Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluation
DOI: 10.54254/2755-2721/2025.PO24674

GCCA: Utilizing Gradient Matching for Open-set Domain Adaptation of Vision-language Models

Mingxi Jiang[1], Shengyong Xu[1*]

[1]*College of Engineering, Huazhong Agricultural University, Wuhan, China*
**Corresponding Author. Email:xsy@mail.hzau.edu.cn*

Abstract. Unsupervised Domain Adaptation (UDA) traditionally focuses on learning domain-invariant representations by aligning source and target domains, yet struggles in open-set scenarios where target domains contain unseen categories. Recent advances in vision-language models like CLIP, which leverage multimodal pre-training on large-scale image-text pairs, exhibit remarkable zero-shot generalization but still face performance degradation under significant domain shifts. In this work, we propose a gradient matching strategy to enhance CLIP's adaptability in open-set UDA. Unlike conventional prompt-based adaptation methods that independently optimize domain-specific parameters, our approach formulates domain alignment as a gradient consensus problem. Specifically, we enforce gradient direction consistency between source and target objectives to reconcile their optimization paths, while applying gradient norm penalization to prevent overfitting to domain-specific noise. This dual mechanism allows CLIP to preserve its inherent cross-modal discriminability while acquiring domain-robust features. Experiments across diverse open-set adaptation benchmarks demonstrate that our gradient matching framework substantially outperforms existing CLIP adaptation baselines, achieving state-of-the-art performance without compromising model stability.

Keywords: Unsupervised Domain Adaptation, Large multimodal models, Out-of-distribution detection

1. Introduction

In real-world machine learning applications, models often face the challenge of a mismatch between the distributions of training data and test data—a phenomenon known as domain shift [1-3]. Unsupervised Domain Adaptation (UDA) [4-6] addresses this issue by transferring knowledge from a labeled source domain to an unlabeled target domain, providing an important framework for mitigating distributional discrepancies. Traditional UDA methods typically focus on learning domain-invariant feature representations [7,8], employing adversarial training [9], statistical alignment [10], or feature disentanglement to minimize inter-domain distribution gaps. However, these approaches suffer from two fundamental limitations: first, an excessive emphasis on domain invariance can weaken the model's ability to learn discriminative features, leading to suboptimal solutions [11] ; second, their closed-set assumption requires that the source and target domains share

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.PO24674

an identical label space, which contradicts real-world scenarios where the target domain may contain unknown classes in an open-set setting [12] . This gap between theory and practice has prompted researchers to reexamine UDA's core assumptions and to seek more practically adaptive solutions.

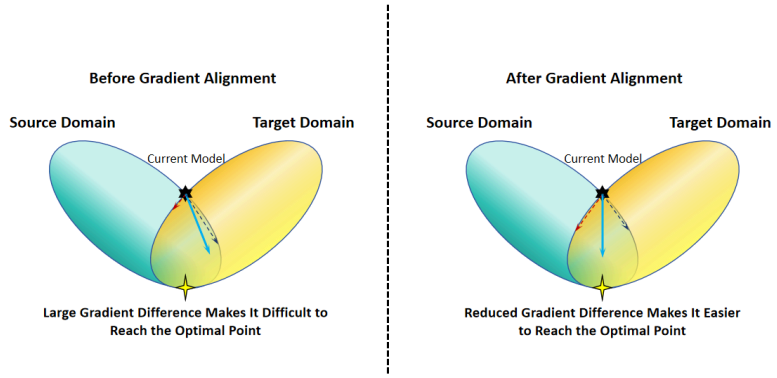

Figure 1: Method diagram

In recent years, large-scale vision–language pretraining models [13-16] exemplified by CLIP [17] have demonstrated breakthrough cross-modal alignment capabilities. By contrastive learning on a dataset of 400 million image–text pairs, they establish a unified semantic space that exhibits astonishing zero-shot classification performance. At its core, CLIP maps images and text into a common embedding space, enabling flexible adaptation to downstream tasks via natural-language prompts. Yet, when applied to domain adaptation tasks [18], CLIP's performance reveals a clear duality: on one hand, its powerful semantic abstraction offers a potential bridge for cross-domain knowledge transfer [19]; on the other hand, its fixed pretrained parameters struggle to dynamically adjust cross-modal alignment under significant domain shifts. This is especially problematic in open-set scenarios [20], where the model must simultaneously handle domain discrepancies among known classes and detect unknown categories. This tension exposes a key bottleneck for current vision–language models in domain adaptation: how to effectively reconcile cross-modal representation learning with domain-specific feature adaptation while maintaining robustness to open-set conditions.

Existing studies have attempted to enhance CLIP's domain adaptation ability through prompt learning [18,20,21], advancing along two main directions: first, designing domain-sensitive prompt templates that adjust the classifier's decision boundary via learnable text prompts; second, constructing a joint vision–language prompt space to jointly fine-tune both the image encoder and the text encoder. Although these methods have made progress on closed-set domain adaptation tasks, they still face limitations along three dimensions. First, most adopt a two-stage optimization strategy —aligning domains before tuning the classifier—yet such decoupled learning fails to capture the synergistic effects of cross-domain and cross-modal interactions. Second, current prompt-learning approaches typically rely on a single objective function (e.g., contrastive loss or domain-adversarial loss) and ignore gradient conflicts between different optimization goals—a problem that is particularly acute in parallel multi-task open-set scenarios. Third, with respect to the special challenges of open-set adaptation, existing methods lack a dynamic mechanism to balance known-class adaptation and unknown-class detection, limiting their generalization under complex domain shifts.

