RCDIB: Robust CLIP Distillation via Information Bottleneck

Zhikui Chen*, Yilong Lin, Yuzhe Li, Xingheng Wan, Zhengyang Tang, Meng Liu
School of Software Technology, Dalian University of Technology, Dalian, China
zkchen@dlut.edu.cn, {2441622724, 1594873855, abc2578880617, tang12091517, liumeng}@mail.dlut.edu.cn

***Abstract*—Recently, distillation methods are proposed to balance performance and parameters of large-scale contrastive vision-language models, which gains encouraging progress. However, these distillation methods primarily align the features between the student network and the teacher network with a simple projection layer. This approach is prone to overfitting when fine-tuning on small datasets, leading to degraded distillation performance. To address the limitation discussed above, this paper proposes a robust CLIP distillation with information bottleneck (RCDIB) via cascading a contrastive distillation module and an information bottleneck distillation module. RCDIB explicitly constrains the student network to learn a compact and smooth representation of the teacher's output by imposing an approximate distribution from the true posterior over the latent space, effectively filtering out task-irrelevant noise and enhancing distillation robustness. Extensive experiments demonstrate that RCDIB achieves cutting-edge performance on both cross modal retrieval and zero-shot classification scenarios. The code is publicly available at https://github.com/10ng1000/RCDIB.**

***Index Terms*—knowledge distillation, CLIP, information bottleneck**

I. Introduction

Large-scale contrastive vision-language models, e.g., CLIP [1], have demonstrated remarkable capabilities in learning transferable visual representations by aligning images and text in a shared embedding space. However, these models typically require a considerable number of parameters. To enable efficient deployment on resource-constrained platforms, model distillation [2] techniques have become critical for compressing CLIP while preserving its cross-modal alignment capability. For example, TinyCLIP [3] leverages knowledge distillation from the teacher CLIP model through affinity mimicking and weight inheritance, while employing multistage progressive distillation to minimize informative weight loss during extreme compression. CLIP-KD [4] shows that feature mimicry with Mean Squared Error loss as well as contrastive learning across teacher and student encoders are useful for distillation. FineCLIP [5] enables CLIP to further enhance fine-grained understanding by preserving global contrastive learning while introducing real-time self-distillation for local feature refinement and semantically rich regional contrastive learning with generated region-text pairs.

However, most distillation methods simply mimic the teacher's behavior, when fine-tuning on a relatively small dataset, it leads the student to overfit to the specific training detail and inherit the teacher's feature noise, ultimately compromising zero-shot generalization. To address these limitations and achieve more robust knowledge transfer, we turn to principles from information theory which explicitly manage the information flow during distillation. Recently, the information bottleneck (IB) principle [6] [7] has been widely used to balance model compression and prediction ability by formalizing the trade-off between preserving task-relevant information and compressing redundant input details. In this context, we propose RCDIB, a novel CLIP distillation approach that incorporates an information bottleneck module to regularize the student model's learning process while improving its task-relevant and zero-shot ability. The main contributions are summarized as follows:

- We propose a cascaded dual-distillation framework fusing contrastive and information bottleneck distillation, mitigating overfitting in low-data CLIP fine-tuning by capturing discriminative cross-modal knowledge and filtering task-irrelevant noise.
- A multi-granularity alignment constraint is proposed via minimizing the divergence of feature distributions and maximizing the consistency of sample representations between teacher and student networks, which ensures unbiased transfer of inherent knowledge.
- Extensive experiments validate that RCDIB achieves state-of-the-art performance across Flickr30K/MSCOCO retrieval and ImageNet zero-shot classification, with superior results for both ViT and ResNet-based student models compared to SOTA distillation methods.

