A Progressive Framework of Vision-language Knowledge Distillation and Alignment for Multilingual Scene

Wenbo Zhang [a 1], **Yifan Zhang** [a 1], **Jianfeng Lin** [b], **Binqiang Huang** [b], **Jinlu Zhang** [b], **Wenhao Yu** [a c *]

[a]School of Geography and Information Engineering, China University of Geosciences, Wuhan, China
[b]Meituan, Beijing, China
[c]National Engineering Research Center for Geographic Information System, China University of Geosciences, Wuhan, China
[1]These authors contributed equally
[*]Corresponding author, email: yuwh@cug.edu.cn

Abstract

Pre-trained vision-language (V-L) models such as CLIP have shown excellent performance in many downstream cross-modal tasks. However, most of them are only applicable to the English context. Subsequent research has focused on this problem and proposed improved models, such as CN-CLIP and AltCLIP, to facilitate their applicability to Chinese and even other languages. Nevertheless, these models suffer from high latency and a large memory footprint in inference, which limits their further deployment on resource-constrained edge devices. In this work, we propose a conceptually simple yet effective multilingual CLIP Compression framework and train a lightweight multilingual vision-language model, called DC-CLIP, for both Chinese and English context. In this framework, we collect high-quality Chinese and English text-image pairs and design two training stages, including multilingual vision-language feature distillation and alignment. During the first stage, lightweight image/text student models are designed to learn robust visual/-multilingual textual feature representation ability from corresponding teacher models, respectively. Subsequently, the multilingual vision-language alignment stage enables effective alignment of visual and multilingual textual features to further improve the model's multilingual performance. Comprehensive experiments in zero-shot image classification, conducted based on the ELEVATER benchmark, showcase that DC-CLIP achieves superior performance in the English context and competitive performance in the Chinese context, even with less training data, when compared to existing models of similar parameter magnitude. The evaluation demonstrates the effectiveness of our designed training mechanism.

Introduction

Learning a good representation in the combined domain of vision and language remains a pivotal goal of foundational model research. The recent release of CLIP (Radford et al. 2021) by OpenAI is a landmark work in the field of visual language. Trained on extensive image-text pairing data, CLIP learns to map images and arbitrary textual descriptions into a shared feature space, thereby enabling the model to comprehend and process a variety of visual tasks, including image classification (Zhou et al. 2022; Zhang et al. 2021b; Khattak et al. 2023), object detection (Gu et al. 2021; Li et al. 2022c; Zhang et al. 2022a), and semantic segmentation (Li et al. 2022a; Xu et al. 2022), without necessitating task-specific training data. This pre-training methodology significantly enhances the model's generalization capabilities and adaptability, facilitating its application across a broad spectrum of visual understanding tasks and showcasing its remarkable zero-shot transfer capability across various downstream tasks.

However, CLIP's reliance on English text data for training introduces performance limitations in non-English linguistic contexts, prompting research into adapting CLIP for diverse language environments. For instance, Chinese CLIP (Yang et al. 2022a) leverages 123 million Chinese image-text pairs to retrain the model, improving the understanding of Chinese visual content. Multilingual CLIP (Carlsson et al. 2022), meanwhile, extends CLIP's applicability to multilingual settings by training multilingual text encoders with machine translation techniques and cross-linguistic teacher-learning methods. AltCLIP (Chen et al. 2022) employs teacher learning and contrastive learning strategies to learn bilingual and multilingual multimodal representation models.

Despite the considerable strides made with the CLIP model and its adaptation under multilingual environment (Yang et al. 2022a; Carlsson et al. 2022; Chen et al. 2022; Zhang et al. 2022b), research on its effective deployment to mobile or edge devices remains scant (Fu et al. 2020; Lin et al. 2023; Mukherjee et al. 2022). This gap primarily arises from the intrinsic constraints of such devices, including limited computational capacity, constrained storage, and stringent energy requirements, which pose formidable challenges to the design and deployment of efficient deep learning models, particularly complex multimodal ones like CLIP. In response, there is a pressing need for novel methodologies and technologies to optimize model size and computational demands, ensuring their operability on resource-constrained devices without compromising performance.