To resolve this, we propose Gradient-Consensus CLIP Adaptation (GCCA), a novel framework that reframes open-set UDA as a multi-objective optimization problem governed by gradient alignment principles. Our core insight stems from two observations: First, in open-set UDA, the

DOI: 10.54254/2755-2721/2025.PO24674

optimization objectives for source and target domains are inherently misaligned due to category mismatches; Second, CLIP's pre-trained weights encode rich cross-modal priors that should be preserved during adaptation. Instead of directly aligning feature representations or prompt embeddings, GCCA operates on the gradient space to establish consensus between domain-specific optimization trajectories. Specifically, we formulate the adaptation process through a bi-level optimization lens: (1) Primary Objectives: For the source domain, we minimize the standard cross-entropy loss between CLIP's predictions and ground-truth labels. For the target domain, we minimize the Chi-Squared Divergence of predictions on known categories while maximizing entropy on confident open-set samples identified as outliers. (2) Gradient Consensus: We compute gradients of KL-divergence with respect to CLIP's text prompt parameters. A gradient direction alignment loss is then imposed to maximize the similarity between source and target gradients, effectively reconciling their optimization directions. (3) Gradient Norm Regularization: To prevent overfitting to noisy pseudo-labels in the target domain, we penalize the nuclear norm of target gradients, enforcing smoother parameter updates.

We evaluate GCCA on open-set UDA benchmark Office-Home compared with comparative SOTA methods like CoOp [13], LoCoOp [22]. Experimental results demonstrate the effectiveness of our methods.

2. Related works

2.1. Out-of-Distribution detection

Out-of-Distribution (OOD) detection is a prominent and critical research area in machine learning, aimed at accurately identifying test data samples (target domain) that exhibit distributional disparities from the training data (source domain). This capability is vital for ensuring model reliability in open-world scenarios. The concept was first introduced by Hendrycks and Gimpel, and over the subsequent years, OOD detection has evolved significantly, progressing from rudimentary Softmax probability-based baselines to contemporary multi-modal solutions.

Hendrycks et al [23]. proposed the Maximum Softmax Probability (MSP) as a baseline for OOD detection. While this method is straightforward and effective for low-complexity tasks, it is prone to overconfidence. Liang et al [24]. introduced the ODIN method, incorporating temperature scaling and input preprocessing to enhance detection accuracy. Hendrycks et al [25]. developed Outlier Exposure, leveraging auxiliary outlier datasets to improve model performance. Li et al [26]. explored the Mahalanobis distance, offering a distance-based perspective for OOD detection. Liu et al [27]. proposed an energy-based OOD detection framework, reformulating the task as a function of energy scores derived from model logits, thereby providing a novel theoretical lens.Lin et al [28]. introduced the Multi-level OOD Detection (MOOD) framework, enabling detection across multiple neural network layers, while Sun et al [29]. proposed DICE, a sparsity-based technique that significantly improves detection performance. The advent of vision-language models [30], such as CLIP, facilitated zero-shot OOD detection, eliminating the need for additional training.OOD research began integrating Open Set Recognition (OSR) to enhance the identification of unknown categories [31]. Few-shot learning approaches [32], particularly those utilizing CLIP for anomaly detection, gained traction in industrial inspection applications.Recent research has increasingly emphasized practical applications [31], introducing the concept of generalized OOD detection and developing new benchmarks, such as ImageNet-OOD, to rigorously evaluate the efficacy of various OOD detection methodologies.

DOI: 10.54254/2755-2721/2025.PO24674

2.2. Open Set Domain Adaptation

Open Set Domain Adaptation (OSDA) is a specialized subfield of domain adaptation that focuses on distinguishing categories present in the target domain but absent in the source domain. This ensures that models do not erroneously classify unknown categories as known ones, thereby enhancing classification accuracy. The concept of OSDA was first introduced by Busto et al. and has since evolved over the years into a robust research area encompassing both comprehensive theoretical frameworks and practical applications.

Busto et al [33]. pioneered the concept of OSDA, highlighting that the traditional closed-set, independent, and identically distributed assumption in domain adaptation often fails in real-world scenarios. They proposed a unified approach addressing both closed-set and open-set challenges.Saito et al [34]. introduced an adversarial training-based method, leveraging discriminators and generators to handle the distinct sample distributions of the source and target domains, significantly improving the model's ability to manage unknown categories in the target domain. OSDA research began transitioning from theoretical exploration to practical applications. Fang et al [35]. introduced the concept of "open set difference" by analyzing the classifier's risk in identifying unknown categories, quantifying the error risk associated with such classifications. Panareda Busto et al [36]. applied OSDA to specific tasks, such as image and action recognition, demonstrating its efficacy and feasibility in cross-domain visual tasks.Xu et al [37]. incorporated Extreme Value Theory into OSDA to address the identification of unknown categories, achieving notable advantages in scenarios with extreme domain shifts. Zhuang et al [38]. proposed the Self-supervised Class-Discovering Adapter (SCDA), which iteratively discovers unknown categories while updating domain-adaptive features. This approach underscored the potential of self-supervised learning in target domains with limited labeled data.Choe et al [39]. extended OSDA to the domain of semantic segmentation, further broadening its application scope and validating its feasibility in pixel-level tasks.

