KDKD: Cross-Architecture Knowledge Distillation for Few-shot Unsupervised Vision-Language Models

Kailai Zhuang[1], Jiajie Fan[2], Rundong Gao[2], Zeming Tian[4] Qingzeng Song[1,*], Yongjiang Xue[1],
[1]Tiangong University, China [2]South China Normal University, China
[3]Shenyang University of Chemical Technology, China [4]Wuhan University, China
*Corresponding Author: qingzengsong@163.com

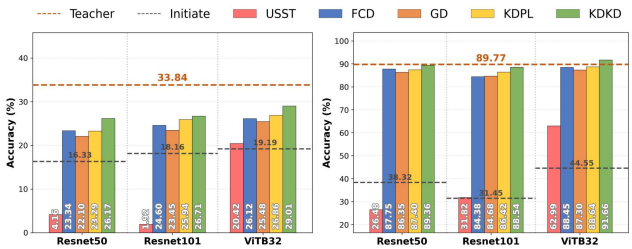
***Abstract*—Vision-Language Models (VLMs) like CLIP have utilized large-scale pre-training to achieve impressive zero-shot capabilities. However, adapting these heavy models to downstream tasks in data-scarce scenarios remains challenging. While existing unsupervised self-training methods have shown promise on Vision Transformers (ViTs), they suffer from catastrophic performance degradation when applied to Convolutional Neural Networks (CNNs), even falling below zero-shot baselines. To bridge this gap, we propose KDKD (Kendall-Driven Knowledge Distillation), a novel framework designed for robust cross-architecture adaptation. We introduce a bidirectional adaptive temperature mechanism guided by the Kendall Correlation Coefficient, which dynamically aligns the prediction rankings between teacher and student models. Furthermore, we propose a soft-weighting strategy, as opposed to hard truncation, to adjust the loss based on sample quality. Extensive experiments on 9 downstream datasets demonstrate that KDKD not only boosts ViT students but effectively rescues ResNet students from degradation, achieving state-of-the-art performance in few-shot unsupervised settings with less computational overhead. Our codes are available at https://github.com/zhuangkailai/KDKD**

***Index Terms*—Vision-Language Models, Knowledge Distillation, Unsupervised Adaptation, Few-Shot Learning, Kendall Correlation, Cross-Architecture Transfer**

I. Introduction

Vision-Language Models(VLMs), as an important paradigm of multimodal learning, achieve cross-modal understanding by jointly modeling visual and textual information. VLMs, through extensive pre-training on large-scale image description datasets, can learn robust visual representations and have been applied to a wide range of downstream tasks [1]–[3], covering visual recognition, cross-modal retrieval, multimodal understanding, and image description. However, the large number of parameters and complex architecture make training visual language models resource-intensive. Furthermore, the availability of large-scale, general, non-fine-grained category datasets for pre-training leaves significant room for improvement in downstream, more specific tasks such as classification and image-text retrieval.

Adapters [4]–[7] and Prompt Learning [8], [9] enable domain adaptation while preserving pre-trained knowledge, yet they rely heavily on human-designed modules and abundant labeled data. In contrast, few-shot unsupervised adaptation [10], [11] reduces reliance on training data and true labels. However, existing unsupervised self-training(USST) approaches are limited to Vision Transformer (ViT) architectures



(a) FGVC (b) EuroSat

Fig. 1: Comparison of Student model accuracy between our proposed KDKD and one Un-supervised Self-training method and three state-of-the-art KD methods. The initiat baselines denote the zero-shot performance derived solely from the pre-trained model's inherent representational capacity, prior to any fine-tuning or distillation.

and suffer from catastrophic degradation on convolutional neural networks (CNNs) like ResNet, even performing worse than the initial pre-trained model which without any downstream adaptation, as shown in the red region of Fig. 1. Moreover, even encoder modules with the same architecture have significant differences in the number of parameters and computational complexity. A lot of exploration and research aims to solve how to effectively transfer knowledge from large models to compact models [12]–[15]. While unimodal distillation operates within a single feature space, distilling multimodal models requires aligning both visual and textual representations along with their cross-modal interactions [16], [17], making the task more complex.

This paper proposes KDKD(Kendall-Driven Knowledge Distillation), a framework for few-shot unsupervised adaptation. We first systematically identify knowledge components that are universal, architecture-agnostic, and dimension-consistent, then introduce an adaptive temperature mechanism based on Kendall Tau correlation to optimize knowledge propagation by aligning sample-wise prediction rankings between teacher and student.