As one of the effective techniques for compressing large models, knowledge distillation (Hinton, Vinyals, and Dean 2015; Buciluǎ, Caruana, and Niculescu-Mizil 2006) involves injecting knowledge from a powerful Teacher model into a smaller Student model without sacrificing much of its generalization ability. It has been extensively researched and

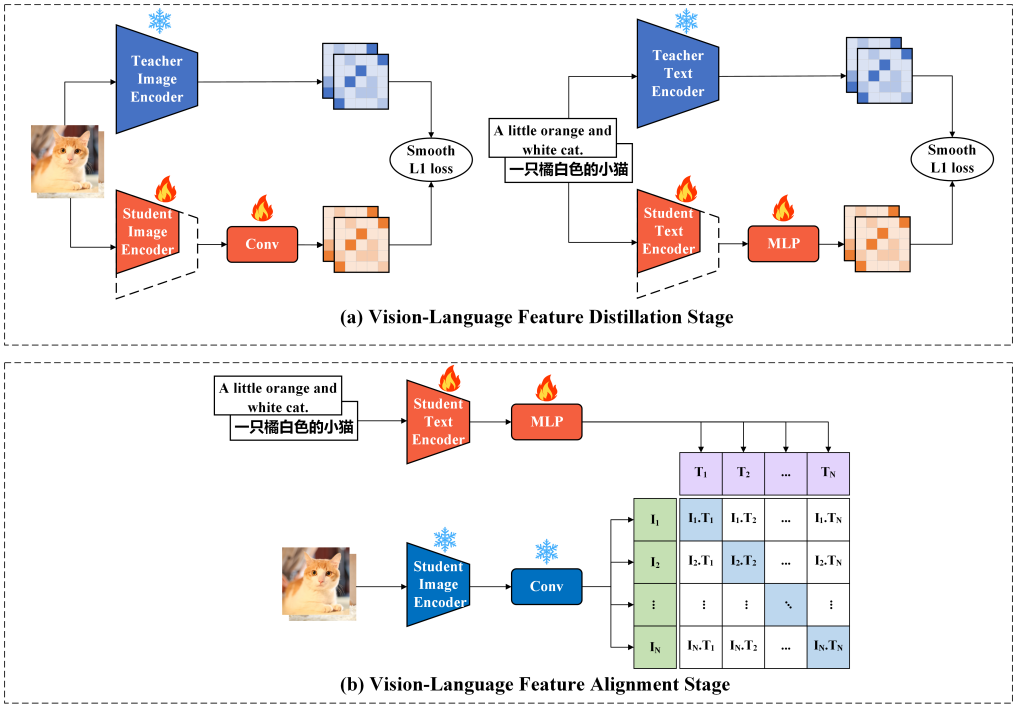

Figure 1: The framework of the proposed method comprises two progressive modules: (a) the vision-language feature distillation module and (b) the vision-language feature alignment module. In the first stage, module (a) extracts key knowledge from the teacher model's image and text encoders and transfers it to the student model. Subsequently, module (b) further aligns the vision-language feature using contrastive learning strategy to enhance the model's performance.

applied in the single-modal environment (Gou et al. 2021). However, its potential in the multimodal domain remains largely untapped. Unlike single-modal models, distilling multimodal models, such as language-image cross-modal models like CLIP, presents distinct challenges. Firstly, these multimodal models, similar to CLIP, typically consist of two branches: an image encoder and a text encoder (Jia et al. 2021; Singh et al. 2022; Thomee et al. 2016; Yuan et al. 2021). When distilling such multi-branch models, it is crucial to consider the information interaction between different modal branches in both the Teacher and Student models. Secondly, the original CLIP model was pre-trained on 400 million image-text pairs for 32 epochs, requiring thousands of GPU-days, which poses a significant challenge for distillation when computational resources are limited.

Addressing this challenge, our study introduces DC-CLIP, a novel, lightweight variant of the CLIP model devised through feature distillation (Srivastava, Greff, and Schmidhuber 2015) and contrastive learning (He et al. 2020). Our approach is geared towards significantly reducing the model's demand for computational resources and storage space without sacrificing the model's performance, and at the same time realizing the application of CLIP in both the Chinese and English contexts. Specifically, the DC-CLIP consists of two main stages: the vision-language feature distillation stage and the vision-language alignment stage. Initially, we take the Chinese and English versions of AltCLIP as teacher models and distill their image encoders and text encoders respectively, which enables us to extract the key knowledge from the original AltCLIP model with a large number of parameters and transfer it effectively to a smaller and more efficient model architecture. Subsequently, the model undergoes refinement through contrastive learning to enhance its precision in image-text alignment within a multilingual framework. This phase predominantly relies on a modest corpus of Chinese and English image-text pairs, optimizing the model via contrastive loss to facilitate image-text matching and comprehension in multilingual contexts.

Through this methodology, DC-CLIP maintains robust performance in Chinese and English settings while adapting to the resource limitations of mobile devices such as smartphones. This breakthrough not only epitomizes technological innovation in multilingual vision-language models but also carves new avenues for deploying efficient and pragmatic multimodal model applications on a plethora of mobile and edge devices in the future. With DC-CLIP, we anticipate offering users a more seamless and effective visual language interaction experience across multilingual landscapes, thereby broadening the applicability and adoption of the CLIP model and its derivatives on mobile platforms.

