Region-based Entropy Separation for One-shot Test-Time Adaptation

Kodai Kawamura
4620034@ed.tus.ac.jp
Shunya Yamagami
4623533@ed.tus.ac.jp
Go Irie
goirie@ieee.org

Tokyo University of Science
Tokyo, Japan

Abstract

In this paper, we address One-shot Test-Time Adaptation, which adapts a classification model using only a given single unlabeled test image. All the existing methods fine-tune the model so that the classification results are consistent for augmented views of a given test image. However, each region of an image has different information; some regions have rich class (object) information, while others express style information essentially irrelevant to the class information. The existing approach based on the image-level classification results is therefore inadequate. To address this problem, we propose a novel One-shot Test-Time Adaptation method based on region-based entropy separation. Specifically, our method aims to obtain style-invariant features by performing global entropy maximization as well as local entropy minimization only on the regions with high confidence values where the class information is considered to be strongly represented. Experimental results on three public benchmark datasets show that the proposed method outperforms the state-of-the-art One-shot Test-Time Adaptation methods. Code is available at: https://github.com/kodaikawamura/Region-basedTTA.

1 Introduction

While pre-trained Vision-Language Models (VLMs), such as CLIP [21] and ALIGN [12], have demonstrated remarkable zero-shot classification performance on various downstream tasks, their performance can be significantly improved by involving task-specific model adaptation [9, 33]. Typical model adaptation methods assume that they are able to access a large amount of downstream training data. However, collecting such a dataset is often difficult in real-world scenarios. To address this problem, Test-Time Adaptation (TTA) has collected much attention recently, where only a pre-trained model and unlabeled test data are available for adaptation [1, 4, 14, 16, 29]. In this paper, we address a more challenging but practical scenario, One-shot TTA (a.k.a. Test-Time Instance Adaptation or Single-Shot Adaptation) of pre-trained VLMs [7, 24, 26]; the task is to adapt a pre-trained VLM classifier such as CLIP [21] to a single unlabeled test image without assuming any other information is available.

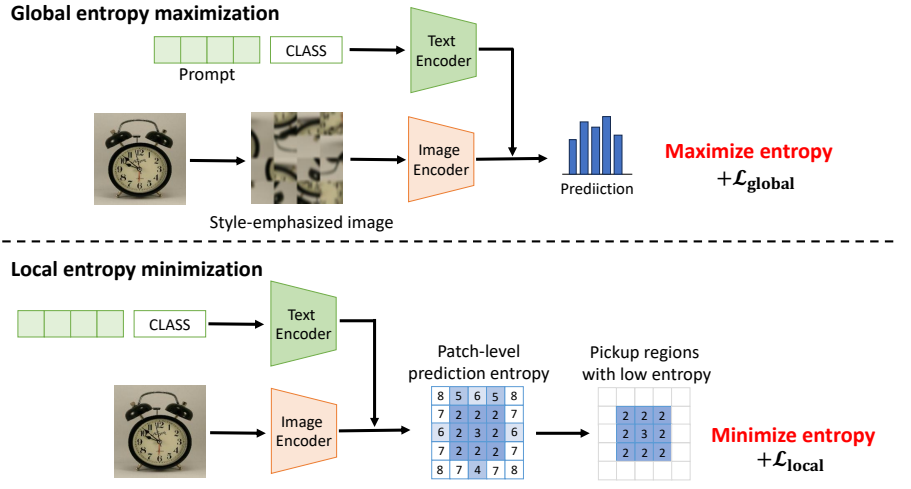


Figure 1: **Method overview.** We propose region-based entropy minimization for One-shot Test-Time Adaptation. More specifically, our method optimizes the text prompt by global entropy maximization and local entropy minimization. Global entropy maximization is performed by first transforming the input test image into style-emphasized one and then maximizing the entropy of the image-level prediction. Local entropy minimization first computes patch-level predictions and then minimizes the entropy only for the high-confident patches.

From the pioneering work on One-shot TTA for VLMs [24], several approaches have been proposed so far [7, 24, 26]. The key challenge is how to tune the model using only a single unlabeled test sample. The common idea is to use random augmentation and unsupervised loss functions; they first randomly generate multiple augmented views of the given test image, and then fine-tune the model (or trainable prompt) by using those augmented views and unsupervised loss functions, e.g., entropy minimization [7, 24, 26].