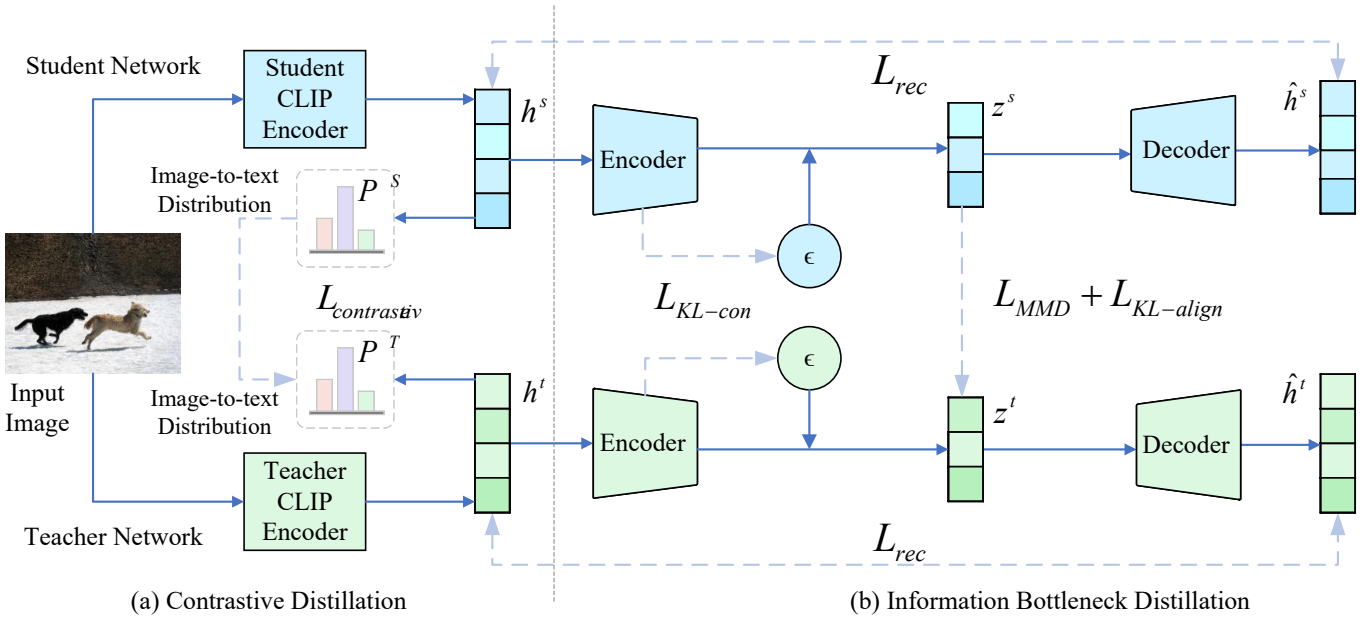
II. Related Work

A. Contrastive Language Image Pretraining

Contrastive language-image pretraining has become a cornerstone for learning transferable multi-modal representations, with CLIP pioneering the paradigm by aligning images and text in a shared embedding space via large-scale contrastive learning. Subsequent works have advanced this framework: ALIGN [8] leverages noisy web-scale image-text data and hard negative mining strategies to strengthen contrastive signals, mitigating the impact of low-quality supervision on representation learning. EVA [9] advances contrastive visual-language pretraining by scaling masked visual representation learning, optimizing training efficiency, and boosting zero-shot

This work was supported by the National Key R&D Program of China under Grant 2025YFF0514800.

*Corresponding author: Zhikui Chen.



(a) Contrastive Distillation (b) Information Bottleneck Distillation

Fig. 1: The overall architecture of our proposed RCDIB method illustrated via its image module. We apply the same structure to the text module to be consistent with the CLIP design.

performance with fewer parameters compared to conventional CLIP models. MobileCLIP [10] achieves fast inference and competitive zero-shot performance via multi-modal reinforced training, boasting fewer parameters and lower latency.

B. Knowledge Distillation

Knowledge distillation (KD) is a core technique for compressing deep neural networks by transferring knowledge from large teacher models to lightweight student models while maintaining performance, pioneered by Hinton et al. [2] via distilling dark knowledge from teacher logits with soft targets. Follow-up studies have extended transferable knowledge into three categories: response-based KD matching final outputs [11] [12], feature-based KD utilizing intermediate layer features maps [13] [14], and relation-based KD capturing inter-feature structural relationships [15] [16]. KD has been broadly applied in computer vision, natural language processing, and speech recognition, facilitating efficient deployment of deep models on resource-constrained devices.