Our contributions are as follows:

- Adaptability to English and Chinese: Innovatively adapts models for both English and Chinese contexts, enhancing linguistic inclusivity and comprehension across these languages.
- Feature Distillation: Employs output feature maps from

pre-trained models as distillation targets, offering a richer knowledge transfer than traditional logits or simple vectors.

- Efficiency with Limited Data: We organized a small-scale, high-quality dataset of Chinese and English image-text pairs. Following training, our model demonstrates competitive performance in the Chinese context compared to the baselines and achieves state-of-the-art performance in the English context. This showcases the effectiveness and practicality of our approach in scenarios with limited data availability.

Related work

The CLIP (Radford et al. 2021) model, pre-trained with large-scale image-text pairs dataset, can effectively align image and text representations in a shared feature space, showcasing strong zero-shot learning performance across various visual tasks. Yet, the fact that its training data is mainly in English leads to a degradation of its performance when processing downstream tasks in other languages. To extend the CLIP model for languages, researchers have carried out various attempts, including customized versions of CLIP for specific languages, such as adaptations for Italian (Bianchi et al. 2021) and Chinese (Changpinyo et al. 2021; Yang et al. 2022a; Zhang et al. 2022b; Gu et al. 2022), augmenting the model's multilingual capability through large-scale bilingual contrastive learning and extending CLIP's multilingual capability under small scale data by employing both teacher learning and contrastive learning approaches. These works have greatly advanced CLIP's application in multilingual downstream tasks, yet they have overlooked the models' high memory usage, which hinders their effective deployment on edge devices in practical, everyday scenarios.

Recently, with the continuous evolution and expansion of deep learning models, the research on deploying large-scale models to edge devices has increasingly become a focal point of interest. Within this domain, Knowledge Distillation (KD) (Hinton, Vinyals, and Dean 2015) has been affirmed as a promising technique for compressing large models (referred to as 'teacher models') into smaller versions (known as 'student models') that require fewer parameters and computational power while still achieving competitive results in downstream tasks. Distillation within singular modalities, such as vision or language, has been extensively explored. For example, in the field of image distillation, KDFM (Chen, Chang, and Lee 2019) proposes a method to improve the effect of knowledge distillation by learning the feature mapping of a teacher network and demonstrates that the model size can be significantly reduced while maintaining the model accuracy by using generative adversarial networks and shared classifiers. This demonstrates that in image processing tasks, distillation can not only compress the model but also maintain or even improve the performance by finely shifting the feature representation. In the text domain, Xu et al. (Xu et al. 2024) summarize how knowledge distillation makes it possible to transfer knowledge from large language models to smaller models, easing the burden of model deployment while preserving language comprehension and generation. These scholarly works provide robust pathways for deploying CLIP and its variants in other languages onto edge devices, providing a sound basis for model refinement for practical deployment in multilingual scenarios.

While considering the challenge of effectively aligning image and text features after the distillation of image and text encoders, we have incorporated insights from related work on aligning these features, adopting a contrastive learning approach to further improve the model's performance. For instance, XMC-GAN (Zhang et al. 2021a) employs contrastive loss to enhance mutual information between images and texts, ensuring not only the realism of images but also their consistency with their descriptions. Moreover, UniCL (Yang et al. 2022b) extends upon this with a bidirectional learning objective, utilizing relationships within image-text-label triplets to optimize alignment across modalities. These methods are especially relevant to our goal of refining distilled features following the distillation of image and text encoders, to ensure their consistency and effective pairing. By understanding these contrastive learning strategies, we aim to address potential alignment issues encountered during distillation and improve the model's overall performance in multilingual settings.

Method

In this work, we aim to ensure the model's adaptability to both English and Chinese contexts by employing the English and Chinese versions of AltCLIP as the teacher model. We introduce an innovative two-stage approach to compress the image-language representation model, making it suitable for deployment on edge devices in multilingual environments. Initially, following the methodology of Wei et al. (Wei et al. 2022), we utilize feature distillation for both the image and text encoders separately. This initial step is crucial for maintaining the teacher model's performance while significantly reducing the model's size and computational demands. Next, we engage in further aligning the image and text features through contrastive learning, using small-scale image-text pairs dataset. Our comprehensive training procedure is depicted in Figure 1.

