Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

GeoRel-CLIP: Boosting CLIP Zero-Shot Recognition via Geometry-Regularized Relational Distillation

Huadong Sun
Harbin University of Commerce
Yexuan Zhao
zhaoyexuan0216@126.com
Harbin University of Commerce
Junwei Tong
Harbin University of Commerce

Research Article

Keywords: Vision-Language Models, CLIP, Zero-Shot Learning, Post-Pretraining, Geometric Distribution Regularization, Relational Distillation

Posted Date: August 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9878849/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

GeoRel-CLIP: Boosting CLIP Zero-Shot Recognition via Geometry-Regularized Relational Distillation

Huadong Sun[1,2], Yexuan Zhao[1*], Junwei Tong[1]

[1*]School of Computer and Information Engineering, Harbin University of Commerce.

[2]Heilongjiang Provincial Key Laboratory of Electronic Commerce and Information Processing, Harbin University of Commerce, Harbin, 150028, China.

*Corresponding author(s). E-mail(s): zhaoyexuan0216@126.com;

Abstract

Vision-language models (VLMs) have demonstrated strong zero-shot learning (ZSL) capability in open-category recognition through joint image-text modeling. However, existing approaches for improving zero-shot performance often rely on large-scale pretraining, prompt learning, or lightweight adaptation, which involve high training costs, strong task dependence, or insufficient optimization of the shared representation space. In particular, during post-pretraining, the modality gap, distributional degradation, and representation collapse may emerge simultaneously, limiting further improvements in zero-shot classification performance. To address these issues, this paper proposes GeoRel-CLIP, a geometry-regularized relational distillation method for CLIP post-pretraining. Instead of solely reducing the distance between image and text features, GeoRel-CLIP jointly optimizes the shared embedding space from two perspectives: global geometric distribution and local relational structure. Specifically, we introduce a Geometric Distribution Regularization (GDR) module to regulate feature spreading and directional occupancy balance on the unit hypersphere, thereby suppressing excessive feature aggregation and representation collapse. Meanwhile, we further propose a Scale-Invariant Relational Distillation (SIRD) module, which standardizes and aligns the cross-modal relational matrices of the teacher and student models, reducing the interference caused by scale drift and enhancing the consistency of structural knowledge transfer. Experimental results show that GeoRel-CLIP outperforms existing SOTA methods on multiple standard zero-shot datasets and achieves further improvement in the average Top-1 accuracy across 11 datasets.

Keywords: Vision-Language Models; CLIP; Zero-Shot Learning; Post-Pretraining; Geometric Distribution Regularization; Relational Distillation.

1 Introduction

In recent years, the rapid development of vision-language models (VLMs) has significantly advanced research in open-vocabulary visual understanding. Representative models such as CLIP [1] and ALIGN [2] learn a unified cross-modal semantic embedding space through contrastive learning on large-scale image-text pairs. This enables models to perform recognition, retrieval, and other tasks directly with natural-language category descriptions, without additional training on target tasks. Despite these advances, the shared embedding spaces learned by these VLMs still exhibit structural limitations that restrict zero-shot classification performance. Previous studies show that image and text representations are often misaligned in terms of geometric distribution and semantic organization [3–5], weakening the reliability of cross-modal similarity estimation and downstream zero-shot generalization.

Recent studies on improving the zero-shot classification performance of vision-language models (VLMs) can be broadly categorized into three perspectives: large-scale pretraining, lightweight adaptation, and modality calibration. Large-scale vision-language pretraining methods, such as CLIP [1], ALIGN [2], SigLIP [6], improve cross-modal representation through enhanced training objectives, architectures, and larger image-text datasets. However, these approaches typically require extensive computational resources and large-scale retraining, limiting their practicality for efficient post-hoc calibration of pretrained models. Lightweight adaptation methods, including CoOp [7], CoCoOp [8], CLIP-Adapter [9], and Tip-Adapter [10], improve downstream performance through prompt learning or lightweight parameter updates. Although effective in few-shot or dataset-specific scenarios, excessive adaptation may introduce representation shift, calibration degradation, and knowledge forgetting, thereby weakening the model's general zero-shot capability [11–14].

To overcome these limitations, recent studies have explored modality calibration and post-pretraining strategies that directly optimize the shared embedding space of VLMs. Existing methods mainly focus on improving modality alignment or output consistency through post-pretraining and knowledge distillation [11, 13–15]. However, simply reducing inter-modal discrepancy does not necessarily produce a well-structured representation space. During optimization, the model may still suffer from representation collapse, anisotropy, and geometric degradation [3, 4, 16–19]. As illustrated in the left part of Figure 1, image and text embeddings may occupy different regions of the hyperspherical feature space, resulting in unreliable cross-modal similarity estimation. Meanwhile, the right part of Figure 1 shows that excessive feature alignment can lead to representation collapse, where features become overly concentrated and lose semantic discriminability. These observations motivate the proposed method, which jointly considers modality calibration, geometric regularization, and relational structure preservation during post-pretraining.

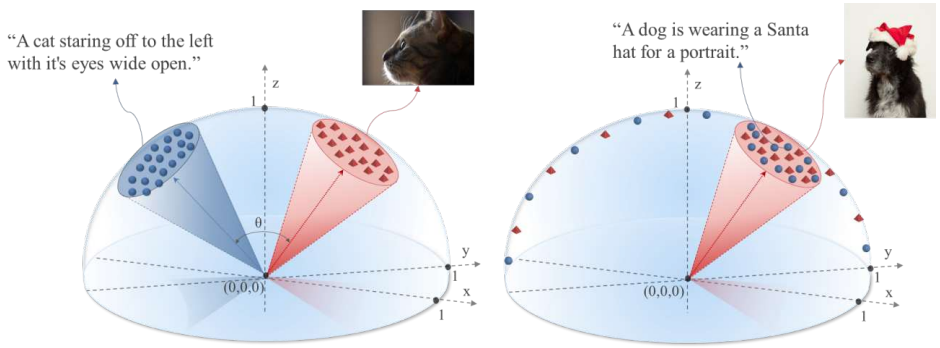


Fig. 1: Illustration of two typical distributional problems in cross-modal alignment.

To this end, we propose GeoRel-CLIP, a geometry-regularized relational distillation framework for vision-language post-pretraining. Instead of solely reducing image-text feature distances, GeoRel-CLIP jointly optimizes the shared embedding space during the post-pretraining process from two complementary perspectives: the global geometric configuration and the local semantic topology. Specifically, we introduce Geometric Distribution Regularization (GDR), which regulates the overall spreading pattern and directional occupancy balance of joint features on the unit hypersphere, thereby suppressing excessive feature aggregation, increased anisotropy, and representation collapse. We further propose Scale-Invariant Relational Distillation (SIRD), which aligns standardized cross-modal relational matrices between the teacher and student models, thereby reducing the interference caused by scale drift and encouraging the model to focus on the consistency of relative relational structures among samples. Through the synergistic effect of these two modules, the model can reduce the modality gap while more effectively preserving the original semantic organization and discriminative capability of the pretrained model.