III. Methods

As illustrated in Fig. 1, our core methodology involves constructing a minimal information bottleneck module to extract essential representations from the teacher model, followed by precise alignment between the student and teacher IB modules. This design enables more efficient and accurate knowledge transfer, as the compressed latent space is inherently more compact than the original output dimension. To further enhance the knowledge distillation process, we incorporated contrastive learning as a supplementary mechanism, which strengthens the model's capacity to assimilate knowledge from the teacher network.

A. Theoretical Foundation

We propose a robust distillation method which is grounded in information-theoretic principles and variational inference, with three core theoretical components:

Information Bottleneck Principle. Information bottleneck compression learns representations that preserve the fundamental trade-off between:

$$\min I(z;h) - \beta I(z;y) \tag{1}$$

where $I(\cdot;\cdot)$ denotes the mutual information measure; z denotes the latent code of the model; h denotes the input embeddings (i.e., the final embeddings output by the CLIP model); y denotes the task-specific target labels; and $\beta > 0$ is a positive hyperparameter that balances the two competing terms in the objective function. Specifically, β regulates the trade-off between two objectives: compressing the latent code z (by minimizing $I(z;h)$ to discard input-irrelevant and redundant information from the input embeddings h) and preserving task-relevant information (by maximizing $I(z;y)$ to retain predictive features that are critical for the task targets y).

Variational Learning. The evidence lower bound [17] provides a variational approximation for learning the latent distribution, which can be formulated as:

$$\begin{aligned}
&\log q_\phi(h) \geq \mathcal{L}(\phi,\theta;h) \\
&= \mathbb{E}_{p_\theta(z|h)}[-\log p_\theta(z|h) + \log q_\phi(h,z)] \\
&= -D_{KL}(p_\theta(z|h)|q(z)) + \mathbb{E}_{p_\theta(z|h)}[\log q_\phi(h|z)]
\end{aligned} \tag{2}$$

where p and q are the variational and generative models, respectively, θ and ϕ are the corresponding parameters, and

$q(z)$ serves as the posterior that regularizes the geometry of the latent space. By optimizing the lower bound $\mathcal{L}(\theta, \phi; h)$, we can effectively learn a latent representation that captures the underlying data distribution.

Sufficient Representation Alignment. In the distillation paradigm, a core theoretical foundation lies in the decomposition of mutual information between the student's output and its intermediate representations. In this article, we denote the teacher and student by superscripts T and S, respectively. Specifically, the identity:

$$I_\theta(z^S; h^S) = I_\theta(z^S; h^S|h^T) + I_\theta(z^S; h^T) \tag{3}$$

holds true, and we provide a rigorous proof sketch for this result as follows:

Proof Sketch:

1) *Chain Rule Application*: By the fundamental chain rule of mutual information, the joint mutual information can be decomposed as:

$$I_\theta(z^S; h^S, h^T) = I_\theta(z^S; h^T) + I_\theta(z^S; h^S|h^T). \tag{4}$$

2) *Distillation Setting Assumption*: In the knowledge distillation framework, our primary objective is to train the student model so that its intermediate representation h^S fully captures all information about the student's output z^S that is present in the teacher's representation h^T. Formally, this means that once h^S is known, h^T provides no additional information about z^S, i.e.,

$$I_\theta(z^S; h^T|h^S) = 0. \tag{5}$$

Applying the chain rule in reverse to the joint mutual information $I_\theta(z^S; h^S, h^T)$, we obtain:

$$\begin{aligned} I_\theta(z^S; h^S, h^T) &= I_\theta(z^S; h^S) + I_\theta(z^S; h^T|h^S) \\ &= I_\theta(z^S; h^S). \end{aligned} \tag{6}$$

3) *Final Substitution*: Substituting the result from Step 2 (Eq. (6)) into the decomposition from Step 1 (Eq. (4)), we directly derive the desired identity:

$$I_\theta(z^S; h^S) = I_\theta(z^S; h^S|h^T) + I_\theta(z^S; h^T). \tag{7}$$

The first term in Eq. (3) represents redundant information specific to the student, which cannot be predicted from the teacher's representation h^T and is thus deemed irrelevant to the distillation task. In our robust distillation framework, this private student information introduces unnecessary noise and hinders the transfer of task-relevant knowledge, so we aim to suppress it. The second term captures task-relevant information shared with the teacher, which is critical for preserving the robust, generalizable knowledge encoded in the teacher CLIP model. To achieve sufficient representation alignment that retains only the task-relevant shared information, we optimize:

$$\max_\theta \left[-I_\theta(z^S; h^S|h^T) + I_\theta(z^S; h^T)\right] \tag{8}$$

