of the hard to see {}.
a low resolution photo of the {}.	a rendering of a {}.	graffiti of a {}.	a bad photo of the {}.
a cropped photo of the {}.	a tattoo of a {}.	the embroidered {}.	a photo of a hard to see {}.
a bright photo of a {}.	a photo of a clean {}.	a photo of a dirty {}.	a dark photo of the {}.
a drawing of a {}.	a photo of my {}.	the plastic {}.	a photo of the cool {}.
a close-up photo of a {}.	a black and white photo of the {}.	a painting of the {}.	a painting of a {}.
a pixelated photo of the {}.	a sculpture of the {}.	a bright photo of the {}.	a cropped photo of a {}.
a plastic {}.	a photo of the dirty {}.	a jpeg corrupted photo of a {}.	a blurry photo of the {}.
a photo of the {}.	a good photo of the {}.	a rendering of the {}.	a {} in a video game.
a photo of one {}.	a doodle of a {}.	a close-up photo of the {}.	a photo of a {}.
the origami {}.	the {} in a video game.	a sketch of a {}.	a doodle of the {}.
a origami {}.	a low resolution photo of a {}.	the toy {}.	a rendition of the {}.
a photo of the clean {}.	a photo of a large {}.	a rendition of a {}.	a photo of a nice {}.
a photo of a weird {}.	a blurry photo of a {}.	a cartoon {}.	art of a {}.
a sketch of the {}.	a embroidered {}.	a pixelated photo of a {}.	itap of the {}.
a jpeg corrupted photo of the {}.	a good photo of a {}.	a plushie {}.	a photo of the nice {}.
a photo of the small {}.	a photo of the weird {}.	the cartoon {}.	art of the {}.
a drawing of the {}.	a photo of the large {}.	a black and white photo of a {}.	the plushie {}.
a dark photo of a {}.	itap of a {}.	graffiti of the {}.	a toy {}.
itap of my {}.	a photo of a cool {}.	a photo of a small {}.	a tattoo of the {}.

A Details of UnInfo

Algorithm 1 lists the procedure of UnInfo.

B Text Prompt Ensemble

Here, we describe the details of the prompt ensemble in the preliminary experiment (Sec. 3). We used template texts listed in Tab. 8 for the ensemble of multiple templates (denoted by "Ensemble" in Tab. 1). We generated the prompts using the templates for each class and encoded them with the text encoder. Then, we calculated the mean of the text embeddings. We normalized the embedding and used it for the class prototype.

For the ensemble of corruption synonyms (denoted by "Corruption prompt" in Tab. 1), we ensembled the corruption synonyms listed in Tab. 9, which are generated by GPT-4o (Hurst et al., 2024) with the instruction "`You are an expert of image processing. List the synonyms of the word "[corruption name]," which represents image quality.`"

C Earth Mover's Distance

In Sec. 3, we evaluated the modality gap between image and text embeddings using the earth mover's distance (EMD). Here, we describe the details of EMD (Villani et al., 2008; Peyré et al., 2019).

[4] `https://github.com/openai/CLIP/blob/main/notebooks/Prompt_Engineering_for_ImageNet.ipynb`

Table 9: Synonyms of the corruption names of ImageNet-C used in the preliminary experiment in Sec. 3.

Defocus blur	Glass blur	Motion blur	Zoom blur	Contrast
defocus blur	glass blur	motion blur	zoom blur	contrast
out-of-focus blur	frosted blur	directional blur	radial blur	tonal contrast
soft focus	glazing blur	linear blur	zooming effect	brightness difference
bokeh	diffuse blur	dynamic blur	dynamic zoom blur	clarity
lens blur	smudged blur	streaking	burst blur	definition
gaussian blur	hazy blur	trail blur	focus expansion blur	distinction
depth blur	translucent blur	speed blur	depth blur	sharpness
background blur	refractive blur	panning blur	lens zoom blur	intensity diffenrece
field blur	distortion blur	motion streak	outward motion blur	dynamic range
focus softness	veiled blur	kinetic blur	radian streak blur	separation
Elastic transform	**Jpeg compression**	**Pixelate**	**Gaussian noise**	**Impulse noise**
elastic transform	jpeg compression	pixelate	Gaussian noise	impulse noise
warping	image compression	blockify	normal noise	salt-and-pepper noise
distortion	lossy compression	rasterize	additive noise	spiky noise
deformation	JPEG encoding	mosaic	white Gaussian noise	outlier noise
stretching	file compression	chunkify	statistical noise	random noise
bending	quantization artifacting	grid effect	random noise	shot noise
geometric transform	data compression	quantization	luminance noise	transitional noise
morphing	image encoding	low-resolution effect	stochastic interference	burst noise
image warping	compression artifacts	bitmapping	signal perturbation	pulsed noise
spatial transform	JPEG artifacts	aliased effect	normal distribution noise	point noise
Shot noise	**Brightness**	**Fog**	**Frost**	**Snow**
shot noise	brightness	fog	frost	snow
photon noise	luminance	haze	frosting	noise
Poisson noise	illumination	mist	glare	grain
quantum noise	lightness	obscuration	haze	salt-and-pepper noise
statistical noise	intensity	cloudiness	mist	static
random noise	radiance	smog	veiling	visual noise
electronic noise	glow	blur	soft-focus	pixel noise
counting noise	shininess	glare	diffusion	random noise
current noise	exposure	veiling	cloudiness	white noise
flicker noise	highlighting	dimming	blur	dither