2.3. Vision-Language Models

Vision-Language Models (VLMs) constitute a pivotal subfield within multimodal artificial intelligence, experiencing substantial advancements in recent years driven by progress in deep learning. Originally conceptualized as theoretical constructs, VLMs have matured into robust technologies widely applied to tasks such as image captioning, visual question answering, and image-text retrieval.

Google introduced the Transformer architecture [40], which achieved remarkable success in natural language processing (NLP). This framework, with its self-attention mechanism, laid the foundation for VLM pretraining and facilitated its extension to multimodal applications.The introduction of BERT marked a significant milestone in NLP [41], heralding the rise of pretrained models. BERT's self-supervised learning paradigm was swiftly adapted to the VLM domain, providing a blueprint for subsequent model training strategies.VLMs began to emphasize joint training of image and text modalities, employing dual-stream architectures to extract visual and linguistic features, which were then fused through interaction layers.The Transformer architecture was first applied to video-language pretraining, addressing tasks such as video action classification.Models such as ERNIE-VIL [42], HERO and UniVL [43] emerged, leveraging larger-scale datasets and more sophisticated training strategies to enhance VLM performance in tasks like visual question answering and caption generation. OpenAI introduced CLIP [44], which utilized contrastive learning on vast image-text pair datasets to achieve unprecedented zero-shot

DOI: 10.54254/2755-2721/2025.PO24674

performance. Concurrently, DALL-E [45] was proposed, enabling high-fidelity text-to-image generation and further diversifying the technical landscape of VLMs. Models like BLIP [46] and OFA [47] introduced joint encoder architectures and sequence-to-sequence learning frameworks, significantly improving training efficiency and generalization capabilities.The introduction of BLIP-2 [48] , which incorporated instruction-based fine-tuning and large-scale training, markedly enhanced model performance in tasks such as computer vision question answering and image captioning. VLM research concurrently shifted toward efficient multimodal processing based on smaller-scale language models, with a growing emphasis on optimizing training efficiency.

3. Methodology

In this section, we elaborate on the proposed Gradient-Consensus CLIP Adaptation (GCCA) framework in detail. We first revisit the fundamental challenges of open-set domain adaptation (OSDA) and motivate our choice of a gradient-space formulation over traditional feature alignment. We then introduce our bi-level optimization framework, which comprises three core components: (1) supervised learning on the source domain with cross-entropy loss, and unsupervised adaptation on the target domain using a divergence-aware objective; (2) gradient consensus enforcement through cosine alignment of task gradients; and (3) gradient norm regularization to suppress noise from unreliable pseudo-labels in the target domain. Notice that our method is based on CoOp [13], where we freeze CLIP's parameters and only fine-tune the learnable text prompts.

3.1. Problem formulation and motivation

Let $\mathscr{D}_s = \{(x_s^i, y_s^i)\}_{i=1}^{N_s}$ denote the labeled source domain drawn from a set of known classes $\mathscr{C}_s$ and $\mathscr{D}_t = \{x_t^i\}_{i=1}^{N_t}$ be the unlabeled target domain. Importantly, $\mathscr{C}_s \subset \mathscr{C}_t$, where $\mathscr{C}_t = \mathscr{C}_k \cup \mathscr{C}_u$, and $\mathscr{C}_{\mathrm{u}}$ denotes unknown (open-set) classes present only in the target domain. The goal of open-set domain adaptation (OSDA) is to accurately classify samples in $\mathscr{C}_k$ while rejecting or abstaining from predictions on $\mathscr{C}_{\mathrm{u}}$.

Traditional domain adaptation methods typically rely on feature space alignment—through minimizing feature distributional discrepancies (e.g., MMD, adversarial training)—but these approaches often break down in ODA due to category shift. In particular, aligning features across known and unknown classes can result in negative transfer, where unknown target samples are erroneously pulled toward known class clusters.

To circumvent this, our key insight is to operate not in the feature space, but in the gradient space, where we interpret adaptation as a process of aligning optimization directions. This is inspired by the following two observations:

· Misaligned Objectives: The optimization landscapes for source and target domains are inherently different. The source domain involves supervised learning on $\mathscr{C}_s$, whereas the target domain must balance learning from unlabeled known-class samples and rejecting open-set outliers. Therefore, direct alignment in the feature space may conflict with the distinct learning signals in each domain.

· Preservation of CLIP Priors: CLIP's pre-trained vision-language model encodes strong cross-modal knowledge. However, aggressive fine-tuning can degrade its generalization ability, especially under noisy pseudo-labels. A gradient-level formulation allows us to steer optimization while preserving pre-trained priors encoded in CLIP's text prompts.

DOI: 10.54254/2755-2721/2025.PO24674

3.2. Bi-level optimization for domain-aware adaptation

We reframe the ODA problem as a bi-level optimization, where the primary objective focuses on task-specific supervision, and a meta-level objective imposes cross-domain gradient alignment constraints.