Extensive experiments on 9 downstream vision tasks across ViT and ResNet architectures demonstrate that KDKD effectively compresses large vision-language models in few-shot scenarios. Our method maintains or surpasses teacher performance while drastically reducing parameters and computation, even under significant teacher-student disparities in architec-

ture, size, performance, and embedding dimensions. Notably, KDKD resolves the catastrophic degradation of ResNet in unsupervised adaptation, demonstrating robust cross-architecture distillation with minimal computational overhead.

II. Related Work

A. Few-Shot Unsupervised Learning for VLMs

MUST [18] employs masked unsupervised self-training, using both pseudo-labels and original images to jointly optimize three objectives, thereby learning both class-level global and pixel-level local features. CALIP [19] strengthens bidirectional modality alignment via a parameter-free approach that updates encoded features based solely on intermediate visual-textual similarity. LaFTer [20] is the first to show how to close the performance gap between zero-shot and supervised fine-tuned classifiers using only unlabeled images and LLM-generated text descriptions, without requiring any labeled or paired vision-language data. DPA [21] introduces dual prototypes as distinct classifiers, generates accurate pseudo-labels via their convex combination, and enhances robust self-training through ranked pseudo-labels while aligning textual and visual prototypes to mitigate visual-text misalignment. However, we observed that existing unsupervised adaptation methods (such as DPA) are designed and validated only for the ViT (Vision Transformer) [22] architecture, and applying them to the ResNet [23] network architecture results in a catastrophic performance degradation. Methods involving visual prompt learnable tokens like VPT [24] , MaPLe [25] , and PromptSRC [26] can only be applied to VLMs equipped with a ViT-based image encoder [27] .

B. Knowledge Distillation for VLMs

Both CLPI-KD [28] and Multi-modal Cooperative Distillation [29] employ embedding alignment loss that demonstrates significant effectiveness. However, the raw model outputs cannot be directly utilized due to dimension mismatch, necessitating an additional projection module that introduces substantial computational overhead. Additionally, the Contrastive Relational Distillation method introduced in clip-kd incorporates an extra text-to-image alignment direction compared to Multi-modal Cooperative Distillation, yet it combines the directional losses through direct summation: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{feat}\to\text{text}} + \mathcal{L}_{\text{text}\to\text{feat}}$. Since gradient information reveals how the model's output changes with respect to inputs, clip-kd exploits this by proposing gradient distillation, which forces the student to mimic the teacher's input-response behavior. KDPL [30] uses dynamic class sampling, retaining only the Top-K classes for distillation, focusing on high-confidence classes and avoiding interference from noisy classes. The above work, in terms of details, simply averages the KL divergence loss for the current batch. Assigning uniform weights to all samples might hinder the adaptive process [21]. Therefore, we calculate a weight based on the similarity between the pseudo-label yand the text and image prototypes, and use this weight to average the loss of the current batch. This is similar to the high-quality sample filtering function in KDPL, except that KDPL filters the elements that generate knowledge before distillation, while our KDKD retains all logits, without discarding any samples within the batch, and uses weights to adjust after knowledge generation. Furthermore, maintaining a constant temperature is usually suboptimal for growing students in the progressive learning stage [31]. Many works have explored adaptive temperature selection and adjustment in single-modal distillation, but existing distillation methods in the multi-modal distillation field ignore the flexible role of temperature in the loss function. We adaptively adjust the distillation temperature between the teacher and student based on the Kendall correlation coefficient. This mechanism eliminates the need for manually exploring and fixing a suitable temperature before training, and can achieve optimal knowledge distillation within a certain range.

III. Methodology

A. The Overall Framework

We present the KDKD framework as a three-stage process: *Knowledge Generation*, *Knowledge Regulation*, and *Knowledge Transfer*. As shown in Fig. 2, the teacher and student models share the same output, obtain visual features via their respective visual encoders, obtain text prototypes via the text encoder, and initialise the classification head. We generate the raw knowledge (*logits*) using the visual features and text prototypes. Then compute the Kendall correlation coefficient to adaptively adjust the distillation temperature, and employ sample-level weighting to re-weight the distillation loss. The modulated knowledge is then transferred to the student through distinct loss combinations for ViT and ResNet students respectively. The following subsections detail each of these components.

B. Knowledge Generation

In the proposed advanced Unsupervised Self-Training (USST) framework, the Teacher ($\mathcal{T}$) and Student ($\mathcal{S}$) models share a unified data augmentation pipeline for processing unlabeled images. For each image x, this pipeline generates a weakly-augmented view $\alpha(x)$ and a strongly-augmented view $\mathcal{A}(x)$. These identical views are processed by the respective image encoders of both models to generate each one's features: the weak feature $\mathbf{f} = E_v(\alpha(x))$ and the strong feature $\mathbf{F} = E_v(\mathcal{A}(x))$, where $\mathbf{f}, \mathbf{F} \in \mathbb{R}^d$.