EMD is computed as the minimum cost of transporting image embeddings $\{\mathbf{z}_1, \dots, \mathbf{z}_N\}$ to match text embeddings $\{\mathbf{t}_1, \dots, \mathbf{t}_C\}$. More formally, we computed EMD between two distributions $p(\mathbf{z}) = (1/N)\sum_{n=1}^{N} \delta(\mathbf{z}-\mathbf{z}_n)$ and $q(\mathbf{z}) = (1/C)\sum_{c=1}^{C} \delta(\mathbf{z}-\mathbf{t}_c)$ over $\mathbb{R}^d$, where $\delta(\cdot)$ is the Dirac function. Specifically, since the embedding vectors are normalized and on the unit hypersphere, we adopted the spherical distance $d(\mathbf{z}_n, \mathbf{t}_c) := \arccos(\mathbf{z}_n^\top \mathbf{t}_c)$ for the transportation cost. We implemented the computation with the Python Optimal Transport (POT) library (Flamary et al., 2021; 2024).

D Additional Experimental Results

D.1 Adaptation Performance

Other CLIP architectures. Similar to Sec. 5.4.1, we conducted experiments with OpenAI ViT-B/32 CLIP[5] and SigLIP (Zhai et al., 2023). Tabs. 10 to 13 show the results. The trend aligns with the cases of the ViT-B/16 CLIP in Sec. 5.4.1. UnInfo outperformed the baselines on most corruption types.

Domain Shifts. We also tested TTA baselines and our UnInfo on domain shift datasets: ImageNet (Deng et al., 2009), ImageNet-A (Hendrycks et al., 2021b), and R (Hendrycks et al., 2021a). Tab. 14 shows the results. Although UnInfo did not outperform the existing TTA baselines since they are specifically designed for domain shifts, UnInfo had slight improvements and no negative effects compared to No-adapt. This implies that the applicability of our method is not limited within sensor degradation by adaptively controlling the balance between entropy and uniformity. Moreover, one may incorporate UnInfo with other methods.

[5] https://github.com/openai/CLIP

Table 10: Test accuracy (%) of OpenAI ViT-B/32 on ImageNet-C. The numbers (1), (5), and (10) presented with the method names are the shot numbers per class n used for the few-shot adaptation methods.

Method	Defocus blur	Glass blur	Motion blur	Zoom blur	Contrast	Elastic transform	Jpeg compression	Pixelate	Gaussian noise	Impulse noise	Shot noise	Brightness	Fog	Frost	Snow	Mean
No-adapt	22.31	11.39	20.07	17.50	17.30	18.13	29.36	30.24	13.01	13.54	13.14	48.38	28.40	24.84	23.81	22.09
Linear probing (1)	7.34	3.98	6.93	6.25	5.60	7.53	10.39	11.38	5.05	5.03	5.24	20.45	10.80	8.10	7.59	8.11
Linear probing (5)	12.27	6.91	11.11	9.82	8.93	12.29	16.34	18.54	7.63	7.88	8.22	31.84	17.36	13.49	12.55	13.01
Linear probing (10)	14.52	8.41	13.08	11.73	10.55	15.38	19.38	22.34	9.55	9.44	9.72	37.52	20.71	16.38	15.21	15.59
Tip-adapter (Zhang et al., 2022b) (1)	10.40	5.67	9.49	8.94	7.45	11.22	15.07	16.36	6.53	6.73	6.97	29.89	15.36	11.98	10.97	11.53
Tip-adapter (5)	14.07	7.94	12.34	11.24	9.39	14.91	18.77	21.68	8.72	9.08	9.12	36.48	19.45	15.83	14.00	14.87
Tip-adapter (10)	16.14	9.40	14.30	12.69	11.23	17.21	21.25	24.76	9.92	10.47	10.48	40.20	22.15	18.04	15.93	16.94
TPT (Shu et al., 2022)	23.29	12.16	21.01	19.62	18.62	19.50	31.74	32.57	12.81	13.24	13.13	50.47	30.06	26.89	25.68	23.39
ZERO (Farina et al., 2024)	21.48	10.17	18.93	20.12	14.33	17.27	30.04	31.22	9.19	9.80	9.31	48.44	29.21	25.33	25.02	21.32
MTA (Zanella & Ben Ayed, 2024b)	20.76	11.19	18.60	18.56	19.42	19.29	30.25	32.09	10.86	11.40	11.43	49.33	29.33	26.22	24.65	22.23
TDA (Karmanov et al., 2024)	23.93	12.97	22.29	19.49	18.22	**21.10**	31.41	32.89	14.65	15.32	15.24	50.49	31.66	27.33	26.02	24.20
UnInfo (ours)	**26.80**	**14.77**	**26.10**	**21.41**	**21.51**	20.94	**34.73**	**36.87**	**16.71**	**18.88**	**17.14**	**51.16**	**33.09**	**27.44**	**27.39**	**26.33**