3.2.1. Bi-level optimization for domain-aware adaptation

The source domain loss is defined as a standard supervised cross-entropy loss:

$$\mathscr{L}_{\text{src}} = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathscr{L}_{\text{CE}} \left(f_\theta \left(x_s^i, p_\phi \right), y_s^i \right),$$

Where $f_\theta \left(\cdot, p_\phi \right)$ is the CLIP visual encoder followed by prompt-conditioned classification logits, and p_ϕ are learnable text prompt embeddings. This term ensures that the model retains discriminative capability on known source categories.

3.2.2. Target domain objectives

Unlike the source domain, the target domain is unlabeled and contains both known and unknown classes. To handle this, we introduce two components:

(a) Known-Class Divergence Minimization

We estimate the distribution of target samples over known classes using a soft pseudo-labeling approach. Let $z_t = f_\theta \left(x_t, p_\phi \right)$ and $\hat{y}_t = \text{softmax} \left(z_t \right)$. We aim to minimize the Chi-Squared Divergence between the predicted target distribution and an empirical known-class prior π_k :

$$\mathscr{L}_{\text{chi}} = \frac{1}{N_t} \sum_{i=1}^{N_t} \sum_{c \in \mathscr{C}_s} \frac{\left(\hat{y}_t^i(c) - \pi_k(c) \right)^2}{\pi_k(c)}.$$

The use of Chi-Squared Divergence over KL-divergence is motivated by its robustness to probability mass shifts caused by unknown categories.

(b) Open-Set Entropy Maximization

To suppress overconfident predictions on open-set samples, we use an entropy maximization term. Let $\mathscr{O} = \left\{ x_t^j \in \mathscr{D}_t \mid \max_c \hat{y}_t^j \left(c \right) < \tau \right\}$ denote target samples with low maximum prediction confidence. We define:

$$\mathscr{L}_{\text{ent}} = -\frac{1}{|\mathscr{O}|} \sum_{x_t^j \in \mathscr{O}} \sum_{c \in \mathscr{C}_s} \hat{y}_t^j \left(c \right) \log \hat{y}_t^j \left(c \right),$$

which encourages the model to be uncertain about potential unknowns, thus increasing robustness.

The combined target objective is:

$$\mathscr{L}_{\text{tgt}} = \mathscr{L}_{\text{chi}} + \lambda_{\text{ent}} \cdot \mathscr{L}_{\text{ent}},$$

where λ_{ent} is a weighting hyperparameter.

DOI: 10.54254/2755-2721/2025.PO24674

3.3. Gradient consensus and regularization

Our main contribution lies in aligning source and target optimization dynamics in the gradient space, where gradients are taken with respect to the prompt parameters p_ϕ . This results in two auxiliary terms: gradient direction alignment and gradient norm regularization.

3.3.1. Gradient direction alignment (consensus)

Let $g_s = \nabla_{p_\phi} \mathscr{L}_{\text{src}}$ and $g_t = \nabla_{p_\phi} \mathscr{L}_{\text{tgt}}$ denote the gradients from source and target objectives, respectively. We define a gradient consensus loss to encourage these directions to align:

$$\mathscr{L}_{\cos} = 1 - \frac{\langle g_s, g_t \rangle}{\|g_s\|_2 \cdot \|g_t\|_2},$$

which minimizes the angle between the gradients, fostering agreement between domain-specific updates. This loss is grounded in multi-objective optimization theory, where aligned gradients are known to improve convergence and task compatibility.

3.3.2. Gradient norm regularization

Given that pseudo-labels on the target domain are inherently noisy, unconstrained gradients may result in overfitting. To alleviate this, we penalize the nuclear norm (i.e., sum of singular values) of the target gradient matrix:

$$\mathscr{L}_{\text{nuc}} = \| G_t \|_*$$

Where $G_t \in \mathbb{R}^{d \times n}$ is a Jacobian-like matrix constructed from gradients of multiple target samples (e.g., batch-wise gradients stacked as columns), and $\| \cdot \|_*$ denotes the nuclear norm. This encourages smoother updates and prevents the model from making abrupt changes due to outlier samples.

3.4. Gradient norm regularization

The total loss for training the model is a weighted sum of the above components:

$$\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{src}} + \alpha \cdot \mathscr{L}_{\text{tgt}} + \beta \cdot \mathscr{L}_{\cos} + \gamma \cdot \mathscr{L}_{\text{nuc}}$$

Where α, β, γ are hyperparameters controlling the trade-off between source learning, target adaptation, gradient consensus, and regularization.

The optimization proceeds iteratively using stochastic gradient descent, where gradients for each component are computed per mini-batch, and prompt embeddings p_ϕ are updated accordingly.