All the existing One-shot TTA methods rely solely on image-level information (e.g., image-level predictions or image-level loss functions) to perform adaptation. For example, Test-Time Prompt Tuning (TPT) [24] first computes the confidence for each augmented image and then minimizes the entropy of their average. However, this is likely to be suboptimal. Image is a spatial medium – different regions may have different information, e.g., some regions of an image may clearly depict a particular object but others may not and instead may represent significant style information. Thus, using only image-level information cannot properly reflect natural locality of images, leading to insufficient adaptation and possibly undesirable degrading classification performance.

In this paper, we propose a novel region-based entropy separation method for One-shot TTA. More specifically, our method simultaneously performs global entropy maximization and local entropy minimization, aiming to obtain classification performance for locally represented class (object) information while avoiding adaptation to misleading style information that appears throughout the entire image. Experimental results with three public benchmark datasets demonstrate that our method successfully outperforms the state-of-the-art One-shot TTA methods.

Setting	One-shot?	Sourse data Available?	Target data available?
Domain Adaptation	-	✓	✓
Domain Generalization	-	✓	-
Test-Time Adaptation	-	-	✓
One-shot TTA	✓	-	✓

Table 1: **Comparison of model adaptation problems.** In this paper, we address One-shot TTA, which is a variant of Test-Time Adaptation where only a single test sample from target data is available for adaptation. Our method is designed for One-shot TTA, hence does not require any source data or multiple test samples, unlike most existing model adaptation methods.

Our contributions in this paper can be summarized as follows:

- We propose region-based entropy separation for the One-shot TTA problem, which performs both global entropy maximization and local entropy minimization for effective model adaptation.
- Our method outperforms the state-of-the-art One-shot TTA methods.

2 Related Work

Model Adaptation for Large Pre-trained Models. While large pre-trained models have shown the promising zero-shot generalization performance, their performance can be further improved by adapting the models to target downstream tasks [9, 35, 38, 39]. Adapter-based adaptation such as CLIP-Adapter [9] and Tip-Adapter [35] achieves adaptation by adding extra parameters to the original model and updating only those parameters. Prompt tuning such as CoOp [39] and CoCoOp [38] does not alter the architecture of the model at all, but instead updates only the learnable tokens for adaptation. In this paper, we rely on the framework of prompt tuning and address the problem of One-shot TTA.

Test-Time Adaptation. Among the typical variants of model adaptation problems (Table 1), Test-Time Adaptation (TTA) [7, 11, 24, 25], which requires adapting models to test samples on the fly, is a challenging and practical setting when we have no access to training samples. One key challenge in this setting is designing an efficient test-time objective. TENT [29] is a pioneering method that introduces entropy minimization as a test-time objective. Entropy minimization is widely used in many prior test-time adaptation methods [7, 8, 24, 34]. For example, MEMO [34] adapts all parameters of a network model by minimizing the marginal entropy of the model's predictions across augmented images. Other prior works such as TPT [24] and DiffTPT [7] optimize the trainable prompt of CLIP by minimizing the entropy of predictions for randomly augmented views.

However, all of these existing works focus only on image-level prediction when they optimize the model. There are different information in an image depending on the regions; some regions have rich class information, while others express strong style information. Therefore, only focusing on image-level prediction is insufficient as it overlooks the importance of different features in an image. To address this problem, our method aims to have a look at local features of an image by leveraging CLIP's local features.

(a) Original (b) Transformed (c) Original (d) Transformed

Figure 2: **Examples of transformed images.** We show the style-emphasized images of "Desk Lamp" from different domains in Office-Home: *art* and *clipart*. Our method applies block shuffle and Gaussian blur to the original images to emphasize style information. The number of grids for block shuffle is 4 in these transformed images.

3 Method

3.1 Background

Pre-trained VLMs, such as CLIP [21], consists of two encoders, the image encoder f and the text encoder g. Given an image $\boldsymbol{x}$, a visual feature $\boldsymbol{f} = f(\boldsymbol{x})$ is obtained by the image encoder f. The textual prompt can be expressed as $\boldsymbol{t}_i$ for class i (e.g., $\boldsymbol{t}_i$="a photo of a [class]") which is then passed on to the text encoder g. Given the prompt $\boldsymbol{t}_i$ as an input, the text encoder g outputs text features as $\boldsymbol{g}_i = g(\boldsymbol{t}_i)$. For zero-shot classification of CLIP, the prediction probability is expressed as:

$$p(y_i|\boldsymbol{x}) = \frac{\exp(\cos(\boldsymbol{f}, \boldsymbol{g}_i)/\tau)}{\sum_{j=1}^{K} \exp(\cos(\boldsymbol{f}, \boldsymbol{g}_j)/\tau)}, \tag{1}$$

where $\cos(\boldsymbol{f}, \boldsymbol{g}_i)$ denotes cosine similarity for class i, and τ is the temperature parameter. To improve the performance of CLIP, prompt tuning methods such as CoOp [39] have been proposed [7, 24, 31, 32, 38, 39]. These methods optimize the trainable text prompts by leveraging training data [31, 32, 38, 39] or test data [7, 24].