We derive a tractable lower bound for this objective through the following steps. For the first term, we have:

$$\begin{aligned} & - I_\theta(z^S; h^S|h^T) \\ &= -\mathbb{E}_{h^S, h^T, z^S}\left[\log \frac{p_\theta(z^S|h^S, h^T)}{p_\theta(z^S|h^T)}\right] \\ &= -\mathbb{E}_{h^S, h^T}\left[D_{KL}(p_\theta(z^S|h^S, h^T)\|p_\theta(z^S|h^T))\right] \\ &\geq -\mathbb{E}_{h^S, h^T}\left[D_{KL}(p_\theta(z^S|h^S)\|p_\theta(z^T|h^T))\right] \\ &= -D_{KL}(p_\theta(z^S|h^S)\|p_\theta(z^T|h^T)) \end{aligned} \tag{9}$$

For the second term, applying the chain rule of mutual information:

$$I_\theta(z^S; h^T) = I_\theta(z^S; z^T) + I_\theta(z^S; h^T|z^T) \tag{10}$$

$$\geq I_\theta(z^S; z^T) \tag{11}$$

We then obtain our distribution alignment lower bound:

$$\begin{aligned} & \max_\theta \left[-I_\theta(z^S; h^S|h^T) + I_\theta(z^S; h^T)\right] \\ & \geq \max_\theta \left[-D_{KL}(p_\theta(z^S|h^S)\|p_\theta(z^T|h^T)) + I_\theta(z^S; z^T)\right] \end{aligned} \tag{12}$$

This lower bound is computable as it replaces the intractable conditional mutual information with a KL divergence and mutual information between latent representations z^S and z^T.

B. Contrastive Distillation

To learn from the teacher in a comprehensive way, we first train the student model to learn the low-level feature of the teacher with contrastive learning. We compute the image-to-text and text-to-image distributions of the teacher:

$$P^T(c|v) = \frac{\exp(\text{sim}(f_V^T(v), f_C^T(c))/\tau)}{\sum_{c' \in C} \exp(\text{sim}(f_V^T(v), f_C^T(c'))/\tau)} \tag{13}$$

$$Q^T(v|c) = \frac{\exp(\text{sim}(f_C^T(c), f_V^T(v))/\tau)}{\sum_{v' \in V} \exp(\text{sim}(f_C^T(c), f_V^T(v'))/\tau)} \tag{14}$$

where V and C denote the sets of images and texts, respectively; v and c represent a given image sample and a given text sample; $f_V^T(v)$ and $f_C^T(c)$ denote the image encoder and text encoder of the teacher model, respectively; $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity function; and τ is the temperature parameter. The student's distributions are calculated in a similar way, which is denoted by $P^S(c|v)$ and $Q^S(v|c)$.

We then employ KL divergence to measure the distribution discrepancy between the student and teacher models:

$$\begin{aligned} \mathcal{L}_{KL} &= \text{KL}(P^T\|P^S) + \text{KL}(Q^T\|Q^S) \\ &= \sum_{v \in V}\sum_{c \in C} P^T(c|v) \log \frac{P^T(c|v)}{P^S(c|v)} \\ &\quad + \sum_{c \in C}\sum_{v \in V} Q^T(v|c) \log \frac{Q^T(v|c)}{Q^S(v|c)} \end{aligned} \tag{15}$$

C. Information Bottleneck Distillation

In the limited-data scenario, using contrastive learning alone is prone to overfitting, because it is easily affected by the specific training datasets. In order to reduce noise and therefore address this problem, we train the information bottleneck on both the student and teacher networks to constrain the model and learn from the sufficient representations.