To align these visual representations with semantic concepts, the framework utilizes two distinct sets of prototypes: Visual Prototypes ($\mathbf{P} \in \mathbb{R}^{C\times d}$) represent the empirical data distribution. For each class j, the prototype $\mathbf{P}_j$ is constructed by averaging the feature vectors of samples assigned to that class via pseudo-labels $\hat{y}$:

$$\mathbf{P}_j = \frac{1}{N_j} \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = j)\mathbf{f}_i \tag{1}$$

where N_j is the count of samples in class j. Textual Prototypes ($\mathbf{Z} \in \mathbb{R}^{C\times d}$) provide semantic grounding and are initialized using the CLIP text encoder. For a specific class c_j, representations from k diverse prompt templates are averaged to form a

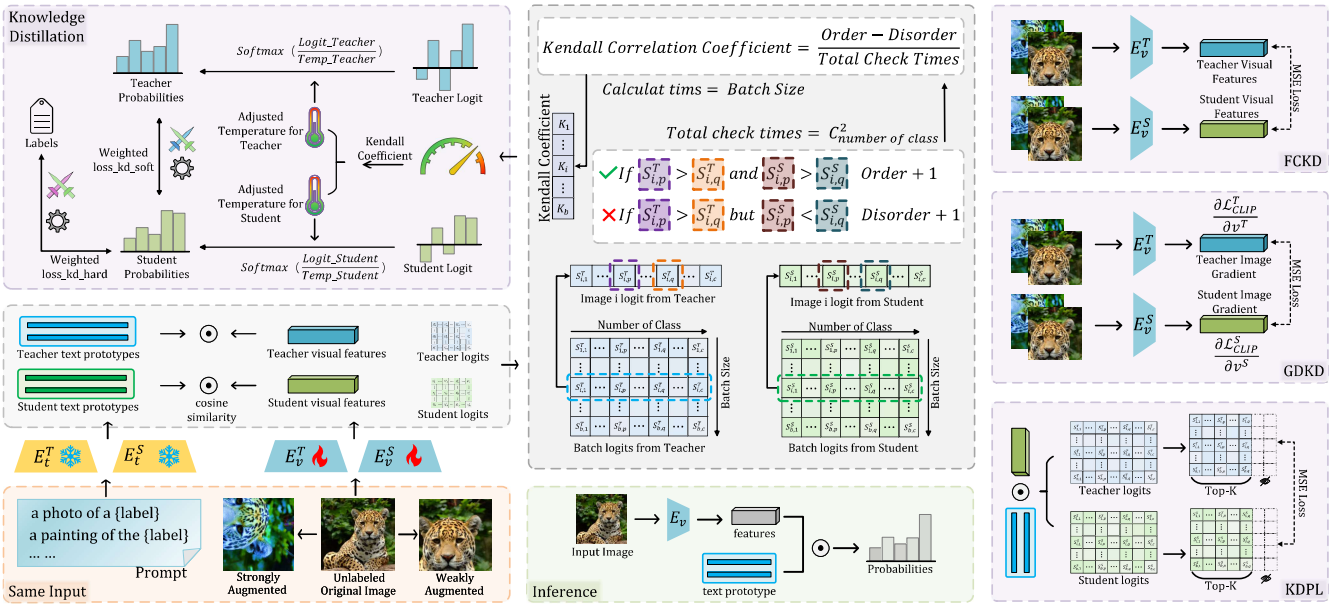


Fig. 2: The overall framework of KDKD and three other advanced knowledge distillation methods for VLMs.

robust textual prototype $\mathbf{Z}_j = \frac{1}{k}\sum_{i=1}^{k} \mathbf{z}_{ji}$.Textual prototypes serve as the fixed or learnable weights of the textual classifier, denoted as $\mathbf{W}_{txt}$.

The prediction logits z for either model are generated via the cosine similarity between the normalized visual features and the textual prototypes:

$$z = \frac{\mathbf{f}}{\|\mathbf{f}\|} \cdot \frac{\mathbf{W}_{txt}^{T}}{\|\mathbf{W}_{txt}\|} \tag{2}$$

Both models execute this process identically to produce the student logits z^s and teacher logits z^t.