Extensive experiments on 11 standard zero-shot classification datasets verified the effectiveness of the proposed method. Using CLIP-ViT-B/32 as the backbone, our GeoRel-CLIP outperforms other representative comparison methods in average Top-1 accuracy, while also achieving more balanced improvements in modality gap, feature uniformity, and hyperspherical distribution quality.

The main contributions of this paper are summarized as follows:

1. We propose **GeoRel-CLIP,** a collaborative post-pretraining framework that jointly optimizes geometric distribution and relational structure for improving zero-shot vision-language learning.
2. We introduce two complementary modules: **Geometric Distribution Regularization (GDR)**, which improves feature spreading and directional balance on the unit hypersphere, and **Scale-Invariant Relational Distillation (SIRD)**, which preserves cross-modal relational consistency through standardized relational alignment.
3. Extensive experiments on 11 standard zero-shot classification benchmarks demonstrate that GeoRel-CLIP consistently improves both representation geometry

and downstream zero-shot classification performance, outperforming representative state-of-the-art post-pretraining methods.

2 Related Work

2.1 Vision-Language Pretraining for Zero-Shot Learning

The zero-shot capability of vision–language models largely relies on large-scale image–text pretraining. CLIP [1] learns a unified cross-modal embedding space by contrastive learning on massive image–text pairs, enabling open-vocabulary classification and retrieval. ALIGN [2] further improves the feature representation using larger-scale, noisy image–text datasets. Subsequent works, including FLAVA [20], ALBEF [21], BLIP/BLIP-2 [22, 23] and ImageBind [24], improve multimodal representation learning through multi-task training, semantic fusion, lightweight bridging modules, and broader modality integration.

Besides, large-scale datasets and optimization strategies are also important for zero-shot learning. LAION-5B [25] and DataComp [26] highlight the importance of data scale and quality, while SigLIP [6], LiT [27], CoCa [28], and PaLI [29] improve training objectives and learning paradigms. Although these methods substantially improve representation quality, they typically require extensive computational resources and large-scale retraining, limiting their applicability to existing pre-trained models. Therefore, efficient post-pretraining calibration remains an important direction for improving zero-shot performance [3–5].

2.2 Prompt Learning and Lightweight Adaptation

Prompting and lightweight adaptation techniques offer a practical alternative to improve zero-shot classification without modifying the pre-trained backbone. CoOp [7] and CoCoOp [8] optimize learnable prompts for downstream tasks, while CLIP-Adapter [9] and Tip-Adapter [10] introduce lightweight feature adaptation mechanisms. Test-Time Prompt Tuning [30] further adapts prompts during inference, and Troika [31] improves compositional zero-shot learning through cross-modal semantic guidance.

These methods achieve strong efficiency-performance trade-offs, but their improvements are often task-dependent and mainly enhance task adaptation rather than the structural quality of the shared embedding space [7–10, 30, 32]. Moreover, adaptation may suffer knowledge forgetting and reduced cross-domain generalization [11–14], leaving structural imbalance in cross-modal representations insufficiently addressed [5, 16, 19].

2.3 Modality Calibration and Relational Distillation

Recent studies increasingly attribute limited zero-shot performance to structural mismatch between image and text representations, particularly modality gaps. Mind the Gap [5] reveals systematic distributional offsets between modalities, while subsequent studies analyze inconsistencies in local structure, global distribution, and open-world

generalization [15–19, 33]. Therefore, different modality calibration and relational distillation methods have been proposed.

Compared to prompt- or adapter-based methods, modality calibration methods directly optimize the shared embedding space [5, 33, 34]. However, existing works mainly focus on reducing average inter-modal distances or mitigating shifts, while paying limited attention to maintaining the overall geometric structure of the space [5, 16, 19]. Reducing modality gaps alone cannot ensure uniformity, separability, and effective representation utilization, since over-concentration along certain directions may still hinder discriminative performance [3, 4, 16, 19].

Post-pretraining (intermediate calibration) strategies offer a cost-effective pathway. By using smaller but higher-quality image–text datasets, pre-trained models can be refined without full retraining [15]. Meanwhile, teacher–student framework-based knowledge distillation can further preserve original knowledge. Studies [35, 36] employ structural or relational distillation, maintaining relative relationships to mitigate performance degradation. Nevertheless, current post-pretraining and distillation methods primarily focus on output consistency, modality alignment, or knowledge retention, while relational scale drift, statistical imbalance, and geometric degradation in the shared embedding space remain underexplored [14, 15, 35–38].

3 Method

Figure 2 illustrates the overall framework of GeoRel-CLIP for representation alignment during vision-language post-pretraining. The student model is initialized with pretrained CLIP weights, while a frozen copy serves as the teacher model. For each training batch, image and text features are extracted by both models, followed by L_2-normalization and projection onto the unit hypersphere. The normalized student features are first optimized using a symmetric contrastive loss to maintain basic cross-modal alignment. On this basis, the proposed Geometric Distribution Regularization (GDR) module regularizes the global spreading behavior and directional occupancy balance of joint representations, alleviating feature over-concentration and directional bias. Meanwhile, the Scale-Invariant Relational Distillation (SIRD) module aligns standardized cross-modal similarity matrices between the teacher and student models, preserving the semantic structures encoded in the pretrained model. Finally, GeoRel-CLIP jointly optimizes the contrastive loss, GDR loss, and SIRD loss, enabling simultaneous cross-modal alignment, geometric regularization, and relational structure preservation during post-pretraining.

3.1 Preliminaries and Notation

Let the image encoder and text encoder be denoted by $f_v(\cdot;\phi)$ and $f_t(\cdot;\phi)$, respectively, where ϕ represents the parameters to be optimized. During post-pretraining, the model is initialized with pretrained weights. Given a mini-batch containing B paired samples $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{B}$, we compute the corresponding feature vectors and project them

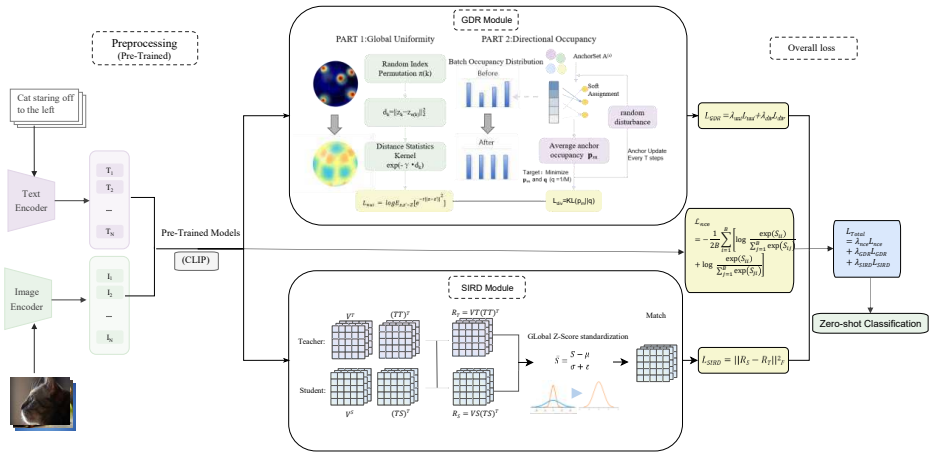


Fig. 2: Overall framework of GeoRel-CLIP.