3.2 Method Overview

Although each region of an image has different information, existing methods only focus on image-level classification results. Therefore, they cannot take account of local features that have diverse attributes depending on the regions, which limits their performance. To tackle this issue, we attempt to apply different learning methods based on the regions in an image.

Our approach is inspired by entropy separation [33], which is originally proposed for universal domain adaptation. Entropy separation aims to distinct images of "known" and "unknown" classes in an unsupervised manner by increasing (resp. decreasing) confidence of the confident (resp. unconfident) image through entropy minimization (resp. maximization).

In our approach, we extend this idea to region-based entropy separation. The overview of our method is shown in Figure 1. Specifically, our method attempts to obtain the representations insensitive to style information by decreasing the confidence for the global feature with strong style information as well as increasing the confidence for the local features with rich class information. To obtain the image with rich style information, our method leverages the style-emphasizing transformation. Moreover, our method utilizes patch-level predictions

Method	Office-Home	VLCS	PACS	Mean
Zero-shot CLIP [21]	82.30	82.40	96.10	86.93
TAF-Cal [36]	67.10	-	86.85	-
Xiao et al. [30]	72.05	-	84.13	-
TPT [24]	76.60	80.23	<u>96.50</u>	84.44
DiffTPT [7]	75.15	82.23	96.28	84.55
PromptStyler [5]	83.58	82.90	**97.23**	87.89
Ours (Bottom-K)	<u>83.70</u>	**84.18**	**97.23**	**88.40**
Ours (Thresholding)	**83.73**	<u>84.13</u>	**97.23**	<u>88.38</u>

Table 2: **Main results.** Accuracies of the model are listed. The highest accuracy is highlighted in bold, and the second highest accuracy is underlined. We test two region selection methods for local entropy minimization: "Bottom-K" picks regions with the K smallest predicted entropy, and "Thresholding" selects regions below a threshold τ. Our method outperforms the state-of-the-art in all the datasets. Although PromptStyler is competitive with our method, it has several limitations which is described in Sec.4.2.

Method	Office-Home	VLCS	PACS
Local entropy maximization	82.13	82.68	96.34
Global entropy maximization	**83.70**	**84.18**	**97.23**

Table 3: **Does local maximization work?** The accuracies of the model when it employs local maximization or global maximization are listed. Local maximization indicates that the model maximizes the entropy of predictions for regions with high entropy. Global maximization denotes that the model maximizes the entropy of predictions for the domain-emphasized image. As we can see from the table above, our method shows much better performance when it employs global entropy maximization rather than local entropy maximization.

to obtain the regions with rich class information by leveraging the value features of CLIP's image encoder.

3.3 Global Entropy Maximization

First, to obtain the style-invariant features, our method attempts to obtain the image with rich style information by leveraging style-emphasizing transformation. Our method is based on the assumption that class and style information in images are independent of each other. Therefore, by transforming an image into "class-destructed" form, it is possible to obtain style-variant and class-invariant representations [17]. To destruct class-variant features, we focus on local structural information, such as object shapes and the interconnections between its parts, that serves as vital representations of class in images. Our method disrupts these significant clues by employing transformation that divides the original image into pixel blocks and subsequently randomizing their positions [3, 17, 20]. Furthermore, we apply Gaussian blur to obfuscate class information. Given that the style information is captured by abstracted information such as the mean and the standard deviation [10], Gaussian blur does not affect the style information but is able to disrupt class information. Examples of the transformation is shown in Figure 2.

Our method then optimizes prompts by using the transformed images. To obtain style-invariant features, the model needs to learn prompts so that style-emphasized images are not

Figure 3: **Visualization of picked up patches.** The original images and the regions that the model picks up to minimize the entropy are shown. The patches that are not selected by the model are colored in gray. We can see that the model correctly selects the regions that possess strong class information.

classified to any class. Therefore, our model decreases the confidence of the prediction for transformed images by maximizing entropy. Formally, the loss function for this is as follows:

$$\mathcal{L}_{\text{global}} = -H(p), \tag{2}$$

where $H(\cdot)$ is the entropy function and p denotes the prediction probabilities of the style-emphasized images.