4. Experiments

4.1. Datasets

The Office-Home dataset is a widely utilized benchmark in domain adaptation research, designed to evaluate the generalization capabilities of models in cross-domain visual classification tasks. Introduced by Venkateswara et al. in 2017, the dataset comprises image data from four distinct

DOI: 10.54254/2755-2721/2025.PO24674

visual domains: Art, Clipart, Product, and Real-World. Each domain encompasses 65 categories, covering common office and household items such as laptops, chairs, and lamps. The images were collected using Python-based web crawling techniques, initially aggregating approximately 100,000 images from multiple search engines and online image repositories, spanning 120 object categories. Through a process of quality filtering and minimum category size requirements, the dataset was refined to a final set of 15,500 images, ensuring its usability and representativeness. This curation process guarantees that each category within each domain contains a sufficient number of images suitable for training and evaluating machine learning tasks.

The Office-Home dataset is particularly well-suited for unsupervised domain adaptation tasks, where labeled data is available in the source domain but absent in the target domain, allowing researchers to assess a model's generalization performance. The significant stylistic and contextual differences among the four domains (e.g., the contrast between abstract artistic images and real-world photographs) make the dataset an ideal testbed for studying domain shift challenges. For instance, images in the Art domain may exhibit greater abstraction, while Real-World images align more closely with practical application scenarios, thereby increasing the complexity of cross-domain learning. For machine learning researchers, particularly those focused on transfer learning and domain adaptation, the Office-Home dataset serves as a critical resource for developing and benchmarking algorithms to enhance model generalization across diverse visual domains. Baseline performance, evaluated using linear SVM and VGG-F network features, ranges from 44.99% to 66.41% classification accuracy (varying by domain), indicating substantial room for improvement and attracting significant research interest.

4.2. Metrics

This experiment employs AUROC, FPR95, and CCR10 as evaluation metrics.

AUROC is a classic metric for assessing the performance of binary classification models. It evaluates the model's ability to distinguish between positive and negative samples across all classification thresholds. The ROC curve plots TPR on the y-axis against FPR on the x-axis, with each point on the curve representing the TPR and FPR at a specific threshold. AUROC, the area under the ROC curve, ranges from [0, 1]. An AUROC of 1 indicates ideal classification ability, perfectly distinguishing all positive and negative samples, while an AUROC of 0.5 suggests the model lacks discriminative power.

FPR95 is a commonly used metric in binary classification tasks requiring high recall. It denotes the FPR when the model's TPR reaches 95%. This metric reflects the probability of incorrectly classifying negative samples as positive in high-recall tasks, thus indicating the model's false positive cost. A lower FPR95 value signifies that the model effectively balances high recall with low false positive rates.

CCR10 represents the probability of correctly classifying ID samples into their corresponding categories when the classification threshold is set to achieve an FPR of 10%. This metric evaluates the model's overall classification accuracy at a fixed FPR of 10%, helping designers find a balance between misclassification and accurate classification.

4.3. Experimental results

To ensure the fairness of comparisons, we strictly adopted the hyperparameter settings of CoOp in all experiments: we used 4 learnable prompts initialized with "a photo of a"; the optimizer was SGD with a base learning rate of 1×10^{-4}, gradually decayed via a cosine annealing schedule, and an

DOI: 10.54254/2755-2721/2025.PO24674

additional warmup learning rate of 1×10^{-5} was applied in the first training epoch; the hyperparameters in the loss function were fixed at α =0.1, β =0.01, and γ =0.01. All models were trained for 5 epochs before evaluation, and the final score was uniformly determined as the maximum softmax output value after temperature scaling.

To validate the performance improvements of the proposed method, we selected three types of comparison baselines: (1) zero-shot CLIP, to demonstrate the native classification capability of CLIP; (2) the prompt learning method CoOp [13], to evaluate the robustness of prompt learning when there is a significant distribution shift between training and test sets; and (3) the state-of-the-art CLIP out-of-distribution detection method LoCoOp [22]. In the experimental design, we paired 4 data domains as source and target domains in a round-robin fashion, resulting in 12 experimental settings. For each setting, we recorded three metrics: AUROC, FPR@TPR95, and CCR@FPR10. The results are summarized in the table below.

Table 1: Experimental Results. In each experiment, the left side of the arrow denotes the Source Domain, and the right side denotes the Target Domain. Specifically, "Pr" represents "Product," "Rw" represents "Real_World," "Cl" represents "Clipart," and "Ar" represents "Art." Regarding evaluation metrics: CCR10 refers to the Correct Classification Rate at FPR=10% (CCR@FPR10), where a higher value indicates better performance; FPR95 refers to the False Positive Rate at TPR=95% (FPR@TPR95), where a lower value indicates better performance; a higher AUROC value indicates superior performance. The best metric in each column is highlighted in bold

	Pr→Rw			Pr→Cl			Pr→Ar			Ar→Pr			Ar→Rw			Ar→Cl		
	Acc10	FPR95	AUROC	Acc10	FPR95	AUROC	Acc10	FPR95	AUROC	Acc10	FPR95	AUROC	Acc10	FPR95	AUROC	Acc10	FPR95	AUROC
CLIP+MCM(T=0.1)	77.65	31.73	92.823	51.683	80.313	79.24	63.873	59.843	86.523	76.823	43.543	92.02	77.65	31.73	92.82	51.68	80.31	79.24
CoOp+MCM(T=0.1)	72.2055	50.06	90.386	44.686	85.346	78.2	49.286	69.566	82.566	76.776	51.146	91.86	78.72	44.45	92.52	52.63	84.68	79.67
LOCOOP+MCM(T=1)	71.83	43.84	89.295	46.095	82.015	77.72	51.935	66.715	81.695	61.015	51.225	87.36	66.7	46.39	88.3	46.09	83.64	76.31
Ours	83.61	32.9	94.092	58.752	81.932	80.74	68.412	56.232	88.282	78.502	44.832	92.69	80.74	34.15	93.56	55.34	81.26	80.62