C. Knowledge Regulation

Our regulation method operates in two aspects: bidirectional temperature adjustment via correlation, and sample-wise weighting. First, to regulate the stiffness of the probability distributions, we compute the Kendall Rank Correlation Coefficient between the logits of the teacher z^t and the student z^s. This metric evaluates the similarity of the class ranking order predicted by the two models. For a prediction over K classes, the coefficient τ is defined as:

$$\tau = \frac{C - D}{\frac{1}{2}Classes(Classes - 1)} \tag{3}$$

where C is the number of *concordant pairs*, and D is the number of *discordant pairs*. A pair of distinct classes (p, q) is considered concordant if the sorting order of their logits agrees between the teacher and student (i.e., $(z_p^t - z_q^t)(z_p^s - z_q^s) > 0$), and discordant otherwise. The term $\frac{1}{2}Classes(Classes - 1)$ represents the total number of unique pairs.

We normalize this coefficient by $\tilde{\tau} = (\tau + 1)/2$ to the range $[0, 1]$ and then bidirectionally adjust the distillation temperatures (T^t and T^s) based on $\tilde{\tau}$:

$$\begin{aligned} T^t &= T^t_{\min} + \tilde{\tau} \cdot (T^t_{\max} - T^t_{\min}) \\ T^s &= T^s_{\min} + (1 - \tilde{\tau}) \cdot (T^s_{\max} - T^s_{\min}) \end{aligned} \tag{4}$$

Second, regarding sample selection, prior work like KDPL utilizes a hard truncation strategy retains only the top-k logits and masks the rest, decisively removing noise but also discarding tail-class information. Instead, we propose a soft weighting mechanism applied after loss calculation, rather than pre-loss truncation. We compute a confidence weight w derived from the consistency between the visual features, text prototypes, and unsupervised cluster centers. For a sample i, the weight is calculated as:

$$w_i = \alpha \cdot p_{center} + (1 - \alpha) \cdot p_{text} \tag{5}$$

where p denotes the softmax probability derived from visual feature with different prototypes. This weight w_i will modulate loss distribution during backpropagation, preserving the full logit structure while rigorously down-weighting low-quality ambiguous samples.

D. Knowledge Transfer

The strategy depends on the student architecture's stability during unsupervised self-training. First, the standard unsupervised loss be defined as $\mathcal{L}_S = \mathcal{L}_{align} + \mathcal{L}_{cls} + \mathcal{L}_{reg}$. Here, $\mathcal{L}_{align}$ aligns visual centers with text, $\mathcal{L}_{cls}$ is the pseudo-label classification loss, and $\mathcal{L}_{reg}$ is a fairness regularization loss. For the distillation component, we define the teacher transfer loss $\mathcal{L}_T$, which comprises a soft KL-divergence term (using

TABLE I: Comparison of Student model accuracy between our proposed KDKD and one Un-supervised Self-training method and three state-of-the-art KD methods. The initiat baselines denote the zero-shot performance derived solely from the pre-trained model's inherent representational capacity, prior to any fine-tuning or distillation.

Backbone	Method	Downstream datasets									
		Cars	DTD	EuroSat	FGVC	Flowers	Food101	Pets	UCF101	Caltech	Average
ResNet50	Initiate	54.60	41.14	38.32	16.33	65.99	77.28	85.80	63.13	83.70	58.48
	USST	2.05	28.84	26.48	4.16	22.60	1.45	53.78	49.16	83.57	30.23
	USST+FCD	63.92	54.71	87.75	23.34	79.12	81.71	84.93	68.40	91.06	70.55
	USST+GD	62.48	53.36	86.35	22.10	78.46	80.30	83.65	67.39	90.68	69.42
	USST+KDPL	65.59	54.22	87.40	23.29	80.73	82.72	85.41	68.13	91.46	70.99
	USST+KDKD	67.76	55.75	89.36	26.17	82.33	84.26	86.75	70.10	93.72	72.91
ResNet101	Initiate	61.07	43.57	31.45	18.16	64.52	81.62	87.04	61.03	89.55	59.78
	USST	1.28	33.76	31.82	1.92	25.76	1.73	58.15	56.03	87.60	33.12
	USST+FCD	68.83	51.36	84.38	24.60	78.03	81.32	84.75	72.14	91.70	70.79
	USST+GD	67.42	50.17	84.68	23.45	78.64	80.16	83.41	71.35	90.45	69.97
	USST+KDPL	70.90	51.35	86.42	25.94	80.54	83.90	85.71	72.01	91.46	72.03
	USST+KDKD	72.87	53.58	88.54	26.71	82.69	84.51	87.85	74.43	93.62	73.87
ViT-B/32	Initiate	59.75	44.11	44.55	19.19	66.35	81.62	87.55	64.05	90.62	61.98
	USST	62.38	55.96	62.99	20.42	76.41	82.96	89.26	70.09	95.35	68.42
	USST+FCD	72.87	58.45	88.45	26.12	84.61	83.75	90.03	76.94	96.19	75.27
	USST+GD	72.60	58.32	87.30	25.48	83.90	83.10	89.42	75.47	95.90	74.61
	USST+KDPL	73.18	59.10	88.64	26.86	85.47	84.62	90.14	77.90	95.60	75.72
	USST+KDKD	76.57	61.44	91.66	29.01	85.62	86.18	92.37	78.86	96.78	77.61
ViT-L/14	Teacher	79.12	60.85	89.77	33.84	82.93	92.57	94.25	79.65	97.57	78.95