Vision-Language Feature Distillation

Distilled Image Encoder During the feature distillation phase for the image encoder, as shown on the left side of Figure 1a, we chose the image encoder of AltCLIP as the teacher model to enable the student model to learn robust image encoding abilities. Additionally, to reduce training time, we opted for the ResNet50 (He et al. 2016) image encoder trained in CLIP as the student model. This phase leverages the teacher model's superior image-understanding capabilities to guide the student model's learning process. Diverging from traditional logic distillation, our strategy employs output feature maps from pre-trained models as distillation targets, allowing the use of any pre-trained model, even those without Logit output. To ensure the comparability of feature maps between teacher and student models, we apply the same augmented view to each raw image. Additionally, a 1 $\times$ 1 convolution layer is applied on top of the student net-

Table 1: Description of datasets

Datasets	Usage	Component
ImageNet	Generic-objects datasets	Humans, animals, vehicles, food, etc.
Caltech101	Generic-objects datasets	Animals, plants, transportation, furniture, etc.
OxfordPets	Pet classification	Different breeds of cats, dogs, etc.
StanfordCars	Automobile classification	Different car makes and models
Flowers102	Flower classification	Varieties of different flowers such as roses, sunflowers, etc.
Food101	Food image classification	Pizza, burgers, sushi, etc.
FGVCAircraft	Fine-grained image classification	Aircraft of different manufacturers and models
EuroSAT	Classification of geological features	Different types of remote sensing images of cities, forests, etc.

Table 2: This table presents a comparative analysis of the performance metrics for the baseline models, original CLIP, teacher model AltCLIP, and DC-CLIP method across eight English datasets. The **best** and second-best results between our model and the baselines are prominently highlighted.

Method	Type	ImageNet	Caltech101	FGVCAircraft	StanfordCars	EuroSAT	Food101	Flowers102	OxfordPets
CLIP(224M)	Original	58.19	80.40	16.95	53.94	38.16	80.81	66.02	85.77
AltCLIP(3.22G)	Large-size	73.83	94.72	29.91	77.19	62.79	90.85	75.72	93.43
CNCLIP(294M)	Tiny-size	11.038	58.09	3.75	20.75	16.2	24.12	13.14	**19.54**
TaiyiCLIP(391M)		14.84	52.87	5.94	**21.45**	13.28	23.71	7.91	10.38
ChineseCLIP(726M)		2.064	23.36	0.93	2.69	10.26	3.35	1.72	4.87
DC-CLIP(301M)		**42.79**	**73.75**	**6.24**	16.54	**64.32**	**46.91**	**26.43**	17.23

work, allowing different dimensions of output feature maps between the teacher and student models.

To precisely gauge feature consistency between teacher and student models, we implement a smooth L1 loss function (Girshick 2015), offering enhanced robustness against outliers compared to conventional L1 loss. This function effectively mitigates potential overfitting issues in the distillation process and incorporates an adjustable hyperparameter for linear attenuation at minor errors, thereby bolstering training process stability and efficiency. Through this distillation approach, the student model assimilates a more nuanced and enriched image representation. The distillation loss is as follows:

$$\begin{aligned} f_s &= g\left(ImageEncoder_{student}\left(I\right)\right) \\ f_t &= ImageEncoder_{teacher}\left(I\right) \\ Loss_i &= SmoothL1\left(f_s, f_t\right) \end{aligned} \tag{1}$$

where I is the input image and g is the convolution layer of 1×1.

Distilled Text Encoder In the distillation of text encoder stage, we distill AltCLIP's text encoder using roughly the same distillation strategy as in the distillation of image encoder stage. The difference is that here, considering that the model needs to be adapted to both Chinese and English contexts, we use the mini-model obtained from the distillation of the XLM-R (Conneau et al. 2019) model by Wang et al. (Wang et al. 2020) as the student model. Meanwhile, to convert the output of the student model to the same output dimension as the teacher encoder, we add a fully connected layer behind the student model. As shown on the right side of Figure 1a, the distillation loss is as follows:

$$\begin{aligned} W_s &= L\left(TextEncoder_{student}\left(T\right)\right) \\ W_t &= TextEncoder_{teacher}\left(T\right) \\ Loss_t &= SmoothL1\left(w_s, w_t\right) \end{aligned} \tag{2}$$

where T is the input text and L is the fully connected layer.