Pr→Ar			Rw→Pr			Rw→Cl			Cl→Rw			Cl→Ar			Cl→Pr			Avg		
Acc10	FPR95	AUROC	Acc10	FPR95	AUROC	Acc10	FPR95	AUROC	Acc10	FPR95	AUROC	Acc10	FPR95	AUROC	Acc10	FPR95	AUROC	Acc10	FPR95	AUROC
63.87	59.84	86.52	76.823	43.542	92.023	51.683	80.313	79.243	77.653	31.733	92.823	63.783	59.843	86.523	76.823	43.543	92.023	67.508	53.858	87.653
60.31	60.08	86.43	82.596	42.746	94.096	50.716	84.416	79.396	80.656	46.246	92.646	67.836	63.286	87.146	80.076	45.046	93.566	66.37	60.587	87.371
52.39	61.93	82.93	75.975	45.085	91.195	47.705	80.675	78.015	73.655	47.265	89.775	59.465	67.905	84.155	75.755	38.265	91.725	60.72	59.583	84.875
67.03	54.07	88.33	79.452	43.252	93.142	52.422	83.452	79.662	81.072	33.212	93.342	63.722	56.532	87.562	79.122	40.502	93.052	70.68	53.528	88.757

DOI: 10.54254/2755-2721/2025.PO24674

5. Conclusion

In this paper, we propose a gradient matching-based strategy to enhance CLIP's adaptation capability in open-set unsupervised domain adaptation (UDA). By enforcing alignment of gradient directions between the source and target domains while applying a gradient norm penalty to prevent overfitting to domain-specific noise, this dual mechanism enables the model to acquire domain-robust features while preserving its inherent cross-modal discriminative ability. Experiments on open-set adaptation benchmarks validate the superiority of our approach, significantly outperforming existing CLIP adaptation baselines, achieving state-of-the-art performance, and maintaining model stability.

References

[1] P. W. Koh et al., “WILDS: a benchmark of in-the-wild distribution shifts, ” in Proceedings of the 38th International Conference on Machine Learning, PMLR, Jul. 2021, pp. 5637–5664. Accessed: May 30, 2025. [Online]. Available: https: //proceedings.mlr.press/v139/koh21a.html

[2] “Do we really have to filter out random noise in pre-training data for language models?, ” presented at the ACL Rolling Review - February 2025, May 2025. Accessed: May 30, 2025. [Online]. Available: https: //openreview.net/forum?id=mpnPR8YQ3d

[3] T. Fang, N. Lu, G. Niu, and M. Sugiyama, “Rethinking importance weighting for deep learning under distribution shift, ” in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2020, pp. 11996–12007. Accessed: May 30, 2025. [Online]. Available: https: //proceedings.neurips.cc/paper_files/paper/2020/hash/8b9e7ab295e87570551db122a04c6f7c-Abstract.html

[4] J. Li, Z. Yu, Z. Du, L. Zhu, and H. T. Shen, “A comprehensive survey on source-free domain adaptation, ” IEEE Trans. Pattern Anal. Mach. Intell., vol. 46, no. 8, pp. 5743–5762, Aug. 2024, doi: 10.1109/TPAMI.2024.3370978.

[5] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, “Analysis of representations for domain adaptation, ” in Advances in Neural Information Processing Systems, MIT Press, 2006. Accessed: May 30, 2025. [Online]. Available: https: //proceedings.neurips.cc/paper/2006/hash/b1b0432ceafb0ce714426e9114852ac7-Abstract.html

[6] J. Ru, J. Tian, C. Xiao, J. Li, and H. T. Shen, “Imbalanced open set domain adaptation via moving-threshold estimation and gradual alignment, ” IEEE Trans. Multimedia, vol. 26, pp. 2504–2514, 2024, doi: 10.1109/TMM.2023.3297768.

[7] K. Wu, J. Li, L. Meng, F. Li, and K. Lu, “Online adaptive fault diagnosis with test-time domain adaptation, ” IEEE Trans. Ind. Informat., vol. 21, no. 1, pp. 107–117, Jan. 2025, doi: 10.1109/TII.2024.3438240.

[8] “Cross-domain mutual information adversarial maximization, ” Eng. Appl. Artif. Intell., vol. 110, p. 104665, Apr. 2022, doi: 10.1016/j.engappai.2022.104665.

[9] Z. Lai et al., “PADCLIP: pseudo-labeling with adaptive debiasing in CLIP for unsupervised domain adaptation, ” presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 16155–16165. Accessed: May 30, 2025. [Online]. Available: https: //openaccess.thecvf.com/content/ICCV2023/html/Lai_PADCLIP_Pseudo-labeling_with_Adaptive_Debiasing_in_CLIP_for_Unsupervised_Domain_ICCV_2023_paper.html

[10] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or, “StyleGAN-NADA: CLIP-guided domain adaptation of image generators, ” ACM Trans, Graph, vol. 41, no. 4, p. 141: 1-141: 13, Jul. 2022, doi: 10.1145/3528223.3530164.