onto the unit hypersphere $\mathbb{S}^{d-1}$ via L_2 normalization:

$$v_i = \frac{f_v(x_i;\phi)}{\|f_v(x_i;\phi)\|_2}, \quad t_i = \frac{f_t(y_i;\phi)}{\|f_t(y_i;\phi)\|_2}. \tag{1}$$

The joint feature set of the current batch is defined as

$$\mathcal{Z} = \{z_k\}_{k=1}^{2B} = \{v_1, \ldots, v_B, t_1, \ldots, t_B\}. \tag{2}$$

Cross-modal alignment is optimized using a symmetric contrastive loss. Let the similarity matrix $\mathbf{S} \in \mathbb{R}^{B\times B}$ be defined as

$$S_{ij} = v_i^\top t_j / \tau_c. \tag{3}$$

The contrastive loss is formulated as

$$\mathcal{L}_{\text{nce}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(S_{ii})}{\sum_{j=1}^{B}\exp(S_{ij})} + \log\frac{\exp(S_{ii})}{\sum_{j=1}^{B}\exp(S_{ji})}\right]. \tag{4}$$

3.2 GDR: Geometric Distribution Regularization

To alleviate feature over-concentration and directional bias on the hypersphere, we propose the GDR constraint. The GDR is composed of a global uniformity term and an anchor-based directional balancing term. Intuitively, GDR directly regularizes the spatial distribution of $\mathcal{Z}$. Specifically, the uniformity term suppresses local aggregation through distance statistics, while the directional balancing term prevents a small

number of directional prototypes from dominating the feature space by encouraging balanced directional occupancy. In this way, GDR improves the global geometric usability of joint representations without modifying the original alignment objective.

3.2.1 Global Uniformity Regularization

We adopt a distance-based kernelized statistic to characterize the degree of local neighborhood aggregation among joint features. In each training iteration, the indices of the joint features in the current batch are randomly shuffled to obtain a mapping π. Based on this mapping, the squared Euclidean distance between randomly paired features is defined as

$$d_k = \|z_k - z_{\pi(k)}\|_2^2. \tag{5}$$

The global uniformity loss is defined as

$$\mathcal{L}_{\text{uni}} = \log \frac{1}{B} \sum_{k=1}^{B} e^{-t\, d_k}, \tag{6}$$

where t is the scale coefficient. By penalizing overly close sample pairs, this term suppresses local aggregation and promotes a dispersed distribution of features on $\mathbb{S}^{d-1}$.

3.2.2 Directional Occupancy Regularization with Anchors

To further quantify and balance the occupancy of joint features over different directional prototypes, we introduce a set of predefined normalized anchors. Let the training iteration step be denoted by s, and the anchor set be written as

$$\mathcal{A}_s = \{a_m^s\}_{m=1}^M \subset \mathbb{S}^{d-1}, \tag{7}$$

where $a_m^s \in \mathbb{R}^d$ denotes the m-th anchor vector at the s-th iteration, and $\|a_m^s\|_2 = 1$. Therefore, $a_m^s \in \mathbb{S}^{d-1}$. The initial anchor set $\mathcal{A}_0$ is generated from random vectors and then L_2-normalized. The anchors do not participate in gradient updates. Instead, every T_{rot} steps, they are randomly perturbed without gradient tracking and then normalized:

$$a_m^{s+1} = \begin{cases} \dfrac{a_m^s + \xi_m^s}{\|a_m^s + \xi_m^s\|_2}, & s \bmod T_{\text{rot}} = 0, \\ a_m^s, & \text{otherwise}, \end{cases} \tag{8}$$

where the random perturbation vector applied to the m-th anchor is $\xi_m^s \in \mathbb{R}^d$ satisfying $\xi_m^s \sim \mathcal{N}(0, \sigma_a^2 \mathbf{I}_d)$, where T_{rot} is the update period, σ_a is the perturbation strength, and $\mathbf{I}_d$ is the $d \times d$ identity matrix.

The soft assignment probability from feature z to an anchor is defined as

$$p_m(z; \mathcal{A}_s) = \frac{\exp\big(z^\top a_m^s / \tau\big)}{\sum_{m'=1}^M \exp(z^\top a_{m'}^s / \tau)}. \tag{9}$$

The average anchor occupancy at the batch level is defined as

$$\bar{p}_m = \frac{1}{2B} \sum_{z \in \mathcal{Z}} p_m(z; \mathcal{A}_s). \tag{10}$$

Directional occupancy is balanced by minimizing the KL divergence between $\bar{\mathbf{p}}$ and the uniform distribution $\mathbf{q}$, where $q_m = 1/M$:

$$\mathcal{L}_{\text{div}} = \text{KL}(\bar{\mathbf{p}} \| \mathbf{q}) = \sum_{m=1}^{M} \bar{p}_m \log \frac{\bar{p}_m}{q_m}. \tag{11}$$

Based on the above terms, the overall GDR regularization term is defined as

$$\mathcal{L}_{\text{GDR}} = \lambda_{\text{uni}} \mathcal{L}_{\text{uni}} + \lambda_{\text{div}} \mathcal{L}_{\text{div}}. \tag{12}$$

where $\lambda_{\text{uni}} \geq 0$ and $\lambda_{\text{div}} \geq 0$ are hyperparameters controlling the strengths of the two constraints.

3.3 SIRD: Scale-Invariant Relational Distillation

Scale-Invariant Relational Distillation (SIRD) is used to preserve the cross-modal relational structure of the teacher model during post-pretraining. Given a mini-batch, both the teacher model and the student model construct image-to-text (I2T) and text-to-image (T2I) similarity matrices to characterize cross-modal semantic relations among samples. However, these relational matrices are sensitive to variations in logit scale and similarity distribution during continued training, which may weaken the effectiveness of relational distillation. To alleviate this issue, SIRD first applies global Z-score standardization to the similarity matrices. The teacher and student relational matrices are then aligned in the standardized relational space using mean squared error (MSE), encouraging the student model to focus on relative semantic structures rather than absolute similarity values. The computation details are as follows.

The student and teacher models first extract features from the input images and texts using their respective encoders. Based on these extracted features, the cross-modal similarity matrices are computed as

$$\mathbf{R}^T = \mathbf{V}^T (\mathbf{T}^T)^\top, \quad \mathbf{R}^S = \mathbf{V}^S (\mathbf{T}^S)^\top, \tag{13}$$

where $\mathbf{V}^T$ and $\mathbf{T}^T$ denote the image and text feature matrices of the teacher model, respectively, while $\mathbf{V}^S$ and $\mathbf{T}^S$ denote the image and text feature matrices of the student model, respectively. Each element R_{ij} in the similarity matrices $\mathbf{R}^T$ and $\mathbf{R}^S$ represents the similarity between the i-th image and the j-th text.