3.4 Local Entropy Minimization

Since we observe the regions that have rich class information in images, our method attempts to increase the confidence for the regions with strong class information to retain this information. To achieve this, our method picks up local regions that possess rich class information from the original image by leveraging CLIP's local features [13, 18, 19, 22, 25, 37, 40].

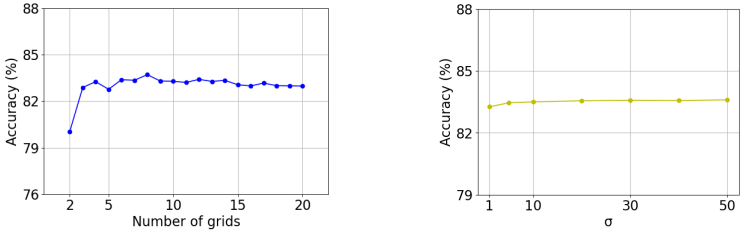
To obtain CLIP's local features, we first project the visual feature $\boldsymbol{f}'_i$ to the textual space for each patch $i \in I = \{0, 1, 2, \ldots, H \times W - 1\}$, where H and W are the height and width of the CLIP's feature map. We can formulate this as follows:

$$\boldsymbol{f}_i = \text{Proj}(v(\boldsymbol{f}'_i)), \tag{3}$$

where v is the value projection and Proj denotes the projections from the visual space to the textual space. Our method uses projections that are inherent in CLIP, therefore they do not require any additional training. Since obtained $\boldsymbol{f}_i$ has a rich local visual and textual alignment [13, 18, 19, 22, 25, 37, 40], our method leverages this to obtain patch-level predictions. The classification prediction probability for each patch i is computed as

$$p_i(y = m|\boldsymbol{x}) = \frac{\exp(\cos(\boldsymbol{f}_i, \boldsymbol{g}_m)/\tau)}{\sum_{m'=1}^{M} \exp(\cos(\boldsymbol{f}_i, \boldsymbol{g}_{m'})\tau)}. \tag{4}$$

To obtain the regions that possess rich class information, we select the patches with low prediction entropy (i.e. high confidence). We employ two ways of selecting the patches



(a) Number of grids (b) σ

Figure 4: **Sensitivity to hyperparameters in style emphasizing transformation.** The sensitivity of the model to hyperparameters in style emphasizing transformation are shown. We can see that 3 or more number of grid is effective to emphasize style information, and our model shows stable performance within a wide range of σ.

with low prediction entropy: picking up patches with K lowest entropy or thresholding with τ. Specifically, our method first computes the prediction entropy of each patch $i \in I = \{0,1,2,\ldots,H \times W - 1\}$. In the first approach, it then selects K indices that have K lowest prediction entropy. With selected K patch-level predictions, our method optimizes prompts by minimizing the sum of K prediction entropy. This loss function is formulated as:

$$\mathcal{L}_{\text{local}} = \sum_{j=1}^{K} H(p_j), \tag{5}$$

where $H(\cdot)$ is the entropy function and p_j is the prediction probabilities for selected patches $j \in \{1,2,\ldots,K\}$. In the second approach, our method selects patches with a prediction entropy below a threshold τ. Formally, it selects patches $j \in \{1,2,\ldots,n\}$ such that $H(p_j) \leq \tau$, and minimize the sum of n prediction entropy. In this case, the formulation is as follows:

$$\mathcal{L}_{\text{local}} = \sum_{j=1}^{n} H(p_j). \tag{6}$$

3.5 Total Loss

The final objective can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{global}} + \lambda \mathcal{L}_{\text{local}}, \tag{7}$$

where λ is a hyperparameter that balances two loss functions.

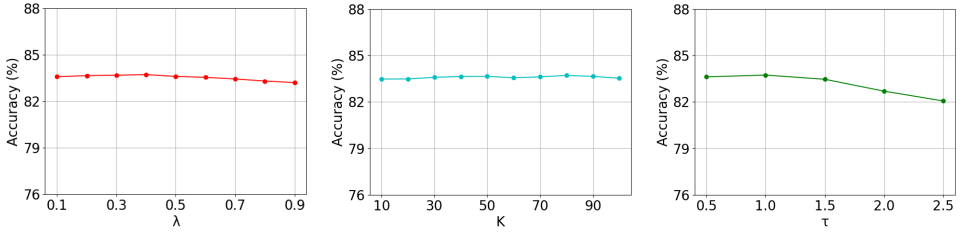
4 Experiments

4.1 Setup

Datasets. We evaluate our method on three public benchmark datasets, i.e., Office-Home [28], VLCS [7], and PACS [15], which are widely used for domain adaptation/generalization. Office-Home has 15,500 samples of 65 object classes from 4 domains (i.e., *art*, *clipart*, *product*, and *real-world*). VLCS consists of 729 images of 5 classes from 4 domains (i.e.,