[11] Y. Yin et al., “ATRI: mitigating multilingual audio text retrieval inconsistencies by reducing data distribution errors, ” Feb. 22, 2025, arXiv: arXiv: 2502.14627. doi: 10.48550/arXiv.2502.14627.

[12] S. Fort, J. Ren, and B. Lakshminarayanan, “Exploring the limits of out-of-distribution detection, ” in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2021, pp. 7068–7081. Accessed: May 30, 2025. [Online]. Available: https: //proceedings.neurips.cc/paper_files/paper/2021/hash/3941c4358616274ac2436eacf67fae05-Abstract.html

[13] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models, ” Int. J. Comput. Vision, vol. 130, no. 9, pp. 2337–2348, Sep. 2022, doi: 10.1007/s11263-022-01653-1.

[14] X. Zhuang et al., “VARGPT-v1.1: Improve visual autoregressive large unified model via iterative instruction tuning and reinforcement learning, ” Apr. 03, 2025, arXiv: arXiv: 2504.02949. doi: 10.48550/arXiv.2504.02949.

DOI: 10.54254/2755-2721/2025.PO24674

[15] J. Li, D. Li, S. Savarese, and S. Hoi, “BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models, ” in Proceedings of the 40th International Conference on Machine Learning, PMLR, Jul. 2023, pp. 19730–19742. Accessed: May 30, 2025. [Online]. Available: https: //proceedings.mlr.press/v202/li23q.html
[16] X. Zhuang et al., “VARGPT: unified understanding and generation in a visual autoregressive multimodal large language model, ” Jan. 21, 2025, arXiv: arXiv: 2501.12327. doi: 10.48550/arXiv.2501.12327.
[17] A. Radford et al., “Learning transferable visual models from natural language supervision, ” in Proceedings of the 38th International Conference on Machine Learning, PMLR, Jul. 2021, pp. 8748–8763. Accessed: May 30, 2025. [Online]. Available: https: //proceedings.mlr.press/v139/radford21a.html
[18] “Domain adaptation via prompt learning | IEEE journals & magazine | IEEE xplore.” Accessed: May 30, 2025. [Online]. Available: https: //ieeexplore.ieee.org/abstract/document/10313995
[19] X. Zhuang, X. Cheng, and Y. Zou, “Towards explainable joint models via information theory for multiple intent detection and slot filling, ” Proc. AAAI Conf. Artif. Intell., vol. 38, no. 17, Art. no. 17, Mar. 2024, doi: 10.1609/aaai.v38i17.29953.
[20] G. Zara, S. Roy, P. Rota, and E. Ricci, “AutoLabel: CLIP-based framework for open-set video domain adaptation, ” presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 11504–11513. Accessed: May 30, 2025. [Online]. Available: https: //openaccess.thecvf.com/content/CVPR2023/html/Zara_AutoLabel_CLIP-Based_Framework_for_Open-Set_Video_Domain_Adaptation_CVPR_2023_paper.html
[21] M. Fahes, T.-H. Vu, A. Bursuc, P. Pérez, and R. de Charette, “PODA: prompt-driven zero-shot domain adaptation, ” presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 18623–18633. Accessed: May 30, 2025. [Online]. Available: https: //openaccess.thecvf.com/content/ICCV2023/html/Fahes_PODA_Prompt-driven_Zero-shot_Domain_Adaptation_ICCV_2023_paper.html
[22] A. Miyai, Q. Yu, G. Irie, and K. Aizawa, “LoCoOp: few-shot out-of-distribution detection via prompt learning, ” Adv. Neural Inf. Process. Syst., vol. 36, pp. 76298–76310, Dec. 2023.
[23] D. Hendrycks and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks, ” presented at the International Conference on Learning Representations, Feb. 2017. Accessed: May 14, 2025. [Online]. Available: https: //openreview.net/forum?id=Hkg4TI9xl
[24] S. Liang, Y. Li, and R. Srikant, “Enhancing the reliability of out-of-distribution image detection in neural networks, ” Aug. 30, 2020, arXiv: arXiv: 1706.02690. doi: 10.48550/arXiv.1706.02690.
[25] D. Hendrycks, M. Mazeika, and T. Dietterich, “Deep anomaly detection with outlier exposure, ” presented at the International Conference on Learning Representations, Sep. 2018. Accessed: May 14, 2025. [Online]. Available: https: //openreview.net/forum?id=HyxCxhRcY7
[26] K. Lee, K. Lee, H. Lee, and J. Shin, “A simple unified framework for detecting out-of-distribution samples and adversarial attacks, ” in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2018. Accessed: May 14, 2025. [Online]. Available: https: //proceedings.neurips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html
[27] W. Liu, X. Wang, J. D. Owens, and Y. Li, “Energy-based out-of-distribution detection, ” Apr. 26, 2021, arXiv: arXiv: 2010.03759. doi: 10.48550/arXiv.2010.03759.
[28] Z. Lin, S. D. Roy, and Y. Li, “MOOD: multi-level out-of-distribution detection”, doi: 10.1109/CVPR46437.2021.01506.
[29] Y. Sun and Y. Li, “DICE: leveraging sparsification for out-of-distribution detection, ” in Computer Vision – ECCV 2022, vol. 13684, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., in Lecture Notes in Computer Science, vol. 13684. , Cham: Springer Nature Switzerland, 2022, pp. 691–708. doi: 10.1007/978-3-031-20053-3_40.
[30] A. Radford et al., “Learning transferable visual models from natural language supervision, ” in Proceedings of the 38th International Conference on Machine Learning, PMLR, Jul. 2021, pp. 8748–8763. Accessed: May 14, 2025. [Online]. Available: https: //proceedings.mlr.press/v139/radford21a.html
[31] J. Yang et al., “OpenOOD: Benchmarking generalized out-of-distribution detection, ” presented at the Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, Jun. 2022. Accessed: May 14, 2025. [Online]. Available: https: //openreview.net/forum?id=gT6j4_tskUt
[32] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, “AnomalyCLIP: object-agnostic prompt learning for zero-shot anomaly detection, ” presented at the The Twelfth International Conference on Learning Representations, Oct. 2023. Accessed: May 14, 2025. [Online]. Available: https: //openreview.net/forum?id=buC4E91xZE