To eliminate the interference caused by scale differences between the teacher and student models, we apply global Z-score standardization to the similarity matrices $\mathbf{R}^T$

and $\mathbf{R}^S$, denoted as follows:

$$\tilde{\mathbf{R}}^T = \frac{\mathbf{R}^T - \mu(\mathbf{R}^T)}{\sigma(\mathbf{R}^T) + \epsilon}, \quad \tilde{\mathbf{R}}^S = \frac{\mathbf{R}^S - \mu(\mathbf{R}^S)}{\sigma(\mathbf{R}^S) + \epsilon}, \tag{14}$$

where $\mu(\mathbf{R}^T)$ and $\mu(\mathbf{R}^S)$ denote the means of the similarity matrices $\mathbf{R}^T$ and $\mathbf{R}^S$, respectively; $\sigma(\mathbf{R}^T)$ and $\sigma(\mathbf{R}^S)$ denote their standard deviations; and ϵ is a small constant used to prevent division by zero.

Finally, the loss function is obtained by computing the Frobenius norm between the standardized similarity matrices:

$$\mathcal{L}_{\text{SIRD}} = \left\| \tilde{\mathbf{R}}^S - \tilde{\mathbf{R}}^T \right\|_F^2. \tag{15}$$

3.4 Optimization Objective and Training Procedure

Based on the above modules, the overall training objective of GeoRel-CLIP consists of three components: the symmetric contrastive loss, the geometric distribution regularization loss, and the scale-invariant relational distillation loss.

Accordingly, the overall objective function for model post-pretraining is defined as

$$\mathcal{L} = \lambda_{\text{nce}} \mathcal{L}_{\text{nce}} + \mathcal{L}_{\text{GDR}} + \lambda_{\text{SIRD}} \mathcal{L}_{\text{SIRD}}. \tag{16}$$

where λ_{nce} and λ_{SIRD} are hyperparameters controlling the strengths of the contrastive learning term and the distillation term, respectively.

In addition, Algorithm 1 summarizes the overall post-pretraining procedure of GeoRel-CLIP. For each training batch, the student model and the frozen teacher model extract image and text features, respectively. The symmetric contrastive loss, the geometric distribution regularization loss, and the scale-invariant relational distillation loss are then computed sequentially. Finally, the student model parameters are updated through joint optimization of all loss terms.

4 Experiments

We evaluate the proposed method on multiple zero-shot classification tasks. In addition, we conduct both quantitative and qualitative analyses of the feature space from the perspectives of modality gap, alignment and uniformity metrics, and PCA visualization.

4.1 Experimental Setup

4.1.1 Dataset

To comprehensively evaluate the generalization ability of the proposed method, we conduct zero-shot transfer evaluation on 11 standard image classification datasets. These datasets cover diverse distributional characteristics and are widely used to assess the zero-shot transfer performance of CLIP-based models [1], including fine-grained

Algorithm 1 Overall post-pretraining procedure of GeoRel-CLIP

Require: Training dataset $\mathcal{D}$, pretrained student parameters (θ_V^p, θ_T^p), frozen teacher parameters (θ_V^t, θ_T^t), anchor set $\mathcal{A}$, batch size B, learning rate η, loss weights λ_{uni}, λ_{div}, and λ_{SIRD}

Ensure: Optimized student parameters (θ_V, θ_T)

Initialize student parameters: $\theta_V \leftarrow \theta_V^p$, $\theta_T \leftarrow \theta_T^p$
while the model has not converged **do**
  Sample a mini-batch $\{(x_i, t_i)\}_{i=1}^B$ from $\mathcal{D}$
  Compute student image and text features $\{z_i^{\text{img}}, z_i^{\text{txt}}\}_{i=1}^B$
  Compute teacher image and text features $\{z_i^{\text{img},t}, z_i^{\text{txt},t}\}_{i=1}^B$
  Apply L_2 normalization to all features
  Construct student cross-modal similarity matrix and compute $\mathcal{L}_{\text{nce}}$
  Construct joint feature set $\mathcal{Z}$ and compute $\mathcal{L}_{\text{uni}}$
  if the anchor update condition is satisfied **then**
    Apply random perturbation to anchor set $\mathcal{A}$ and normalize anchors
  end if
  Compute soft assignment probabilities from joint features to anchors
  Estimate average anchor occupancy and compute $\mathcal{L}_{\text{div}}$
  Construct teacher cross-modal similarity matrix
  Apply global Z-score standardization to student and teacher relational matrices
  Compute $\mathcal{L}_{\text{SIRD}}$
  Compute total loss: $\mathcal{L} = \lambda_{\text{nce}}\mathcal{L}_{\text{nce}} + \mathcal{L}_{\text{GDR}} + \lambda_{\text{SIRD}}\mathcal{L}_{\text{SIRD}}$
  Backpropagate $\mathcal{L}$ and update (θ_V, θ_T)
end while
return (θ_V, θ_T)

classification datasets such as Aircraft (Air) [39], Bird [40], Car [41], Flower (Flo) [42], Food [43], and Pet [44]; material and texture recognition datasets such as DTD [45]; remote sensing land-cover recognition datasets such as EuroSAT (Euro) [46]; and general scene and object recognition datasets including Caltech-101 (Cal) [47], ImageNet (IN) [48], and SUN397 [49]. The evaluation protocol follows the official split of each dataset.

4.1.2 Implementation Details

Pretrained backbone and initialization: We adopt CLIP-ViT-B/32 [1] as the default backbone network for both the visual encoder and the text encoder, and initialize the model with the official pretrained weights provided by OpenAI.

Post-pretraining data and training details: In the post-pretraining stage, we use COCO Caption [50] image–text pairs as the training data. The model is trained only for one epoch using the AdamW optimizer [51], with a learning rate of 1.0×10^{-6} and a default mini-batch size of 512. All experiments are conducted on a single NVIDIA H20 GPU.

Testing and evaluation: During feature computation and geometric metric evaluation, we strictly follow the standard CLIP processing protocol [1] and apply L_2-normalization to the output feature vectors, thereby projecting them onto the unit hypersphere. This operation ensures that all geometric measurements and similarity statistics are computed in a unified scale space, avoiding the interference of norm drift in the evaluation metrics.

4.1.3 Metrics

Zero-shot transfer performance metric: In zero-shot classification tasks, we report Top-1 Accuracy as the primary metric for evaluating the model's generalization performance.

Representation geometry metrics: To quantitatively characterize the geometric topology of post-pretrained representations in the shared embedding space, we introduce three metrics: Modality Gap, Alignment, and Uniformity.

Let the visual feature be $f_V(x_i)$, the text feature be $f_T(t_i)$, and the total number of evaluation samples be N_{test}.

Modality Gap [5] measures the geometric separation between the centers of the visual and textual modalities in the embedding space:

$$\text{Modality Gap} = \left\| \bar{f}_V - \bar{f}_T \right\|_2^2, \tag{17}$$

where $\bar{f}_\cdot$ denotes the mean feature vector of the corresponding modality over all evaluation samples. A smaller value is preferred, indicating reduced inter-modal bias.

Alignment [3] measures the compactness, or degree of alignment, of matched image–text pairs in the feature space:

$$\text{Alignment} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \left\| f_V(x_i) - f_T(t_i) \right\|_2^2, \tag{18}$$

A smaller value is preferred, indicating that matched image–text pairs are closer in the embedding space and exhibit stronger semantic consistency.