For variational learning (Eq. (2)), we reconstruct the original representation by minimizing the MSE loss of the reconstructed embedding and the original CLIP embedding:

$$\mathcal{L}_{\text{recon}} = \|h^{\mathcal{T}} - q_\phi^T(z^{\mathcal{T}})\|_2^2 + \|h^{\mathcal{S}} - q_\phi^S(z^{\mathcal{S}})\|_2^2 \tag{16}$$

We use the KL divergence term to constraint the model, ensuring that the distribution of latent embedding is consistent with their intractable true posterior:

$$\mathcal{L}_{\text{KL-con}} = D_{\text{KL}}(p_\theta^T(z|h^{\mathcal{T}})\|q(z)) + D_{\text{KL}}(p_\theta^S(z|h^{\mathcal{S}})\|q(z)) \tag{17}$$

To sufficiently transfer knowledge between teacher and student networks, we implement Eq. (11) by matching their latent distributions using Maximum Mean Discrepancy (MMD) [18] and KL Divergence. It measures the distance between distributions in a reproducing kernel Hilbert space:

$$\mathcal{L}_{\text{MMD}} = \left\|\mathbb{E}[\phi(z^{\mathcal{T}})] - \mathbb{E}[\phi(z^{\mathcal{S}})]\right\|_{\mathcal{H}}^2 \tag{18}$$

where $\phi(\cdot)$ is the feature mapping associated with the characteristic kernel $k(z_i, z_j) = \langle \phi(z_i), \phi(z_j) \rangle_{\mathcal{H}}$. We implement this with an RBF kernel:

$$k(z_i, z_j) = \exp\left(-\frac{\|z_i - z_j\|^2}{2\sigma^2}\right) \tag{19}$$

Additionally, we employ a KL divergence to align the distribution parameters as in Eq. (9):

$$\mathcal{L}_{\text{KL-align}} = D_{KL}(p_\theta(z^S|h^S)\|p_\theta(z^T|h^T)) \tag{20}$$

This combined approach leverages the nonparametric strengths of MMD for capturing complex distributional similarities while utilizing KL divergence for exact probabilistic alignment, enabling more comprehensive knowledge transfer. In this approach, we implement Eq. (8) for sufficient representation alignment.

D. Overall Objective

The training objective combines all the aforementioned components with the balancing hyperparameters λ_1–λ_5:

$$\begin{aligned}\mathcal{L}_{\text{RCDIB}} = \mathcal{L}_{\text{recon}} + \lambda_1\mathcal{L}_{\text{KL-con}} + \lambda_2\mathcal{L}_{\text{MMD}} \\ + \lambda_3\mathcal{L}_{\text{KL-align}} + \lambda_4\mathcal{L}_{\text{contrast}}\end{aligned} \tag{21}$$

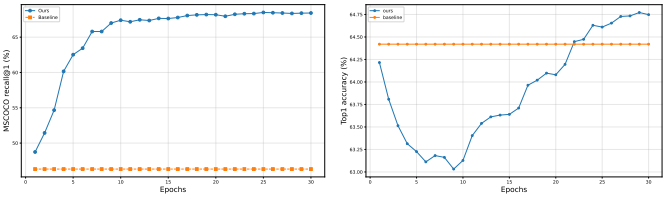
These objectives complement each other by operating in both the latent bottleneck space and the original output space.

IV. Experiments

A. Experimental Setup

Datasets and Metrics. We adapt the widely used Flickr30K [19] and MSCOCO [20]datasets for distillation training, following the common practice in CLIP distillation literature. Specifically, we sample 50,000 image-text pairs for training, simulating a low-resource fine-tuning scenario. For cross-modal retrieval, we report results on the full test splits of Flickr30K and MSCOCO (1,000 test images for Flickr30K and 5,000 test images for MSCOCO). For zero-shot classification, we evaluate on ImageNet-1K [21] and its variants, including ImageNet-V2 (IN-V2) [22], ImageNet-Sketch (IN-S) [23], and ImageNet-R (IN-R) [24], to comprehensively assess model generalization.

Training Settings. All experiments use the ViT-L-14 model pre-trained on DataComp-1B [25], [26] as the teacher. The visual and text encoders use the default configurations of open_clip [26] codebase setting. We set the hyperparameters for the loss function as follows: $\lambda_1 = 1$, $\lambda_2 = 0.01$, $\lambda_3 = 10$, and $\lambda_4 = 1$. We train the student model for 30 epochs using a total batch size of 64 (on 2 V100 GPUs) with the AdamW optimizer [27]. The learning rate is set to 1e-6 with a cosine decay schedule and a weight decay of 0.1.