the adaptive temperatures regulated in Sec. III-C) and a hard cross-entropy term using teacher pseudo-labels $\hat{y}^t$:

$$\mathcal{L}_{soft} = (T_s)^2 \cdot KL(\frac{z^s}{T_s} || \frac{z^t}{T_t}) \tag{6}$$

$$\mathcal{L}_{hard} = CE(z^s, \hat{y}^t) \tag{7}$$

$$\mathcal{L}_T = w \cdot (\lambda \mathcal{L}_{soft} + (1-\lambda)\mathcal{L}_{hard}) \tag{8}$$

When the student is a ResNet, standard USST destructively degrades accuracy even below the initial zero-shot baseline. Thus, we discard loss from USST $\mathcal{L}_S$ entirely and rely solely on the teacher's guidance:

$$\mathcal{L}_{KT1} = \mathcal{L}_{KD} \tag{9}$$

which is denoted as Knowledge Transfer Method 1 (KT1). When the student is a ViT, we retain the student's self-learning capability while augmenting it with teacher knowledge:

$$\mathcal{L}_{KT2} = \mathcal{L}_S + \mathcal{L}_{KD} \tag{10}$$

which is denoted as Knowledge Transfer Method2 (KT2).

IV. Experiments

A. Datasets and Baselines

We extensively evaluate on 8 diverse datasets: Standford-Cars [32], DTD [33], EuroSAT [34], FGVCAircraft [35], Caltech101 [36], Flowers102 [37], Food101 [38], OxfordPets [39] and UCF101 [40]. We perform comparative analysis with two SOTA knowledge distillation methods designed for unsupervised adaptive models: CLIP-KD [28], KDPL [30].

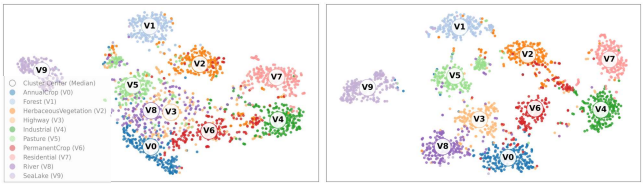
B. Implementation Details

We employ CLIP-ViT-L/14 as the Teacher model ($\mathcal{T}$). For the Student model ($\mathcal{S}$), we evaluate performance across three distinct architectures: ViT-B/32, ResNet-50, and ResNet-101.

For the knowledge transfer method 1 (KT1), we set the ratio of soft to hard label losses ($\mathcal{L}_{soft} : \mathcal{L}_{hard}$) to 7 : 3. In method 2 (KT2), the weight ratio between the self-training loss ($\mathcal{L}_S$) and the transfer loss ($\mathcal{L}_T$) is set to 1 : 1. The adaptive distillation temperature ranges are set to [2, 5] for the teacher model and [3, 7] for the student model. We set both teacher and student models are trained for 10 epochs each, while the order of training can be manually flexibly adjusted.

C. Results

Table I and Fig. 1 demonstrate that KDKD successfully rescues ResNet from catastrophic accuracy degradation, achieving average improvements of 40.68% for ResNet50, 40.75% for ResNet101, and 9.19% for ViT-B/32, which significantly outperforms other knowledge distillation approaches. The t-SNE visualization in Fig. 3 further illustrates the enhanced classification capability of the unsupervised model.

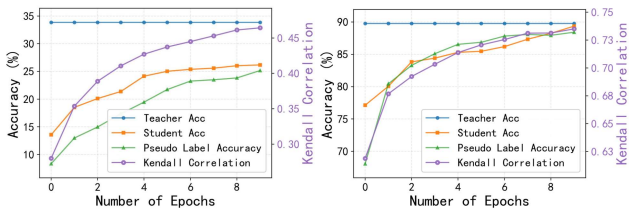


(a) Train Resnet50 by USST (b) Train Resnet50 by KDKD

Fig. 3: Comparison of t-SNE projections for student Resnet50 trained by USST and KDKD.