Uniformity [3] measures the distributional uniformity of features on the unit hypersphere based on the Gaussian kernel potential:

$$\text{Uniformity} = \log \frac{1}{2N_{\text{test}}^2} \sum_{\substack{f_1, f_2 \in \mathcal{F} \\ f_1 \neq f_2}} \exp\Big(-2 \left\| f_1 - f_2 \right\|_2^2 \Big), \tag{19}$$

where the feature set is defined as $\mathcal{F} = \{f_V(x_i)\}_{i=1}^{N_{\text{test}}} \cup \{f_T(t_i)\}_{i=1}^{N_{\text{test}}}$. A smaller value is preferred, indicating that the feature distribution is more uniform.

To reduce randomness, all experiments are repeated three times, and the average metric values are reported.

4.2 Comparison with SOTA Methods

To verify the effectiveness of the proposed method, we compare GeoRel-CLIP with several representative methods, including Pre-trained CLIP [1], which directly uses pretrained weights for prediction; the Contrastive method [2], which performs post-pretraining using only the contrastive loss; m2-mix [52], which combines a modified contrastive loss with multimodal mixup; Self-KD [15], which only adopts self-distillation constraints; and CLIP-Refine [15], which introduces random feature alignment and mixed contrastive distillation.

4.2.1 Comparison of Zero-Shot Classification Accuracy

Table 1 presents the zero-shot classification results of different methods on 11 downstream datasets.

The results show that post-pretraining with only the contrastive loss does not lead to stable improvements, and even causes a decrease in Top-1 accuracy on several datasets. This is likely due to the much smaller batch size used in post-pretraining compared with the original CLIP training stage (512 vs. 32,768). With fewer negative samples, contrastive learning becomes more prone to overfitting and may weaken the general knowledge learned during pretraining.

In contrast, Self-KD preserves the output distribution of the pretrained model through distillation constraints and achieves better overall performance than the contrastive-only method, indicating the importance of maintaining the original knowledge during post-pretraining. CLIP-Refine further introduces random feature alignment and mixed contrastive distillation, achieving improvements in reducing the modality gap and maintaining feature uniformity.

Compared with the above methods, GeoRel-CLIP achieves better zero-shot classification performance on most datasets and obtains the best average Top-1 accuracy across the 11 datasets. In particular, the proposed method shows more evident advantages on DTD, Flower, ImageNet, and SUN397. These results suggest that effective post-pretraining cannot rely solely on the contrastive objective, nor be limited to simple knowledge preservation or modality alignment. Instead, it requires a more balanced synergy among knowledge preservation, modality alignment, and spatial structure regularization.

4.2.2 Comparison of Geometric Metrics in the Shared Embedding Space

To further analyze the influence of post-pretraining on the structure of the shared embedding space, Table 2 reports the results of different methods in terms of three geometric metrics.

From the quantitative results, post-training based solely on the contrastive objective does not bring desirable geometric improvements, and some metrics even deteriorate. This indicates that directly reusing the contrastive optimization objective from the pretraining stage is insufficient for stable calibration of the shared space. In contrast, GeoRel-CLIP exhibits a more coordinated optimization trend across all three metrics. Specifically, the Modality Gap is substantially reduced, indicating that

Table 1: Zero-shot classification accuracy (%) of different methods on 11 datasets

Method	Air	Bird	Cal	Car	DTD	Euro	Flo	Food	IN	Pet	SUN	Avg.
Pre-trained [1]	20.34	51.95	85.51	60.03	38.67	38.19	59.72	80.20	59.03	81.79	60.89	57.85
Contrastive [2]	15.48	42.75	85.51	50.32	36.49	33.86	47.28	73.32	54.13	70.78	54.79	51.33
m2-mix [52]	14.64	40.76	23.49	49.86	36.49	34.65	47.42	71.27	53.58	70.32	56.51	46.48
Self-KD [15]	18.96	51.52	23.94	60.04	41.17	38.36	59.55	80.02	59.06	81.85	60.68	52.94
CLIP-Refine [15]	20.74	50.59	86.20	59.95	41.49	40.33	61.37	80.39	59.94	82.91	63.37	58.84
GeoRel-CLIP (Ours)	**21.24**	**52.54**	**89.32**	**60.37**	**45.32**	**40.87**	**64.29**	**82.72**	**62.70**	**84.62**	**65.06**	**60.82**

the global center shift between the image and text modalities is effectively alleviated. Uniformity is also significantly improved, suggesting that the joint representations are more evenly distributed on the unit hypersphere, while local crowding and directional collapse are suppressed. Meanwhile, Alignment remains generally stable, showing that such geometric correction is not achieved at the expense of weakening the semantic consistency between matched image–text pairs.

Overall, the performance improvement brought by GeoRel-CLIP does not merely come from local contraction of image–text feature distances, but rather reflects an improvement in the overall structural quality of the shared embedding space. In other words, effective optimization during post-pretraining requires not only reducing the systematic shift between modalities, but also maintaining a uniformly spread distribution of joint features. Only when modality calibration and geometric distribution optimization are simultaneously improved can the enhancement of zero-shot transfer performance be supported by a more reliable representational foundation.

Table 2: Geometric diagnostic results of the shared embedding space for different methods

Method	Modality Gap (↓)	Alignment (↓)	Uniformity (↓)
Pre-trained [1]	0.6805	1.3902	208.8000
Contrastive [2]	0.8396	1.3946	239.9639
m2-mix [52]	0.9971	1.3928	230.7469
Self-KD [15]	0.6804	1.3897	208.1625
CLIP-Refine [15]	0.2413	1.4065	183.2942
GeoRel-CLIP (Ours)	**0.1645**	**1.3954**	**64.6166**

4.3 Analysis of the Proposed Method

4.3.1 Visualization Analysis

The modality gap refers to the systematic geometric misalignment between image and text embeddings in the shared feature space, which can reduce the comparability of cross-modal similarities. To further verify the convergence of the modality gap from the perspective of sample-level directional relationships, we analyze the changes in the modality angle before and after training on the top-K image–text samples, where $\Delta\cos$ shows the largest improvement and satisfies $\Delta\cos > 0$, as shown in Figure 3. The left side of Figure 3 presents the probability density histogram of the modality angle, where the horizontal axis denotes the angle θ between the normalized image and text features, and the vertical axis denotes the probability density within the corresponding interval. The right side presents the cumulative distribution function $P(\Theta \leq \theta)$, which is used to characterize the proportion of samples that have completed alignment within a given angular threshold.

As shown in Figure 3, the angle distribution after training shifts overall toward the smaller-angle region, and the cumulative distribution curve after training is located

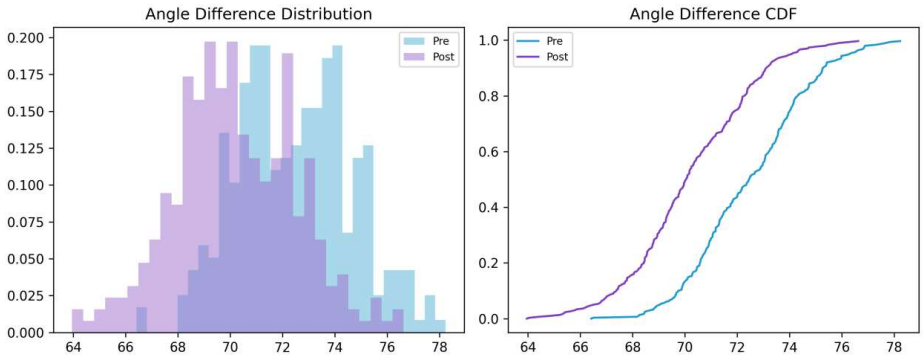


Fig. 3: Changes in the modality angle distribution and cumulative distribution before and after training on top-K image–text samples.

closer to the upper-left region than that before training. This indicates that, under the same angular threshold, a higher proportion of samples satisfy the condition that "the angle does not exceed the threshold" after training. In other words, the average directional deviation between image and text features is reduced, and the modality gap is effectively narrowed in the angular space.