(a) Text-to-image retrieval (b) Zero-shot classification

Fig. 2: Results against training epochs.

B. Comparisons with State-of-the-arts

We compare our distilled models with state-of-the-art CLIP variants in both the retrieval and zero-shot classification tasks. As shown in Table I, under 50,000 training data, our distilled CLIP models gain significant improvement in task-specific cross-modal retrieval while maintaining zero-shot classification ability, better than the SOTA methods. Specifically, the ViT-based variants not only attain the most significant improvement in task-specific retrieval, but also performs better than the student CLIP in ImageNet zero-shot classification. We do not provide the results of ResNet [28] for TinyCLIP, as it requires the teacher and student networks to share the same encoder architecture. While featuring a different image encoder architecture from the teacher model, our approach brings the model with ResNet encoder close to the zero-shot classification capability of the base CLIP model. Notably, our approach incorporates lightweight IB modules with merely 5M parameters, which is significantly smaller than the ViT-B/16 student with 155M parameters. This minimal architectural addition incurs only marginal computational and memory

TABLE I: Performance comparisons with the state-of-the-art methods. We evaluate cross-modal retrieval by Recall@1 (R@1) and zero-shot classification by top-1 and top-5 accuracy. The best results for each image encoder are in **bold**.

Method	Image Encoder	MSCOCO		Flickr30K		ImageNet-1K		ImageNet-R		ImageNet-S		ImageNetV2	
		I→T	T→I	I→T	T→I	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
CLIP	ViT-L-14	74.22	68.06	97.20	97.40	78.11	95.51	90.30	97.70	66.89	88.73	70.79	92.30
CLIP	ViT-B/16	53.10	46.32	82.70	80.40	64.42	89.72	73.51	90.99	44.21	72.08	57.86	85.02
CLIP-KD	ViT-B/16	70.62	66.44	92.80	89.30	38.09	67.97	47.54	73.85	23.17	47.07	33.63	61.67
TinyCLIP	ViT-B/16	75.40	67.34	95.50	95.20	65.29	90.03	74.80	90.91	45.44	73.10	59.56	85.68
RCDIB(Ours)	ViT-B/16	**76.04**	**68.32**	**96.20**	**95.40**	**67.05**	**91.54**	**76.18**	**92.48**	**46.29**	**74.55**	**60.57**	**87.16**
CLIP	ViT-B/32	50.16	42.24	74.40	72.50	59.59	86.34	67.06	87.18	38.93	66.90	52.87	80.89
CLIP-KD	ViT-B/32	66.70	61.80	88.80	86.80	30.46	59.08	37.40	61.80	18.62	40.41	26.30	52.71
TinyCLIP	ViT-B/32	71.56	63.78	92.10	89.80	60.11	86.32	66.11	85.73	38.93	66.39	52.98	81.05
RCDIB(Ours)	ViT-B/32	**73.72**	**67.26**	**93.60**	**91.00**	**61.32**	**87.91**	**67.53**	**87.40**	**40.33**	**68.04**	**54.02**	**82.68**
CLIP	RESNET101	58.50	44.86	86.30	79.10	60.45	87.31	**67.64**	**87.76**	39.83	67.75	54.24	82.28
CLIP-KD	RESNET101	67.56	60.66	90.60	88.40	39.86	70.44	48.53	74.29	23.77	48.58	35.13	64.78
RCDIB(Ours)	RESNET101	**72.98**	**65.36**	**94.00**	**92.30**	**60.69**	**87.67**	67.32	87.68	**40.29**	**68.10**	**54.38**	**82.35**
CLIP	RESNET50	52.78	46.32	85.50	89.40	57.90	85.50	**60.26**	**83.14**	34.61	62.54	50.85	80.35
CLIP-KD	RESNET50	66.54	57.28	90.40	88.40	41.70	72.18	42.24	69.86	20.59	45.33	36.42	66.66
RCDIB(Ours)	RESNET50	**70.40**	**62.44**	**92.80**	**95.00**	**58.21**	**86.10**	59.51	83.10	**34.79**	**62.94**	**51.15**	**80.76**

overhead, while contributing meaningfully to the retrieval performance and zero-shot generalization.