Moreover, as depicted in Fig. 4, the relationship between the Kendall correlation coefficient and student performance during training indicate a positive correlation: as training progresses, both the student's accuracy and pseudo-label quality rise in tandem with the Kendall coefficient. This trend suggests

that the student effectively internalizes knowledge from the teacher, leading to more reliable pseudo-labels compared to self-training alone.

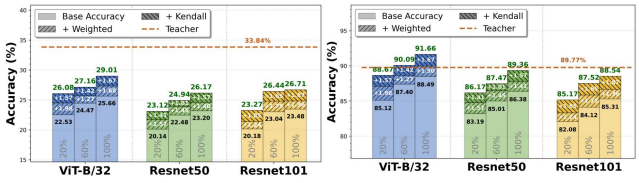


(a) FGVC (b) EuroSat

Fig. 4: Kendall correlation coefficient in Relation to Student Accuracy and Student-generated Pseudo-Label Accuracy.

D. Ablation and Analysis Experiments

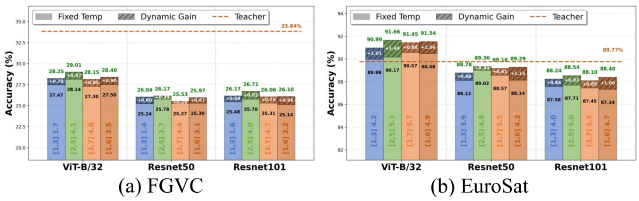
Weighted and Kendall. We experiment with different truncation ratios, retaining the top k% of classes for logit distillation. Performance improves as more classes are included, reaching the best result when all classes are used (100% retention). While full logit distillation appears optimal, it does not preclude sample quality selection. Instead of hard truncation (as in KDPL), we apply soft weighting after computing the full loss, which proves more effective than class clipping. Furthermore, Kendall correlation consistently enhances distillation for both truncated and full logits, as illustrated in Fig. 5.



(a) FGVC (b) EuroSat

Fig. 5: Accuracy Comparison Across Logit Truncation Ratios with Soft Weighting and Adaptive Temperature Tuning.

Temperature for Knowledge Distillation. We explored multiple temperature ranges in our experiments. For each range, we identified the epoch with the highest student accuracy and fixed the corresponding temperatures as $T_{\text{fixed}}^{\text{student}}$ and $T_{\text{fixed}}^{\text{teacher}}$. Keeping all other training configurations unchanged, we then performed an additional 10 epochs of knowledge distillation using these fixed temperatures. The results are shown in Fig. 6. Different temperature ranges indeed have a slight impact on the results. However, regardless of the chosen range, enabling dynamic adjustment based on the Kendall correlation coefficient consistently yields better performance than selecting a fixed value a posteriori within the range.This demonstrates the flexibility and practicality of our method: regardless of the initial choice of temperature range, which is difficult to determine optimally in advance, the dynamic strategy serves as the most effective companion within any given range.



(a) FGVC (b) EuroSat

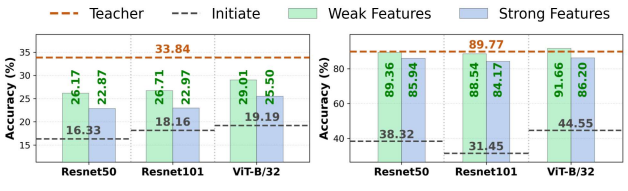
Fig. 6: Performance Comparison Across Temperature Ranges and Fixed Values from each one.

Original USST loss for Resnet. Are the three unsupervised self-training losses really harmful for ResNet? We added the three USST loss terms separately and found that all three losses are actually useful. As shown in table II, using all three original USST losses alone has a negative effect, but when built on the knowledge distillation from a strong teacher model, the three losses show a certain positive effect. The function of these losses relies on the support of a strong teacher; if trained independently, they are ineffective.

TABLE II: Performance Comparison of ResNet with Different Loss Configurations. ● represents with and ○ represents without. PLAcc is the accuracy of the pseudo labels generated by the student, Acc is the highest accuracy achieved by the student after training.