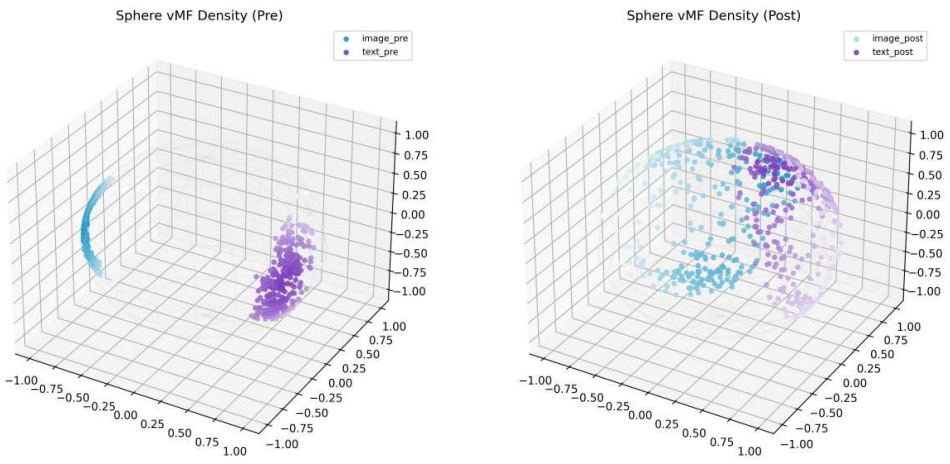
In addition to angular statistics, we further project the top-K image and text features into a unified three-dimensional space and normalize them onto the unit sphere. Figure 4 presents the vMF density distribution of the top-K joint representations on the unit sphere before and after training. In the figure, the color intensity indicates the local density estimated using a vMF-style kernel density estimator. Darker colors indicate that more features are concentrated around the corresponding direction, whereas lighter colors indicate fewer samples or distributional holes in that region.

Before training, the features are mainly concentrated in a few local high-density regions, with limited coverage over the sphere. After training, the hotspot regions become more dispersed, and previously sparse regions are partially filled, leading to a more balanced overall distribution. The proposed method not only reduces the modality gap but also improves the occupancy pattern of joint representations in the spherical space, gradually shifting the features from local over-concentration toward a more uniform and more complete global distribution. This provides a more stable structural basis for improving the generalization ability of the model.

4.3.2 Ablation Study

To further analyze the contribution of each loss term, Table 3 reports the average zero-shot classification accuracy under different loss combinations. It can be observed that when only the contrastive loss is used, the average accuracy of the model is 51.33%. Adding $\mathcal{L}_{\text{div}}$, $\mathcal{L}_{\text{SIRD}}$, or $\mathcal{L}_{\text{uni}}$ individually improves the performance, indicating that all three constraints are effective for post-pretraining.

Furthermore, when multiple losses are jointly used, the overall performance is generally better than that with only a single additional term. The best result is achieved



(a) (b)

Fig. 4: vMF density distribution of top-K joint representations on the unit sphere before and after training.

when all four losses are enabled simultaneously. This suggests that these modules are complementary to some extent: $\mathcal{L}_{\text{div}}$ mainly acts on the global distributional geometry of the shared space, $\mathcal{L}_{\text{SIRD}}$ preserves the stability of cross-sample relational structures, and $\mathcal{L}_{\text{uni}}$ helps improve the overall spreading pattern of features. With the joint effect of these three terms, the model obtains more sufficient structural calibration during post-pretraining, leading to better zero-shot classification performance.

Table 3: Average zero-shot accuracy (%) under different loss combinations

Method	$\mathcal{L}_{\text{nce}}$	$\mathcal{L}_{\text{div}}$	$\mathcal{L}_{\text{SIRD}}$	$\mathcal{L}_{\text{uni}}$	Avg.
E1	✓				51.33
E2	✓	✓			60.01
E3	✓		✓		59.98
E4	✓			✓	59.92
E5	✓	✓	✓		60.36
E6	✓	✓		✓	60.29
E7	✓		✓	✓	60.25
E8 (GeoRel-CLIP)	✓	✓	✓	✓	**60.82**

4.3.3 Parameter Sensitivity Analysis

Influence of different loss weights: After determining the complete model structure, we further examine the influence of different loss weights on zero-shot classification performance, as shown in Figure 5.

All three auxiliary losses improve performance within suitable weight ranges, while excessively large weights may lead to performance saturation or degradation. Specifically, $\mathcal{L}_{\text{uni}}$ achieves the best performance when its weight is set to 0.2, while $\mathcal{L}_{\text{div}}$ and $\mathcal{L}_{\text{SIRD}}$ achieve optimal or near-optimal performance when their weights are set to 0.05. Further increasing these weights does not bring additional gains. By comparison, the weight of InfoNCE has a more pronounced influence on performance, and the average accuracy reaches its best value when the weight is set to 0.2. The default weight settings adopted in this paper balance the discriminative ability of the main task and the strength of geometric calibration, and therefore represent a reasonable trade-off.

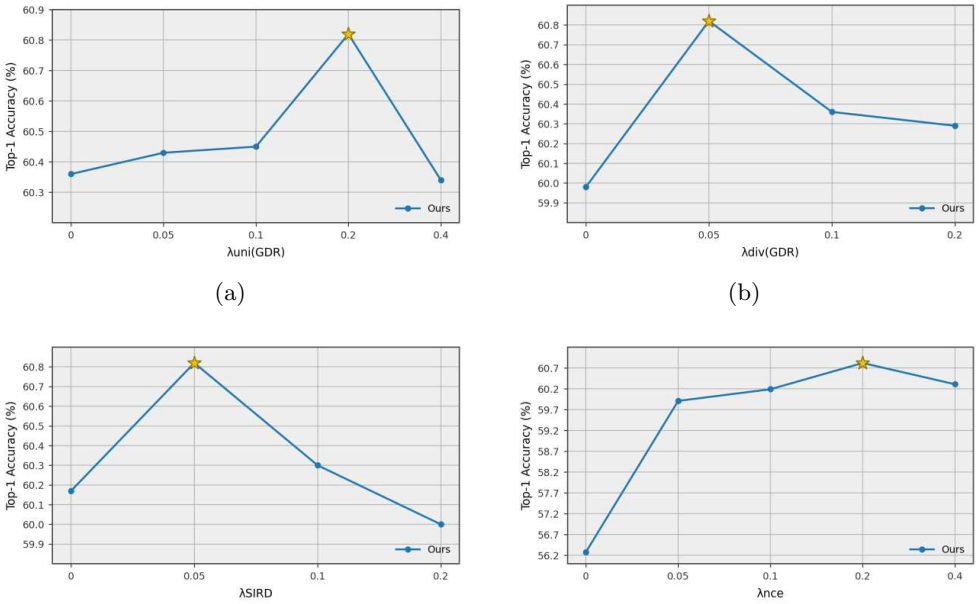


Fig. 5: Average zero-shot accuracy (%) under different loss weight settings.