To further validate the training dynamics and stability of RCDIB, we plot the top-1 accuracy on ImageNet-1K zero-shot classification and Recall@1 on MSCOCO text-to-image retrieval against training epochs, with ViT-B/16 as the baseline model, as shown in Fig. 2. In the early training epochs, the model prioritizes the improvement of task-specific retrieval capability, which is accompanied by a temporary decline in zero-shot classification performance. As the training proceeds, the information bottleneck module gradually exerts its regularization effect, effectively alleviating the overfitting phenomenon. Consequently, the zero-shot classification performance of RCDIB rebounds and surpasses the baseline model in the later epochs. This training behavior demonstrates that RCDIB can dynamically balance the learning of task-specific knowledge and generalizable representations, achieving superior overall performance in both cross-modal retrieval and zero-shot classification tasks.

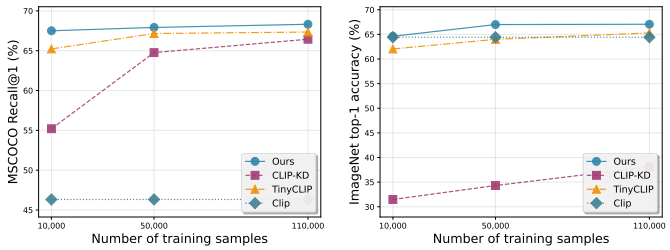
C. Ablation Study

We analyze the contribution of each loss of RCDIB with ViT-B/16 as the student image encoder. Table II shows that the addition of $\mathcal{L}_{\text{contrast}}$ brings substantial gains in retrieval, confirming its role in transferring discriminative knowledge. Further incorporating the other loss components brings additional gains in zero-shot ability, indicating that the IB module complements the contrastive loss by enhancing the feature alignment and generalization capability.

Notably, removing $\mathcal{L}_{\text{KL-con}}$ causes a slight degradation in zero-shot accuracy, suggesting that the variational constraint is essential for learning smoother and more robust latent representations, which in turn improves out-of-distribution generalization. When Ablating $\mathcal{L}_{\text{contrast}}$, although the student remains a strong zero-shot capability, it only has limited improvement in the retrieval task. Using only $\mathcal{L}_{\text{KL-align}}$ or $\mathcal{L}_{\text{MMD}}$ for alignment also weakens the performance. The complete combination of losses proves essential, enabling the student to achieve strong performance in both retrieval and zero-shot tasks while maintaining high generalization capability.

TABLE II: Ablation study on the objective components.

$\mathcal{L}_{\text{recon}}$	$\mathcal{L}_{\text{KL-con}}$	$\mathcal{L}_{\text{MMD}}$	$\mathcal{L}_{\text{KL-align}}$	$\mathcal{L}_{\text{contrast}}$	MSCOCO		ImageNet-1K	
					I→T	T→I	Top1	Top5
				✓	75.38	67.64	65.10	90.54
✓	✓	✓	✓		75.10	67.72	66.96	91.46
✓		✓	✓	✓	75.86	67.96	66.34	91.03
✓	✓		✓	✓	75.96	68.04	65.90	90.34
✓	✓	✓		✓	75.94	68.02	66.76	91.34
✓	✓	✓	✓	✓	**76.04**	**68.32**	**67.05**	**91.54**



(a) Text-to-image retrieval (b) Zero-shot classification

Fig. 3: Model performance on datasets in three different scales.

D. Comparisons on Scaled Dataset

To further validate the scalability and robustness of our proposed RCDIB method under different data scales, we conduct extensive experiments by varying the number of training samples from 10,000 to 110,000, with ViT-B/16

as the student model. As depicted in Fig. 3, our RCDIB method consistently outperforms all competing methods across different training data scales. For zero-shot classification on ImageNet-1K, RCDIB maintains the highest top-1 accuracy even when the training data is limited to 10,000 samples. As the number of training samples increases, the performance gap between RCDIB and other methods remains stable, which demonstrates the effectiveness of our information bottleneck module in preserving task-relevant knowledge and avoiding overfitting. For text-to-image retrieval on MSCOCO, RCDIB achieves the best Recall@1 scores across all data scales. The advantage is particularly significant in the low-data regime, which verifies that our contrastive distillation combined with information bottleneck can effectively transfer discriminative cross-modal alignment capabilities from the teacher model to the student model.