Feature Alignment

This training stage aims to further align image and text features through contrastive learning with Chinese and English image-text pairs dataset. As shown in Figure 1b, following the LiT (Zhai et al. 2022) approach, we freeze the distilled image encoder at training time and only update the distilled text encoder's parameters. Alignment between the output projections of the image and text encoders is achieved using a contrastive loss, known as InfoNCE loss (Oord, Li, and Vinyals 2018), paralleling prior research methodologies. Specifically, given N image text pairs (x, y), where x is the image and y is the text description. Then the contrastive loss from image to text is:

$$\mathcal{L}_{\text{I-T}} = -\log \frac{\exp(\cos(f_{\text{s}}(x), w_{\text{s}}(y)))}{\sum_{y' \in Y} \exp(\cos(f_{\text{s}}(x), w_{\text{s}}(y')))} \tag{3}$$

where $f_s\left(x\right)$ is the image feature, $w_s\left(y\right)$ is the text feature, $cos\left(.,.\right)$ denotes the cosine similarity, and Y is the set of text descriptions. The contrastive loss from text to image is:

$$\mathcal{L}_{\text{T-I}} = -\log \frac{\exp(\cos(w_{\text{s}}(y), f_{\text{s}}(x)))}{\sum_{x' \in X} \exp(\text{sim}(w_{\text{s}}(y), f_{\text{s}}(x')))} \tag{4}$$

Table 3: This table presents a comparative analysis of the performance metrics for the baseline models, original CLIP, teacher model AltCLIP, and DC-CLIP method across eight Chinese datasets. The **best** and second-best results between our model and the baselines are prominently highlighted.

Method	Type	ImageNet	Caltech101	FGVCAircraft	StanfordCars	EuroSAT	Food101	Flowers102	OxfordPets
CLIP(224M)	Original	1.428	9.12	6.72	16.55	26.64	1.49	2.04	4.71
AltCLIP(3.22G)	Large-size	56.914	81.04	29.19	69.56	55.38	73.91	49.44	76.94
CNCLIP(294M)	Tiny-size	32.262	74.37	6.9	27.71	22.52	39.93	33.91	54.05
TaiyiCLIP(391M)		**40.50**	71.67	7.13	**34.75**	48.4	42.37	**48.11**	**56.25**
ChineseCLIP(726M)		17.394	52.33	2.91	11.82	28.64	19.81	8.60	10.46
DC-CLIP(301M)		38.142	**75.44**	**7.62**	13.90	**53.42**	**43.89**	16.57	15.67

Table 4: This table presents a performance comparison between DC-CLIP and TinyCLIP trained on different training datasets, where TinyCLIP* denotes the model trained on the English image-text pairs dataset organized in this paper.

Method	ImageNet	Caltech101	FGVCAircraft	StanfordCars	EuroSAT	Food101	Flowers102	OxfordPets
TinyCLIP	41.08	71.48	**6.87**	7.01	22.64	**57.05**	**57.90**	**45.22**
TinyCLIP*	16.32	56.25	2.46	1.77	19.70	11.31	15.63	40.37
DC-CLIP	**42.782**	**73.75**	6.24	**16.54**	**64.32**	46.91	26.43	17.23

where $f_s\left(x\right)$ is the image feature,$w_s\left(y\right)$ is the text feature, $cos\left(.,.\right)$ denotes the cosine similarity, and X is the set of images. The total loss of comparative learning is:

$$\mathcal{L}\wr\int\int_{\mathrm{CL}} = \frac{1}{2}\left(\mathcal{L}_{\text{T-I}} + \mathcal{L}_{\text{T-I}}\right) \tag{5}$$

By minimizing the loss of contrastive learning, the features between images and text are further aligned, thus improving the performance of the model.

Experiment

Experiment Setting

Training Datasets During the distillation of image encoder stage, we organized the training sets of Imagenet-1k (Deng et al. 2009), Caltech101 (Fei-Fei, Fergus, and Perona 2004), FGVC Aircraft (Maji et al. 2013), Stanford Cars (Krause et al. 2013), EuroSAT (Helber et al. 2019), Food-101 (Bossard, Guillaumin, and Van Gool 2014), UCF101 (Soomro, Zamir, and Shah 2012), and Oxford Flowers 102 (Nilsback and Zisserman 2008) as the dataset for training the image encoder. The distillation text encoder stage utilizes the translation2019zh Chinese-English translation dataset, which consists of approximately 5.2 million Chinese-English parallel corpora. The training set contains 5.16 million corpora, while the validation set contains 39,000 corpora. Both the training set and validation set are formatted in JSON. In the comparative learning stage, we organized 1.4 million Chinese image-text pairs from CLIP-Chinese and 1.5 million English image-text pairs from LAION 5B. The English image-text pairs dataset from LAION 5B (Schuhmann et al. 2022) is a randomly selected subset of data that has been filtered by the Laion Aesthetic v2 model with a threshold score exceeding 6.25 producing a small-scale, high-quality dataset of Chinese-English image-text pairs.