Table 11: Test accuracy (%) of SigLIP on ImageNet-C. The numbers (1), (5), and (10) presented with the method names are the shot numbers per class n used for the few-shot adaptation methods.

Method	Defocus blur	Glass blur	Motion blur	Zoom blur	Contrast	Elastic transform	Jpeg compression	Pixelate	Gaussian noise	Impulse noise	Shot noise	Brightness	Fog	Frost	Snow	Mean
No-adapt	26.49	12.33	16.60	20.38	16.48	16.09	36.61	45.29	6.18	8.09	8.03	62.21	42.49	32.75	31.94	25.46
Linear probing (1)	9.79	5.80	6.56	9.75	6.76	8.33	14.46	20.78	3.13	3.23	3.34	31.11	19.56	10.68	13.48	11.12
Linear probing (5)	18.15	9.83	12.28	15.72	10.50	14.17	23.23	32.55	1.61	0.10	2.26	49.34	30.55	18.31	21.20	17.32
Linear probing (10)	21.64	12.08	14.54	18.97	12.62	17.40	27.96	37.78	0.10	0.10	2.85	55.23	35.82	22.93	24.72	20.31
Tip-adapter (Zhang et al., 2022b) (1)	15.16	7.86	10.10	13.73	9.25	11.93	21.88	29.03	4.49	5.51	5.53	44.67	27.45	16.71	18.62	16.13
Tip-adapter (5)	20.04	10.99	13.81	17.90	11.88	16.26	26.36	36.26	5.72	7.01	6.95	52.24	33.39	20.28	23.22	20.15
Tip-adapter (10)	22.86	12.91	15.80	20.27	13.44	18.60	29.41	39.62	6.47	7.88	8.06	55.65	36.67	23.37	25.94	22.46
TPT (Shu et al., 2022)	11.32	4.87	6.56	8.94	22.77	2.68	9.08	18.02	0.90	0.97	0.88	2.81	33.48	3.44	0.44	8.48
ZERO (Farina et al., 2024)	10.20	3.74	5.63	9.09	21.11	3.65	8.10	19.41	0.81	0.89	0.69	2.65	36.91	3.86	0.36	8.47
MTA (Zanella & Ben Ayed, 2024b)	11.35	5.04	6.27	9.04	**26.00**	4.33	11.18	21.69	1.24	1.26	1.13	5.10	38.29	4.93	0.70	9.84
TDA (Karmanov et al., 2024)	29.07	14.75	19.09	23.57	18.48	**19.95**	38.35	47.45	7.87	9.76	**9.71**	64.13	45.62	35.45	34.81	27.87
UnInfo (ours)	**30.82**	**15.28**	**20.41**	**24.48**	20.80	19.89	**41.73**	**49.98**	**9.72**	**11.41**	9.09	**64.96**	**48.22**	**36.73**	**37.11**	**29.38**

Table 12: Test accuracy (%) of OpenAI ViT-B/32 on ImageNet-C-bar. The numbers (1), (5), and (10) presented with the method names are the shot numbers per class n used for the few-shot adaptation methods.