Backbone	Initiate	$\mathcal{L}_{kd}$	$\mathcal{L}_{cls}$	$\mathcal{L}_{align}$	$\mathcal{L}_{reg}$	Kendall τ	PLAcc (%)	Acc (%)
ResNet50	38.32	●	○	○	○	0.75	88.00	86.35
	38.32	●	●	●	●	0.78	87.74	89.36
	38.32	○	●	●	●	0.21	27.10	26.48
	38.32	●	●	○	○	0.72	86.52	88.12
	38.32	●	○	●	○	0.70	88.13	85.33
	38.32	●	○	○	●	0.75	88.40	88.50
ResNet101	31.45	●	○	○	○	0.75	88.20	86.17
	31.45	●	●	●	●	0.77	86.49	88.54
	31.45	○	●	●	●	0.24	31.03	31.84
	31.45	●	●	○	○	0.73	86.90	88.04
	31.45	●	○	●	○	0.70	88.14	85.17
	31.45	●	○	○	●	0.76	88.21	87.97

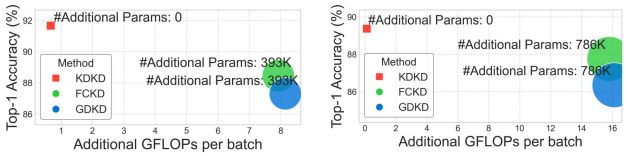
Other sources of knowledge. In addition to weak features, USST also generates strong features during its forward propagation. We investigate whether these elements also contain transferable knowledge. As shown in Fig. 7, distilling the logits derived from the interaction between strong features with textual prototypes can also improve student performance, but less effectively than using weak features.



(a) FGVC (b) EuroSat

Fig. 7: Accuracy Comparison of Knowledge Distillation from Weak Features vs. Strong Features.

Efficiency comparison. Fig. 8 presents a comparison of computational resources across methods. Notably, KDKD achieves the highest distillation performance without introducing extra parameters and with considerably lower GFLOPs, underscoring its efficiency.



(a) ViT-L/14 train ViT-B/32 (b) ViT-L/14 train Resnet50

Fig. 8: Additional Computational Cost Comparison of Three Knowledge Distillation Methods.

V. Conclusion

In this paper, we propose KDKD, a cross-architecture knowledge distillation framework for Vision-Language Models under the few-shot unsupervised training scenario, aimed at bridging the performance gap in unsupervised adaptation—especially for CNN-based students that suffer from severe degradation in self-training. By introducing Kendall-correlation-guided adaptive temperature modulation and a soft-weighted loss combination, KDKD leverages only raw, generic inputs without auxiliary modules to project, efficiently generate, regulate, and transfer knowledge. Extensive experiments demonstrate that KDKD outperforms existing state-of-the-art methods, enabling robust heterogeneous distillation without reliance on labeled data.

References

[1] J. Li, R. R. Selvaraju, ...“Align before Fuse: Vision and Language Representation Learning with Momentum Distillation.” 2021. [Online]. Available: https://arxiv.org/abs/2107.07651

[2] L. Yao et al. “FILIP: Fine-grained Interactive Language-Image Pre-Training.” 2021. [Online]. Available: https://arxiv.org/abs/2111.07783

[3] J. Yu, Z. Wang, V. Vasudevan ... “CoCa: Contrastive Captioners are Image-Text Foundation Models.” 2022. [Online]. Available: https://arxiv.org/abs/2205.01917

[4] Gao, Peng, Geng, Shijie, Zhang, Renrui, Ma, Teli ...“CLIP-Adapter: Better Vision-Language Models with Feature Adapters”. arXiv preprint arXiv:2110.04544, 2021

[5] R. Zhang, R. Fang, W. Zhang ...“Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling”. arXiv:2111.03930, 2021.

[6] Y. Wang, P. Ding, L. Li, C. Cui, Z. Ge, ...“VLA-Adapter: An Effective Paradigm for Tiny-Scale Vision-Language-Action Model” 2025. [Online]. Available: https://arxiv.org/abs/2509.09372

[7] R. Zhang, R. Fang, W. Zhang, ...“Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling” arXiv:2111.03930, 2021.

[8] Zhou, J. Yang, C. C. Loy, and Z. Liu... “Conditional Prompt Learning for Vision-Language Models.” 2022. [Online]. Available: https://arxiv.org/abs/2203.05557

[9] G. Chen, W. Yao, X. Song, ... “PLOT: Prompt Learning with Optimal Transport for Vision-Language Models” arXiv:2210.01253, 2023.

[10] Zhang, Y., Zhang, R., Fang, R., Gao, P., Li, K., Dai, J., ... “Candidate pseudo-label learning for unsupervised fine-tuning of vision-language models” arXiv preprint arXiv:2211.13745.

[11] Ali, M., Touvron, H. .... “Data-efficient fine-tuning of vision-language models with prototype-based pseudo-labeling”. arXiv preprint arXiv:2305.14483.