V. Conclusion

In this paper, we proposed RCDIB, a novel CLIP distillation framework that leverages information bottleneck to preserve the generalization capabilities of the larger CLIP model. By cascading contrastive distillation and information bottleneck modules, our method effectively filters out task-irrelevant noise and mitigates overfitting issues in low-data fine-tuning scenarios. Extensive experiments show that the distilled models perform strongly in retrieval and zero-shot tasks with limited training data, outperforming state-of-the-art distillation methods. This provides an efficient solution for deploying lightweight yet robust vision-language models on resource-constrained platforms.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[2] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[3] K. Wu, H. Peng, Z. Zhou, B. Xiao, M. Liu, L. Yuan, H. Xuan, M. Valenzuela, X. S. Chen, X. Wang *et al.*, "Tinyclip: Clip distillation via affinity mimicking and weight inheritance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 970–21 980.

[4] C. Yang, Z. An, L. Huang, J. Bi, X. Yu, H. Yang, B. Diao, and Y. Xu, "Clip-kd: An empirical study of clip model distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 952–15 962.

[5] D. Jing, X. He, Y. Luo, N. Fei, W. Wei, H. Zhao, Z. Lu *et al.*, "Fineclip: Self-distilled region-based clip for better fine-grained understanding," *Advances in Neural Information Processing Systems*, vol. 37, pp. 27 896–27 918, 2024.

[6] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[7] J. Gao, M. Liu, P. Li, A. A. Laghari, A. R. Javed, N. Victor, and T. R. Gadekallu, "Deep incomplete multiview clustering via information bottleneck for pattern mining of data in extreme-environment iot," *IEEE Internet of Things Journal*, vol. 11, no. 16, pp. 26 700–26 712, 2024.

[8] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.

[9] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva: Exploring the limits of masked visual representation learning at scale," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 358–19 369.

[10] P. K. A. Vasu, H. Pouransari, F. Faghri, R. Vemulapalli, and O. Tuzel, "Mobileclip: Fast image-text models through multi-modal reinforced training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 963–15 974.

[11] G. Lu, H. Yin, Z. Shu, J. Wang, and G. Luo, "Bdckd: Unlocking the power of brownian distance covariance in knowledge distillation," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[12] Y. Ding, G. Yang, S. Yin, J. Zhang, X. Fang, and W. Yang, "Generous teacher: Good at distilling knowledge for student learning," *Image and Vision Computing*, vol. 150, p. 105199, 2024.

[13] T. Liu, C. Chen, X. Yang, and W. Tan, "Rethinking knowledge distillation with raw features for semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 1155–1164.

[14] J. Yuan, M. H. Phan, L. Liu, and Y. Liu, "Fakd: Feature augmented knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 595–605.

[15] T. Huang, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge distillation from a stronger teacher," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 716–33 727, 2022.

[16] C. Wang, J. Zhong, Q. Dai, Y. Qi, Q. Yu, F. Shi, R. Li, X. Li, and B. Fang, "Channel correlation distillation for compact semantic segmentation," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 37, no. 03, p. 2350004, 2023.

[17] D. P. Kingma, M. Welling *et al.*, "Auto-encoding variational bayes," 2013.

[18] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel method for the two-sample-problem," *Advances in neural information processing systems*, vol. 19, 2006.

[19] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the association for computational linguistics*, vol. 2, pp. 67–78, 2014.

[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[22] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International conference on machine learning*. PMLR, 2019, pp. 5389–5400.

[23] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning robust global representations by penalizing local predictive power," *Advances in neural information processing systems*, vol. 32, 2019.

[24] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8340–8349.

[25] S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang *et al.*, "Datacomp: In search of the next generation of multimodal datasets," *Advances in Neural Information Processing Systems*, vol. 36, pp. 27 092–27 112, 2023.

[26] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Openclip," Jul. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5143773

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.