Test Datasets Our DC-CLIP was tested on eight public image classification datasets to showcase its effectiveness in zero-shot classification experiments. The datasets consist of two generic-objects datasets, ImageNet and Caltech101; five fine-grained classification datasets, OxfordPets (Parkhi et al. 2012), StanfordCars, Flowers102, Food101, and FGVCAircraft; and a satellite image dataset, EuroSAT. Details of the datasets are listed in Table 1. To evaluate the Chinese CLIP on the dataset, we first transform the datasets for Chinese models by translating labels and prompts into Chinese. Specifically, we translate the text descriptions of the labels and the templates of the manual prompts into Chinese. The labels in caltech101, such as "car" and "dog," are translated into Chinese manually. Special cases, like the labels in FGVC-Aircraft, can be challenging to translate or transliterate. We research the names on Google to discover the optimal Chinese name for each label. However, we cannot guarantee that we have the best Chinese translation. Additionally, Chinese pre-trained models may struggle to comprehend certain concepts, resulting in subpar performance in related tasks. We utilize template translations from the ELEVATER (Li et al. 2022b) toolkit for certain datasets and from the OpenAI CLIP templates for other datasets.

Baselines To evaluate our approach, we compared DC-CLIP against the original CLIP, AltCLIP, and other variants of comparable parameter scale on both Chinese and English datasets. Specifically, we analyzed three CLIP variants adapted to different linguistic contexts: Taiyi-CLIP, CN-CLIP, and ChineseCLIP. Taiyi-CLIP employs Chinese-roberta-wwm (Cui et al. 2021) as its text encoder and ViT-B/32 (Dosovitskiy et al. 2020) from CLIP as its image encoder. CN-CLIP utilizes the RBT3 architecture for its text encoder and the ResNet50 architecture for its image encoder. ChineseCLIP integrates a Bert architecture for text encoding, pairing it with ViT-B/32 from CLIP for image encoding. The memory footprint of DC-CLIP and these variant CLIP

Table 5: This table shows the comparative analysis of the performance metrics of the DC-CLIP method before and after the comparative learning on the eight datasets, with the best performing results highlighted.

Method	language	ImageNet	Caltech101	FGVCAircraft	StanfordCars	EuroSAT	Food101	Flowers102	OxfordPets
DC-CLIP-P	English	41.912	71.87	5.55	15.69	62.75	45.56	26.12	12.7
DC-CLIP		**42.782**	**73.75**	**6.24**	**16.54**	**64.32**	**46.91**	**26.43**	**17.23**
DC-CLIP-P	Chinese	37.666	73.59	6.24	10.81	49.76	43.57	16.43	11.04
DC-CLIP		**38.142**	**75.44**	**7.62**	**13.9**	**53.42**	**43.89**	**16.57**	**15.67**

Table 6: This table shows the results of the first stage distillation image encoder after distillation using different image encoders, with the best-performing result highlighted.

Method	ImageNet	Caltech101	FGVCAircraft	StanfordCars	EuroSAT	Food101	Flowers102	OxfordPets
ViT	40.438	55.375	1.41	2.08	22.08	0.59	6.04	66.45
ResNet50	**67.324**	**82.15**	**8.04**	**15.96**	**19.56**	**42.17**	**16.20**	**71.42**

models is approximately 500M. Meanwhile, to verify that DC-CLIP achieves good performance with only small-scale data, we compared the model replicated on our small-scale English image-text pairs dataset, referred to as TinyCLIP*, with the original TinyCLIP trained on a large-scale dataset. This comparison was conducted on the English datasets.

Parameter Settings In the stage of distilling the image encoder, we adopted the hyperparameter settings as outlined by Wei et al., setting the image size to 224, batch size to 128, epochs to 30, learning rate to 3e-4, and weight decay to 0.05. For the text encoder distillation stage, the batch size was set to 256, with epochs set to 10 and a learning rate of 8e-5. In the contrastive learning phase, we set a batch size of 256, processing the data once, with a learning rate of 2e-6. All models were trained using two NVIDIA RTX4090 GPUs.