Influence of different mini-batch size: In the post-pretraining stage, we train the model with different mini-batch sizes $\{4, 32, 64, 128, 256, 512, 1024\}$ and evaluate the corresponding performance. For the batch size of 1024, the gradient accumulation strategy provided by OpenCLIP [53] is used during training. Figure 6 shows the variation trend of the average zero-shot accuracy under different mini-batch sizes. Overall, as the batch size gradually increases from smaller values, model performance continues to improve. This indicates that, during post-pretraining, a larger batch can provide

richer cross-sample relational information, which is beneficial for the stable estimation of relational distillation and geometric statistics. When the batch size increases to 512, the model achieves the best result. When the batch size is further increased to 1024, the performance shows a slight decline. Considering both training stability and computational cost, we adopt 512 as the default configuration.

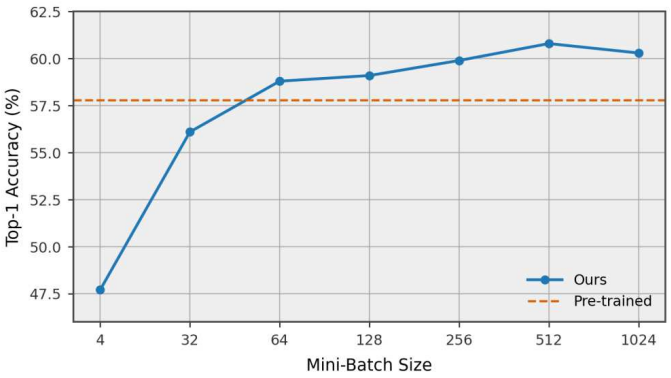


Fig. 6: Effect of different mini-batch sizes on average zero-shot accuracy.

Influence of different numbers of anchors: Figure 7 shows the average zero-shot accuracy under different numbers of anchors. As the number of anchors increases from 128 to 1024, the model performance generally improves, indicating that an appropriate increase in the number of anchors helps more sufficiently characterize the semantic directions and local structures in the shared embedding space.

When the number of anchors further increases to 2048, the performance decreases. This indicates that more anchors do not necessarily lead to better performance. Although an excessive number of anchors improves the granularity of directional partitioning, it may also introduce additional noise and redundant constraints, thereby affecting the overall optimization effect. Considering both performance and computational complexity, we adopt 1024 as the default number of anchors.

5 Conclusion

This paper proposes a collaborative optimization method, GeoRel-CLIP, for the post-pretraining stage of vision-language models, aiming to alleviate the modality gap between image and text features and improve distributional degradation in the shared embedding space. The proposed method consists of two components: Geometric Distribution Regularization (GDR) and Scale-Invariant Relational Distillation (SIRD). GDR constrains the distributional geometry of joint features on the unit hypersphere, encouraging representations to occupy the shared space in a more uniform and stable manner. SIRD performs standardized distillation on cross-modal relational matrices, preserving the relational knowledge of the pretrained model while reducing the interference caused by scale differences.

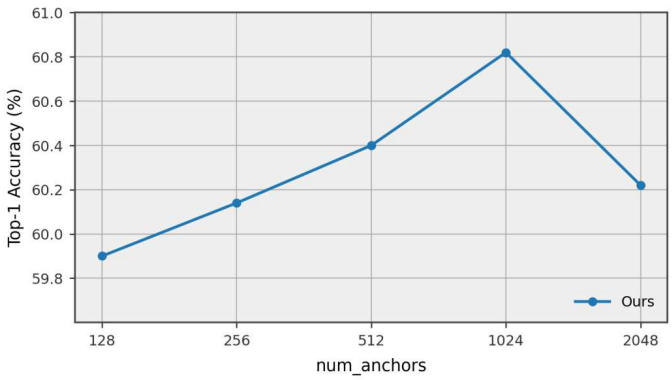


Fig. 7: Effect of different numbers of anchors on average zero-shot accuracy.

Through extensive experiments, this paper demonstrates that the proposed method can effectively reduce the modality gap, improve the zero-shot classification performance of pretrained CLIP, and enhance the uniformity and geometric health of the feature space. These results indicate that post-pretraining is not only an effective optimization strategy with a lower cost than full-scale pretraining, but also provides a new perspective for geometric calibration and performance enhancement of vision-language models.

Declarations

Funding. This work was supported by the Fundamental Research Funds in Universities of Heilongjiang Province (Grant No. 2025-KYYWF-ZR0082).

Conflict of interest. The authors declare that they have no conflict of interest.

Data availability. The datasets used in this study are publicly available.

Code availability. The code used in this study is available from the corresponding author upon reasonable request.

References

[1] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)

[2] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning, pp. 4904–4916. PMLR (2021)

[3] Jing, L., Vincent, P., LeCun, Y., Tian, Y.: Understanding dimensional collapse in contrastive self-supervised learning (2021). https://arxiv.org/abs/2110.09348

[4] Liang, V.W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.Y.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. Advances in Neural Information Processing Systems **35**, 17612–17625 (2022)

[5] Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning, pp. 9929–9939. PMLR (2020)

[6] Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (2023). https://doi.org/10.1109/ICCV51070.2023.01100

[7] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. Int J Comput Vis **130** (2022) https://doi.org/10.1007/s11263-022-01653-1

[8] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16816–16825 (2022). https://doi.org/10.1109/CVPR52688.2022.01631

[9] Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: CLIP-adapter: Better vision-language models with feature adapters. Int J Comput Vis **132**, 581–595 (2024) https://doi.org/10.1007/s11263-023-01891-x

[10] Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free adaption of CLIP for few-shot classification. In: European Conference on Computer Vision, pp. 493–510. Springer (2022). https://doi.org/10.1007/978-3-031-19833-5_29

[11] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A.: Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci U S A **114**, 3521–3526 (2017) https://doi.org/10.1073/pnas.1611835114

[12] Li, Z., Hoiem, D.: Learning without forgetting. IEEE Trans Pattern Anal Mach Intell **40**, 2935–2947 (2017) https://doi.org/10.1109/TPAMI.2017.2773081

[13] Murugesan, B., Silva-Rodríguez, J., Ayed, I.B., Dolz, J.: Robust calibration of large vision-language adapters. In: European Conference on Computer Vision, pp. 147–165. Springer (2024). https://doi.org/10.1007/978-3-031-72691-0_9

[14] Zheng, M., Tang, Y., Hao, Z., Han, K., Wang, Y., Xu, C.: Adapt without forgetting: Distill proximity from dual teachers in vision-language models. In:

European Conference on Computer Vision, pp. 109–125. Springer (2024). https://doi.org/10.1007/978-3-031-72949-2_7

[15] Yamaguchi, S., Feng, D., Kanai, S., Adachi, K., Chijiwa, D.: Post-pre-training for modality alignment in vision-language foundation models. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 4256–4266 (2025). https://doi.org/10.1109/CVPR52734.2025.00402