$+\mathcal{L}_{\text{global}}$	$+\mathcal{L}_{\text{local}}$ (Bottom-K)	$+\mathcal{L}_{\text{local}}$ (Thresholding)	Office-Home	VLCS	PACS
-	-	-	82.30	82.40	96.10
✓	-	-	83.60	84.09	97.18
-	✓	-	82.30	83.53	96.10
-	-	✓	82.40	83.17	96.23
✓	✓	-	<u>83.70</u>	**84.18**	**97.23**
✓	-	✓	**83.73**	<u>84.13</u>	**97.23**

Table 4: **Ablation of loss functions.** "Bottom-K" indicates picking up regions with K lowest prediction entropy, and "Thresholding" denotes selecting regions with a prediction entropy below a threshold τ when minimizing entropy. Although our method achieves improvement only with $\mathcal{L}_{\text{global}}$, by employing both of $\mathcal{L}_{\text{global}}$ and $\mathcal{L}_{\text{local}}$, it shows higher performance.



(a) λ (b) K (c) τ

Figure 5: **Sensitivity to hyperparameters.** The sensitivity of the model to hyperparameters are shown. We can see that the model is not very sensitive to hyperparameters.

Caltech, *LabelMe*, *SUN09*, and *VOC2007*). PACS has 9,991 samples of 7 categories from 4 domains (i.e., *photo*, *art-painting*, *cartoon*, and *sketch*). We report the average results over three runs using different random seeds. Since we focus on One-shot TTA in our experiments, we use only a single unlabeled test sample.

Implementation Details. We initialize the prompt to "a photo of a [class]" and fine-tune the four tokens corresponding to "a photo of a", using the SGD with learning rate of 0.00008 for Office-Home and PACS, and 0.00001 for VLCS. We use ViT-B/16 [6] as a backbone. We consistently set the number and the threshold for picking up regions in entropy minimization $K = 80$, $\tau = 1.0$, and set the weight in loss function $\lambda = 0.4$. For style-emphasizing transformation, we set hyperparameters as follows: the numbers of grids for block shuffle are 8 for Office-Home, 17 for VLCS and 3 for PACS, and kernel size and sigma for Gaussian blur are (5,5) and $\sigma \sim U(10, 30)$ respectively.

Baselines. We compare our method with four groups of the state-of-the-art methods, namely; **(a)** zero-shot CLIP, **(b)** TAF-Cal [36] and [30], the state-of-the-art One-Shot TTA methods, not for VLMs, **(c)** TPT [24] and DiffTPT [7], the state-of-the-art One-Shot TTA methods for VLMs, and **(d)** PromptStyler [5], the state-of-the-art source-free domain generalization method. For fair comparisons, we employ ViT-B/16 for CLIP image encoder for all the methods compared (except for those in **(b)**).

Block shuffle	Gaussian Blur	Office-Home	VLCS	PACS
-	-	76.00	81.30	93.80
✓	-	77.43	82.55	94.50
-	✓	79.78	82.95	95.18
✓	✓	**83.70**	**84.18**	**97.23**

Table 5: **Ablation of style-emphasizing transformation.** Our method successfully emphasize the style information by employing both of block shuffle and Gaussian blur. However, employing either of them fails in emphasizing the style information, which leads to degradation of the model.

4.2 Main Results

Comparative results are shown in Table 2. Our method outperforms all the baselines in all the datasets. In particular, our method improves the state-of-the-art One-Shot TTA methods, TPT and DiffTPT, by more than 7% in Office-Home, which shows clear significance of our method. The only very recent domain generalization method, PromptStyler, is highly competitive with our method; the benefit of our method is that we do not change anything of the VLM-based classification framework, unlike PromptStyler that uses an additional linear classifier. Moreover, PromptStyler has a limitation that it needs to fix the number of styles beforehand.