Performance Evaluation

The results obtained by different methods on 8 English datasets and 8 Chinese datasets are listed in Table 2 and Table 3. From the results, compared with the baselines, DC-CLIP significantly outperforms the baseline model in accuracy across the majority of datasets in the English context (in 6 out of 8 cases), with only slight underperformance noted in the StanfordCars and OxfordPets datasets. But the performance is only slightly degraded, which may be caused by the lack of text about the StanfordCars and OxfordPets fine-grained categorical data in the English text data when distilling the text encoder. In the Chinese context, DC-CLIP demonstrates superior accuracy over CN-CLIP on five datasets (ImageNet, Caltech101, FGVC, EuroSAT, Food101), exceeds Taiyi-CLIP in four datasets (Caltech101, FGVCAircraft, EuroSAT, Food101), and comprehensively outperforms ChineseCLIP across all assessed datasets. A possible reason for the weak performance on the other evaluated datasets is that the Chinese dataset in the distillation text encoder stage is of poorer quality and contains less finely categorized text. In conclusion, the results show that our model holds competitive advantages over baseline models in the Chinese context and exhibits superior performance in the English context, particularly on the generic-objects dataset of ImageNet-1k. In conclusion, these results validate the effectiveness of our idea.

Furthermore, to validate the advantages of our approach in achieving competitive performance with only small-scale dataset and the effectiveness of the two-stage compression model strategy, we replicated the TinyCLIP (Wu et al. 2023) model, referred to as TinyCLIP*, on the English image-text pairs dataset we organized. We then compared its performance across eight English evaluation datasets with the original TinyCLIP model trained on a dataset of over a billion samples. As shown in Table 4, compared to TinyCLIP*, DC-CLIP exhibits lower performance only on the OxfordPets dataset, outperforming TinyCLIP* on the other datasets. In comparison to TinyCLIP, DC-CLIP demonstrates performance advantages on four datasets (ImageNet, Caltech101, StanfordCars, EuroSAT), showcasing a certain level of competitiveness. In conclusion, DC-CLIP holds the advantage of achieving good performance with only a small-scale dataset and further validates the effectiveness of the two-stage compression model strategy.

Ablation Studies

Impact of contrastive learning To assess the efficacy of the contrastive learning, we analyzed the performance of the DC-CLIP model both with and without the contrastive learning on 8 different test datasets. Table 5 illustrates that DC-CLIP performs better than DC-CLIP prior to contrastive learning on all test sets. This demonstrates the effectiveness of contrastive learning, which aligns image and text features, improves the model's ability to generalize across datasets, mitigates data bias, and enhances the model's robustness and generalization performance. At the same time, it also proves the importance of the contrastive learning strategy adopted by DC-CLIP.

Impact of different image encoder In the stage of distilling the image encoder, in order to explore the effect of choosing different image encoders as the student model on the distillation effect, we chose the vit architecture with 6-layer transformer and Resnet50 as the image encoder respectively. The distillation results are shown in Table 6, which

shows that under the same size of training data, the selection of Resnet50 as the student model is more effective. This indicates that Resnet50 can learn the effective knowledge of the teacher model better under a relatively small training dataset. Meanwhile, in order to enable the model to converge faster, we choose the pre-trained image encoder ResNet50 of CLIP as the initialized student model.

Conclusion

In this study, we developed and evaluated a comprehensive model distillation approach, DC-CLIP, that emphasizes bilingual adaptability, innovatively uses output feature mapping as a distillation target, and combines contrastive learning to improve model performance on a limited training dataset. Our findings reveal that this methodology not only facilitates the adaptation of models to both English and Chinese contexts but also significantly enriches the knowledge transfer process between teacher and student models. By focusing on output feature maps rather than traditional logits or simplified feature vectors, we were able to capture a more nuanced representation of information. The inclusion of contrastive learning further augmented this process, enabling a more precise alignment of features across different modalities and languages.

However, there are some limitations to our approach. While our DC-CLIP achieves state-of-the-art performance in the English environment, it only attains competitive performance compared to the baseline model in the Chinese environment. Therefore, our future work could delve deeper into optimizing feature mapping distillation techniques and assess the suitability of our approach in more resource-constrained computing environments. Additionally, exploring the extension of this approach to other languages and domains is warranted. The potential for enhancing model understanding and facilitating interaction across diverse cultures and languages is vast, and our work serves as a foundational step for further explorations in this direction.

References

Bianchi, F.; Attanasio, G.; Pisoni, R.; Terragni, S.; Sarti, G.; and Lakshmi, S. 2021. Contrastive language-image pre-training for the italian language. *arXiv preprint arXiv:2108.08688*.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, 446–461. Springer.

Buciluǎ, C.; Caruana, R.; and Niculescu-Mizil, A. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 535–541.

Carlsson, F.; Eisen, P.; Rekathati, F.; and Sahlgren, M. 2022. Cross-lingual and multilingual clip. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 6848–6854.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3558–3568.

Chen, W.-C.; Chang, C.-C.; and Lee, C.-R. 2019. Knowledge distillation with feature maps for image classification. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, 200–215. Springer.