Method	Blue noise sample	Brownish noise	Caustic refraction	Checkerboard cutout	Concentric sine waves	Inverse sparkles	Perlin noise	Plasma noise	Single frequency greyscale	Sparkles	Mean
No-adapt	29.44	40.46	32.98	35.76	10.59	14.09	44.83	16.58	26.53	44.70	29.60
Linear probing (1)	11.85	17.61	12.22	14.22	3.17	4.76	17.89	6.70	6.94	19.51	11.49
Linear probing (5)	18.91	26.94	19.19	23.10	5.08	7.71	27.78	9.97	10.80	30.39	17.99
Linear probing (10)	22.96	30.90	23.29	27.44	6.63	9.32	33.01	11.52	13.51	35.42	21.40
Tip-adapter (Zhang et al., 2022b) (1)	17.16	24.37	18.03	20.65	4.94	6.83	26.05	9.00	11.84	27.31	16.62
Tip-adapter (5)	21.75	30.07	22.59	26.24	6.02	8.96	32.16	11.05	13.95	33.69	20.65
Tip-adapter (10)	24.19	32.85	25.15	29.52	7.10	10.42	35.13	12.59	16.01	36.78	22.97
TPT (Shu et al., 2022)	30.43	42.77	35.87	38.12	11.32	16.35	46.75	18.03	28.87	47.69	31.62
ZERO (Farina et al., 2024)	23.87	40.90	**36.66**	39.51	9.68	19.63	44.81	17.72	28.03	46.27	30.71
MTA (Zanella & Ben Ayed, 2024b)	28.94	40.85	35.61	37.84	11.17	16.68	45.65	16.61	29.65	46.46	30.95
TDA (Karmanov et al., 2024)	33.49	43.47	35.45	39.97	**12.42**	15.61	46.94	**18.96**	27.94	**48.50**	32.27
UnInfo (ours)	**35.88**	**43.86**	36.34	**40.03**	12.34	**17.09**	**48.46**	18.82	**29.67**	48.03	**33.05**

Table 13: Test accuracy (%) of SigLIP on ImageNet-C-bar. The numbers (1), (5), and (10) presented with the method names are the shot numbers per class n used for the few-shot adaptation methods.

Method	Blue noise sample	Brownish noise	Caustic refraction	Checkerboard cutout	Concentric sine waves	Inverse sparkles	Perlin noise	Plasma noise	Single frequency greyscale	Sparkles	Mean
No-adapt	32.20	55.95	46.04	56.68	13.33	24.72	59.48	27.46	21.30	60.34	39.75
Linear probing (1)	14.39	28.10	20.19	26.71	4.46	9.99	29.67	10.57	6.95	30.44	18.15
Linear probing (5)	22.61	43.51	32.23	43.69	7.74	16.64	45.60	17.11	11.80	47.98	28.89
Linear probing (10)	26.64	49.88	38.19	49.51	9.89	19.82	51.88	20.63	15.22	54.09	33.57
Tip-adapter (Zhang et al., 2022b) (1)	19.91	39.06	29.38	39.06	6.93	14.89	41.90	15.63	11.58	42.54	26.09
Tip-adapter (5)	25.00	46.29	35.35	46.78	8.57	18.50	48.91	18.72	13.56	50.05	31.17
Tip-adapter (10)	27.78	49.58	38.62	50.29	10.08	20.43	52.48	20.86	15.78	53.58	33.95
TPT (Shu et al., 2022)	4.86	17.57	16.77	22.19	2.40	0.66	37.78	8.47	2.23	31.88	14.48
ZERO (Farina et al., 2024)	6.32	19.31	22.32	27.12	2.09	4.64	43.11	8.54	5.53	35.37	17.44
MTA (Zanella & Ben Ayed, 2024b)	8.42	20.05	21.69	27.69	2.29	2.12	44.93	8.62	6.92	36.89	17.96
TDA (Karmanov et al., 2024)	35.11	59.13	48.80	58.94	**16.30**	28.05	61.59	30.90	23.55	63.03	42.54
UnInfo (ours)	**38.76**	**60.79**	**50.93**	**60.92**	15.74	**28.82**	**63.03**	**31.70**	**25.57**	**63.57**	**43.98**

Table 14: Test accuracy (%) on domain shifts. The numbers (1), (5), and (10) presented with the method names are the shot numbers per class n used for the few-shot adaptation methods. '-' represents that the setting is infeasible due to the dataset size.

Method	ImageNet	ImageNet-A	ImageNet-R	Mean
No-adapt	66.97	32.91	74.36	58.08
Linear probing (1)	37.86	14.07	23.48	25.13
Linear probing (5)	55.95	-	52.91	-
Linear probing (10)	62.50	-	62.51	-
Tip-adapter (Zhang et al., 2022b) (1)	57.55	30.80	63.70	50.68
Tip-adapter (Zhang et al., 2022b) (5)	61.76	-	63.82	-
Tip-adapter (Zhang et al., 2022b) (10)	64.00	-	66.08	-
TPT (Shu et al., 2022)	70.70	19.92	80.26	56.96
ZERO (Farina et al., 2024)	69.02	43.23	75.53	62.59
MTA (Zanella & Ben Ayed, 2024b)	68.25	35.93	74.72	59.63
TDA (Karmanov et al., 2024)	70.33	33.24	76.81	60.13
UnInfo (ours)	67.82	33.13	76.53	59.16