[16] Bianchi, L., Carrara, F., Messina, N., Falchi, F.: Is CLIP the main roadblock for fine-grained open-world perception? In: 2024 International Conference on Content-Based Multimedia Indexing (CBMI), pp. 1–8. IEEE (2024). https://doi.org/10.1109/CBMI62980.2024.10859215

[17] Huang, Z., Chen, H., Wen, Z., Zhang, C., Li, H., Wang, B., Chen, C.: Model-aware contrastive learning: Towards escaping the dilemmas. In: International Conference on Machine Learning, pp. 13774–13790. PMLR (2023)

[18] Jing, D., He, X., Luo, Y., Fei, N., Wei, W., Zhao, H., Lu, Z.: FineCLIP: Self-distilled region-based CLIP for better fine-grained understanding. Advances in Neural Information Processing Systems **37**, 27896–27918 (2024) https://doi.org/10.52202/079017-0875

[19] Wang, F., Liu, H.: Understanding the behaviour of contrastive loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2495–2504 (2021). https://doi.org/10.1109/CVPR46437.2021.00252

[20] Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: FLAVA: A foundational language and vision alignment model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15638–15650 (2022). https://doi.org/10.1109/CVPR52688.2022.01519

[21] Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in Neural Information Processing Systems **34**, 9694–9705 (2021)

[22] Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning, pp. 19730–19742. PMLR (2023)

[23] Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, pp. 12888–12900. PMLR (2022)

[24] Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: ImageBind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15180–15190 (2023). https://doi.org/10.1109/CVPR52729.2023.01457

[25] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M.: LAION-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)

[26] Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J.: DataComp: In search of the next generation of multimodal datasets. Advances in Neural Information Processing Systems **36**, 27092–27112 (2023)

[27] Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: LiT: Zero-shot transfer with locked-image text tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18123–18133 (2022). https://doi.org/10.1109/CVPR52688.2022.01759

[28] Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: CoCa: Contrastive captioners are image-text foundation models (2022). https://arxiv.org/abs/2205.01917

[29] Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A.J., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L.: PaLI: A jointly-scaled multilingual language-image model (2022). https://arxiv.org/abs/2209.06794

[30] Shu, M., Nie, W., Huang, D.-A., Yu, Z., Goldstein, T., Anandkumar, A., Xiao, C.: Test-time prompt tuning for zero-shot generalization in vision-language models. Advances in Neural Information Processing Systems **35**, 14274–14289 (2022)

[31] Huang, S., Gong, B., Feng, Y., Zhang, M., Lv, Y., Wang, D.: Troika: Multi-path cross-modal traction for compositional zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24005–24014 (2024). https://doi.org/10.1109/CVPR52733.2024.02266

[32] Liu, T., Yang, C., Chen, G., Li, W.: Open-vocabulary multi-label classification with visual and textual features fusion. Vis Comput **41**, 6027–6039 (2025) https://doi.org/10.1007/s00371-024-03769-6

[33] Jiang, Q., Chen, C., Zhao, H., Chen, L., Ping, Q., Tran, S.D., Xu, Y., Zeng, B., Chilimbi, T.: Understanding and constructing latent modality structures in multi-modal representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7661–7671 (2023). https://doi.org/10.1109/CVPR52729.2023.00740

[34] Zhu, L., Han, D., Shen, X., Chen, C., Li, K.-C.: Enhancing image-text matching through multi-level semantic consistency alignment. Vis Comput **41**, 9555–9570 (2025) https://doi.org/10.1007/s00371-025-03981-y

[35] Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network

(2015). https://arxiv.org/abs/1503.02531

[36] Yang, Y., Xi, W., Zhou, L., Tang, J.: Rebalanced vision-language retrieval considering structure-aware distillation. IEEE Trans Image Process (2024) https://doi.org/10.1109/TIP.2024.3518759

[37] Huang, C.-W., Chen, Y.-N.: PairDistill: Pairwise relevance distillation for dense retrieval. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 18225–18237 (2024). https://doi.org/10.18653/v1/2024.emnlp-main.1013

[38] Sun, S., Ren, W., Li, J., Wang, R., Cao, X.: Logit standardization in knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15731–15740 (2024). https://doi.org/10.1109/CVPR52733.2024.01489

[39] Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft (2013). https://arxiv.org/abs/1306.5151

[40] Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-UCSD birds 200. Technical report, California Institute of Technology (2010)

[41] Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 554–561 (2013). https://doi.org/10.1109/ICCVW.2013.77

[42] Nilsback, M.-E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics and Image Processing, pp. 722–729. IEEE (2008). https://doi.org/10.1109/ICVGIP.2008.47

[43] Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: European Conference on Computer Vision, pp. 446–461. Springer (2014). https://doi.org/10.1007/978-3-319-10599-4_29

[44] Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3498–3505. IEEE (2012). https://doi.org/10.1109/CVPR.2012.6248092

[45] Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3606–3613 (2014). https://doi.org/10.1109/CVPR.2014.461

[46] Helber, P., Bischke, B., Dengel, A., Borth, D.: EuroSAT: A novel dataset and

deep learning benchmark for land use and land cover classification. IEEE J Sel Top Appl Earth Obs Remote Sens **12**, 2217–2226 (2019) https://doi.org/10.1109/JSTARS.2019.2918242

[47] Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Comput Vis Image Underst **106**, 59–70 (2007) https://doi.org/10.1016/j.cviu.2005.09.012

[48] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.: ImageNet large scale visual recognition challenge. Int J Comput Vis **115**, 211–252 (2015) https://doi.org/10.1007/s11263-015-0816-y

[49] Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3485–3492. IEEE (2010). https://doi.org/10.1109/CVPR.2010.5539970

[50] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision, pp. 740–755. Springer (2014). https://doi.org/10.1007/978-3-319-10602-1_48

[51] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)

[52] Oh, C., So, J., Byun, H., Lim, Y., Shin, M., Jeon, J.-J., Song, K.: Geodesic multimodal mixup for robust fine-tuning. Advances in Neural Information Processing Systems **36**, 52326–52341 (2023)

[53] Qian, Q., Xu, Y., Hu, J.: Intra-modal proxy learning for zero-shot visual categorization with CLIP. Advances in Neural Information Processing Systems **36**, 25461–25474 (2023)



Huadong Sun was born in Heilongjiang, China, in 1980. He received the Ph.D. degree in communication engineering from Harbin Institute of Technology, Harbin, China, in 2009. He joined Harbin University of Commerce in 2010, where he is currently a Professor and Master's Supervisor. His research interests include image processing and signal processing. His recent research interests include digital image processing and computer vision.



Yexuan Zhao has been a master's student at the School of Computer and Information Engineering, Harbin University of Commerce, Harbin, China, since 2024. Her recent research interests include zero-shot learning, vision-language models, and computer vision, with a focus on improving the generalization ability of visual recognition models in open-category scenarios.



Junwei Tong has been a master's student at the School of Computer and Information Engineering, Harbin University of Commerce, Harbin, China, since 2024. Her recent research interests include few-shot object detection, computer vision, and visual recognition, with a focus on improving object detection performance under limited annotation and limited training-sample conditions.