Chen, Z.; Liu, G.; Zhang, B.-W.; Ye, F.; Yang, Q.; and Wu, L. 2022. Altclip: Altering the language encoder in clip for extended language capabilities. *arXiv preprint arXiv:2211.06679*.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Cui, Y.; Che, W.; Liu, T.; Qin, B.; and Yang, Z. 2021. Pre-Training with Whole Word Masking for Chinese BERT.

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

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. 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*, 178–178. IEEE.

Fu, Z.; Yang, J.; Bai, C.; Chen, X.; Zhang, C.; Zhang, Y.; and Wang, D. 2020. Astraea: Deploy ai services at the edge in elegant ways. In *2020 IEEE International Conference on Edge Computing (EDGE)*, 49–53. IEEE.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6): 1789–1819.

Gu, J.; Meng, X.; Lu, G.; Hou, L.; Minzhe, N.; Liang, X.; Yao, L.; Huang, R.; Zhang, W.; Jiang, X.; et al. 2022. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35: 26418–26431.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, 770–778.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226.

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

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

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19113–19122.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 554–561.

Li, B.; Weinberger, K. Q.; Belongie, S.; Koltun, V.; and Ranftl, R. 2022a. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*.

Li, C.; Liu, H.; Li, L.; Zhang, P.; Aneja, J.; Yang, J.; Jin, P.; Hu, H.; Liu, Z.; Lee, Y. J.; et al. 2022b. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems*, 35: 9287–9301.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022c. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10965–10975.

Lin, J.; Zhu, L.; Chen, W.-M.; Wang, W.-C.; and Han, S. 2023. Tiny Machine Learning: Progress and Futures [Feature]. *IEEE Circuits and Systems Magazine*, 23(3): 8–34.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Mukherjee, A.; Ukil, A.; Dey, S.; and Kulkarni, G. 2022. TinyML Techniques for running Machine Learning models on Edge Devices. In *Proceedings of the Second International Conference on AI-ML Systems*, 1–2.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, 722–729. IEEE.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, 3498–3505. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

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

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15638–15650.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Srivastava, R. K.; Greff, K.; and Schmidhuber, J. 2015. Training very deep networks. *Advances in neural information processing systems*, 28.

Thomee, B.; Shamma, D. A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; and Li, L.-J. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2): 64–73.

Wang, W.; Bao, H.; Huang, S.; Dong, L.; and Wei, F. 2020. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. *arXiv preprint arXiv:2012.15828*.

Wei, Y.; Hu, H.; Xie, Z.; Zhang, Z.; Cao, Y.; Bao, J.; Chen, D.; and Guo, B. 2022. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*.

Wu, K.; Peng, H.; Zhou, Z.; Xiao, B.; Liu, M.; Yuan, L.; Xuan, H.; Valenzuela, M.; Chen, X. S.; Wang, X.; et al. 2023. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21970–21980.

Xu, J.; De Mello, S.; Liu, S.; Byeon, W.; Breuel, T.; Kautz, J.; and Wang, X. 2022. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18134–18144.

Xu, X.; Li, M.; Tao, C.; Shen, T.; Cheng, R.; Li, J.; Xu, C.; Tao, D.; and Zhou, T. 2024. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

Yang, A.; Pan, J.; Lin, J.; Men, R.; Zhang, Y.; Zhou, J.; and Zhou, C. 2022a. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*.

Yang, J.; Li, C.; Zhang, P.; Xiao, B.; Liu, C.; Yuan, L.; and Gao, J. 2022b. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19163–19173.

Yuan, L.; Chen, D.; Chen, Y.-L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Zhai, X.; Wang, X.; Mustafa, B.; Steiner, A.; Keysers, D.; Kolesnikov, A.; and Beyer, L. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18123–18133.

Zhang, H.; Koh, J. Y.; Baldridge, J.; Lee, H.; and Yang, Y. 2021a. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 833–842.

Zhang, H.; Zhang, P.; Hu, X.; Chen, Y.-C.; Li, L.; Dai, X.; Wang, L.; Yuan, L.; Hwang, J.-N.; and Gao, J. 2022a. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35: 36067–36080.

Zhang, J.; Gan, R.; Wang, J.; Zhang, Y.; Zhang, L.; Yang, P.; Gao, X.; Wu, Z.; Dong, X.; He, J.; Zhuo, J.; Yang, Q.; Huang, Y.; Li, X.; Wu, Y.; Lu, J.; Zhu, X.; Chen, W.; Han, T.; Pan, K.; Wang, R.; Wang, H.; Wu, X.; Zeng, Z.; and Chen, C. 2022b. Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence. *CoRR*, abs/2209.02970.

Zhang, R.; Fang, R.; Zhang, W.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2021b. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.