DOI: 10.54254/2755-2721/2025.PO24674

[33] P. Panareda Busto and J. Gall, “Open set domain adaptation, ” presented at the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 754–763. Accessed: May 14, 2025. [Online]. Available: https: //openaccess.thecvf.com/content_iccv_2017/html/Busto_Open_Set_Domain_ICCV_2017_paper.html
[34] K. Saito, S. Yamamoto, Y. Ushiku, and T. Harada, “Open set domain adaptation by backpropagation, ” presented at the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 153–168. Accessed: May 14, 2025. [Online]. Available: https: //openaccess.thecvf.com/content_ECCV_2018/html/Kuniaki_Saito_Adversarial_Open_Set_ECCV_2018_paper.html
[35] “Open set domain adaptation: theoretical bound and algorithm | IEEE journals & magazine | IEEE xplore.” Accessed: May 14, 2025. [Online]. Available: https: //ieeexplore.ieee.org/abstract/document/9186366
[36] P. Panareda Busto, A. Iqbal, and J. Gall, “Open set domain adaptation for image and action recognition, ” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 413–429, Feb. 2020, doi: 10.1109/TPAMI.2018.2880750.
[37] Y. Xu and D. Klabjan, “Open set domain adaptation by extreme value theory, ” Dec. 22, 2020, arXiv: arXiv: 2101.02561. doi: 10.48550/arXiv.2101.02561.
[38] “Discovering implicit classes achieves open set domain adaptation | IEEE conference publication | IEEE xplore.” Accessed: May 14, 2025. [Online]. Available: https: //ieeexplore.ieee.org/abstract/document/9859594
[39] S.-A. Choe, A.-H. Shin, K.-H. Park, J. Choi, and G.-M. Park, “Open-set domain adaptation for semantic segmentation, ” presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 23943–23953. Accessed: May 14, 2025. [Online]. Available: https: //openaccess.thecvf.com/content/CVPR2024/html/Choe_Open-Set_Domain_Adaptation_for_Semantic_Segmentation_CVPR_2024_paper.html
[40] A. Vaswani et al., “Attention is all you need, ” Aug. 02, 2023, arXiv: arXiv: 1706.03762. doi: 10.48550/arXiv.1706.03762.
[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding, ” May 24, 2019, arXiv: arXiv: 1810.04805. doi: 10.48550/arXiv.1810.04805.
[42] F. Yu et al., “ERNIE-ViL: knowledge enhanced vision-language representations through scene graph, ” Mar. 19, 2021, arXiv: arXiv: 2006.16934. doi: 10.48550/arXiv.2006.16934.
[43] H. Luo et al., “UniVL: a unified video and language pre-training model for multimodal understanding and generation, ” Sep. 15, 2020, arXiv: arXiv: 2002.06353. doi: 10.48550/arXiv.2002.06353.
[44] A. Radford et al., “Learning transferable visual models from natural language supervision, ” Feb. 26, 2021, arXiv: arXiv: 2103.00020. doi: 10.48550/arXiv.2103.00020.
[45] A. Ramesh et al., “Zero-shot text-to-image generation, ” in Proceedings of the 38th International Conference on Machine Learning, PMLR, Jul. 2021, pp. 8821–8831. Accessed: May 14, 2025. [Online]. Available: https: //proceedings.mlr.press/v139/ramesh21a.html
[46] J. Li, D. Li, C. Xiong, and S. Hoi, “BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation, ” Feb. 15, 2022, arXiv: arXiv: 2201.12086. doi: 10.48550/arXiv.2201.12086.
[47] P. Wang et al., “OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, ” Jun. 01, 2022, arXiv: arXiv: 2202.03052. doi: 10.48550/arXiv.2202.03052.
[48] J. Li, D. Li, S. Savarese, and S. Hoi, “BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models, ” Jun. 15, 2023, arXiv: arXiv: 2301.12597. doi: 10.48550/arXiv.2301.12597.