[12] Hinton, G., Vinyals, O., ...(2015). “Distilling the knowledge in a neural network”. arXiv preprint arXiv:1503.02531.

[13] Y. Tian, D. Krishnan, and P. Isola,“Contrastive Representation Distillation” arXiv:1910.10699, 2022.

[14] Romero, A., Ballas, N., Kahou, S. E., Chassang, A. ... (2014). “Fitnets: Hints for thin deep nets” arXiv preprint arXiv:1412.6550.

[15] S. Mirzadeh, M. Farajtabar, ...“Improved Knowledge Distillation via Teacher Assistant” arXiv:1902.03393, 2019.

[16] Gu, X., Lin, T. Y... “Open-vocabulary object detection via vision and language knowledge distillation” arXiv preprint arXiv:2104.13921.

[17] Xu, Y., Zhang, J. ... “EfficientVLM: Efficient vision-language models via cross-modal pruning and knowledge distillation” arXiv preprint arXiv:2303.08368.

[18] J. Li, S. Savarese, and S. C. H. Hoi, “Masked Unsupervised Self-training for Label-free Image Classification” arXiv:2206.02967, 2023.

[19] Z. Guo, R. Zhang ... “CALIP: Zero-Shot Enhancement of CLIP with Parameter-free Attention” arXiv:2209.14169, 2022.

[20] M. J. Mirza, L. Karlinsky, ... “LaFTer: Label-Free Tuning of Zero-shot Classifier using Language and Unlabeled Image Collections” arXiv:2305.18287, 2023.

[21] E. Ali, S. Silva, and M. H. Khan, “DPA: Dual Prototypes Alignment for Unsupervised Adaptation of Vision-Language Models” arXiv:2408.08855, 2024.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov ... “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale” arXiv:2010.11929, 2021.

[23] K. He, X. Zhang, S. Ren, and J. Sun “Deep Residual Learning for Image Recognition” arXiv:1512.03385, 2015.

[24] M. Jia, L. Tang, B.-C. Chen ... “Visual Prompt Tuning” arXiv:2203.12119, 2022.

[25] M. U. Khattak, H. Rasheed ... “MaPLe: Multi-modal Prompt Learning” arXiv:2210.03117, 2023.

[26] M. U. Khattak, S. T. Wasim,... “Self-regulating Prompts: Foundational Model Adaptation without Forgetting” arXiv:2307.06948, 2023.

[27] M. Mistretta, A. Baldrati ... “Improving Zero-shot Generalization of Learned Prompts via Unsupervised Knowledge Distillation” arXiv:2407.03056, 2024.

[28] C. Yang, Z. An, L. Huang ... “CLIP-KD: An Empirical Study of CLIP Model Distillation” arXiv:2307.12732, 2024.

[29] Y. Wang and Y. Chen, “Multi-Modal Cooperative Distillation for Zero-Shot Multi-Label Classification” in (ICIVC), 2025, pp. 112-116.

[30] M. Mistretta, A. Baldrati... “Improving Zero-shot Generalization of Learned Prompts via Unsupervised Knowledge Distillation” arXiv:2407.03056, 2024.

[31] Z. Li, X. Li, L. Yang ... “Curriculum Temperature for Knowledge Distillation” arXiv:2211.16231, 2022.

[32] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, “3D Object Representations for Fine-Grained Categorization” in 2013 IEEE pp. 554-561.

[33] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, “Describing Textures in the Wild” 2014 IEEE pp. 3606-3613, 2013.

[34] P. Helber, B. Bischke, A. Dengel, and D. Borth, “EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification” IEEE Journal Topics in Applied Earth Observations and Remote Sensing, vol. 12, no. 7, pp. 2217-2226, 2019.

[35] Subhransu Maji, Esa Rahtu ... “Fine-Grained Visual Classification of Aircraft” arXiv:1306.5151, 2013.

[36] L. Fei-Fei, R. Fergus, and P. Perona, “Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories” in 2004 Conference on Computer Vision and Pattern Recognition Workshop, pp. 178-178.

[37] Nilsback, Maria-Elena and Zisserman, Andrew,“Automated Flower Classification over a Large Number of Classes” 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, 2008.

[38] L. Bossard, M. Guillaumin, and L. Van Gool,“Food-101 - Mining Discriminative Components with Random Forests” in *Computer Vision - ECCV 2014*, 2014, pp. 446–461.

[39] O. M. Parkhi, A. Vedaldi ... “Cats and dogs” in 2012 IEEE 2012, pp. 3498-3505.

[40] K. Soomro, A. Zamir, and M. Shah, “UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild” arXiv:1212.0402, 2012.