4.3 Analysis

Does local entropy maximization work? Although our method applies entropy maximization to obtain style-invariant features, it might seem more straightforward to employ local entropy maximization (i.e., maximizing the entropy of predictions for regions with low confidence) rather than global entropy maximization. However, as shown in Table 3, global entropy maximization works much better than local entropy maximization. This is because style-emphasizing transformation effectively emphasizes the style information and makes it possible to obtain style-variant features.

Visualization of picked up patches. We show the visualization of patches that are selected by the model in Figure 3. We can see that our model correctly picks up the regions with rich class information.

Ablation of loss functions. We show the impact of $\mathcal{L}_{\text{global}}$ and $\mathcal{L}_{\text{local}}$ in Table 4. As the table shows, $\mathcal{L}_{\text{global}}$ itself achieves the performance improvement, but by incorporating both of two loss functions, our method achieves higher performance.

Ablation of style-emphasizing transformation. In style-emphasizing transformation, we use several techniques, such as block shuffle and Gaussian blur, to emphasize style information. To evaluate the effect of the number of grids in block shuffle and σ for Gaussian blur, we show the accuracy of the model for Office-Home when we change the number of grids and σ in Figure 4. This evaluations shows that 3 or more number of grids for block shuffle effectively emphasize the style information. Ablation studies on the components in the transformation is shown in Table 5. As ablation studies show, by employing both block shuffle and Gaussian blur, the transformation emphasizes style information effectively, which leads to higher performance.

Sensitivity to hyperparameters. λ, K and τ are major hyperparameters to control the

impact of L_{local} in Eq. (5). We evaluate the sensitivity to these hyperparameters in Figure 5. We can see that the model shows stable performance within a wide range of hyperparameters.

5 Conclusions

We proposed a One-shot TTA method based on region-based entropy separation. Our method jointly performs global entropy maximization to obtain style invariance and local entropy minimization to improve adaptive classification ability. Experimental results showed that our method outperformed the state-of-the-art One-shot TTA methods.

References

[1] Fatemeh Azimi, Sebastian Palacio, Federico Raue, Jörn Hees, Luca Bertinetto, and Andreas Dengel. Self-supervised test-time adaptation on video data. In *Proc. WACV*, 2022.

[2] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*, 2020.

[3] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019.

[4] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proc. CVPR*, 2022.

[5] Junhyeong Cho, Gilhyun Nam, Sungyeon Kim, Hunmin Yang, and Suha Kwak. Promptstyler: Prompt-driven style generation for source-free domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15702–15712, 2023.

[6] 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.

[7] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2704–2714, 2023.

[8] François Fleuret et al. Test time adaptation through perturbation robustness. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

[9] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, 132(2):581–595, 2024.

[10] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[11] Yujia Huang, James Gornet, Sihui Dai, Zhiding Yu, Tan Nguyen, Doris Tsao, and Anima Anandkumar. Neural networks with recurrent generative feedback. *Advances in Neural Information Processing Systems*, 33:535–545, 2020.

[12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[13] Kwanyoung Kim, Yujin Oh, and Jong Chul Ye. Zegot: Zero-shot segmentation through optimal transport of text prompts. *arXiv preprint arXiv:2301.12171*, 2023.

[14] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proc. CVPR*, 2020.

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

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

[17] Yu Mitsuzumi, Go Irie, Daiki Ikami, and Takashi Shibata. Generalized domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1084–1093, 2021.

[18] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Zero-shot in-distribution detection in multi-object settings using vision-language foundation models. *arXiv preprint arXiv:2304.04521*, 2023.

[19] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[20] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[22] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022.

[23] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3118–3126, 2018.

[24] 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*, 35:14274–14289, 2022.

[25] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *Advances in Neural Information Processing Systems*, 35:30569–30582, 2022.

[26] Kowshik Thopalli, Rakshith Subramanyam, Pavan K. Turaga, and Jayaraman J. Thiagarajan. Target-aware generative augmentations for single-shot adaptation. In *Proc. ICML*, 2023.

[27] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[28] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[29] 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.

[30] Zehao Xiao, Xiantong Zhen, Ling Shao, and Cees GM Snoek. Learning to generalize across domains on single test samples. *arXiv preprint arXiv:2202.08045*, 2022.

[31] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023.

[32] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021.

[33] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2720–2729, 2019.

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

[35] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

[36] Xingchen Zhao, Chang Liu, Anthony Sicilia, Seong Jae Hwang, and Yun Fu. Test-time fourier style calibration for domain generalization. *arXiv preprint arXiv:2205.06427*, 2022.

[37] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022.

[38] 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)*, pages 16816–16825, June 2022.

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

[40] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023.