2024 International Joint Conference on Neural Networks (IJCNN) | 979-8-3503-5931-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCNN60899.2024.10650723

A Lightweight and Effective Multi-View Knowledge Distillation Framework for Text-Image Retrieval

Yuxiang Song[1], Yuxuan Zheng[1,2], Shangqing Zhao[1], Shu Liu[1], Xinlin Zhuang[1], Zhaoguang Long[1],
Changzhi Sun[1], Aimin Zhou[1], and Man Lan[1,3,*]
[1] *School of Computer Science and Technology, East China Normal University, Shanghai, China*
[2] *China Jiangxi Radio and TV Station, Nanchang, China*
[3] *Shanghai Institute of AI for Education, East China Normal University, Shanghai, China*
{yxsong, yuxuanzheng, sqzhao, shuliu, xinlinzhuang, 51265901014}@stu.ecnu.edu.cn,
czsun.cs@gmail.com, {amzhou, mlan}@cs.ecnu.edu.cn

***Abstract*—Large-scale dual-stream Vision-Language Pre-training (VLP) models provide an efficient solution for text-image retrieval tasks. Despite this, their performance often falls short of the most current single-stream models, primarily due to limited fine-grained text-image interactions. Recent trends indicate a union of these two types of networks. Some methods adopt a *retrieve and rerank* strategy, their performance improvements largely hinge on the single-stream encoder during inference. Other approaches utilize knowledge distillation to strengthen either the single-stream encoder or the dual-stream encoder, surpassing their previous capabilities. However, existing distillation techniques typically focus on a single knowledge type, neglecting the richer insights available in the teacher model. To bridge this gap, we introduce a Lightweight and Effective Multi-View Knowledge Distillation approach, named LEMKD, for text-image retrieval. This method effectively utilizes response-based, feature-based and relation-based knowledge, transferring the knowledge from the single-stream encoder to the dual-stream encoder. Our approach is executed on the widely used MS-COCO and Flickr30K datasets. Results demonstrate that LEMKD not only matches the exceptional performance of the most advanced single-stream models but also excels in dual-stream encoder performance amidst the recent integration of single-stream and dual-stream models.**

***Index Terms*—text-image retrieval, knowledge distillation, multi-modal**

I. Introduction

Text-Image Retrieval (TIR) represents a critical task in cross-modal learning, involving the retrieval of pertinent samples from one modality by utilizing another. This process typically embraces two subtasks: Image-to-Text (i2t) and Text-to-Image (t2i) retrieval. With the rapid developments within the field of deep learning, alongside the proliferation of data interaction, TIR has evolved to become a research focus within cross-modal learning. Consequently, it has found practical applications in areas such as search engines, recommendation systems, and question-answering systems [1]. Recent advancements in Vision-Language Pre-training (VLP) [2]–[6] have significantly propelled text-image retrieval tasks forward. Current approaches divide into dual-stream and single-stream models, each with distinct architectural characteristics as depicted in Fig. 1.

Dual-stream models [6], [27], [30] allow for the pre-computation of representations in both modalities, enabling their persistent reuse. Despite this efficiency, their performance often lags behind single-stream models due to limited text-image interaction depth. In contrast, single-stream models [11], [31], [32] excel in capturing intricate details between modalities, thereby enhancing cross-modal alignment and retrieval effectiveness. However, they face challenges in retrieval latency during the inference phase, as each text-image pair requires online processing through fusion modules. The need for dual-stream models that are both lightweight and highly effective is acute in practical applications. Bridging the gap between the benefits of dual-stream and single-stream models has thus emerged as a crucial research focus.

Some studies adopt a two-stage retrieve-and-rerank methodology. For instance, LightningDot [33], initially employs a dual-stream encoder to identify the $top-M$ candidates (where M is significantly smaller than the database size) and subsequently reranks these pairs using a slower, more powerful single-stream encoder. [34] extends this approach, combining dual-stream and single-stream encoders into a shared-weight model with a parameter-efficient joint fine-tuning strategy. However, these methods heavily depend on the single-stream encoder's performance, imposing substantial computational demands, while the dual-stream encoder simply serves as a foundational stepping stone in the structure.

As previously mentioned, our attention should be paid to improving the efficient dual-stream encoder by utilizing the powerful single-stream encoder. Knowledge Distillation (KD), introduced by [12], is a promising approach to address this issue. The single-stream encoder, acting as a teacher, generates binary classification scores, while the dual-stream encoder, acting as a student, produces cosine similarity scores. Both types of scores can be used in the distillation process. Some researchers have followed this strategy to transfer knowledge from the single-stream encoder to the dual-stream encoder. For instance, LoopITR [13], based on the same model architecture as ALBEF [3], employs the original classification task-specific

We appreciate the support from National Natural Science Foundation of China with the Main Research Project on Machine Behavior and HumanMachine Collaborated Decision Making Methodology (72192820 & 72192824), Pudong New Area Science & Technology Development Fund (PKX2021-R05), Science and Technology Commission of Shanghai Municipality (22DZ2229004), Shanghai Trusted Industry Internet Software Collaborative Innovation Center and East China Normal University International Conference Grant Programme. (* Corresponding author: Man Lan)

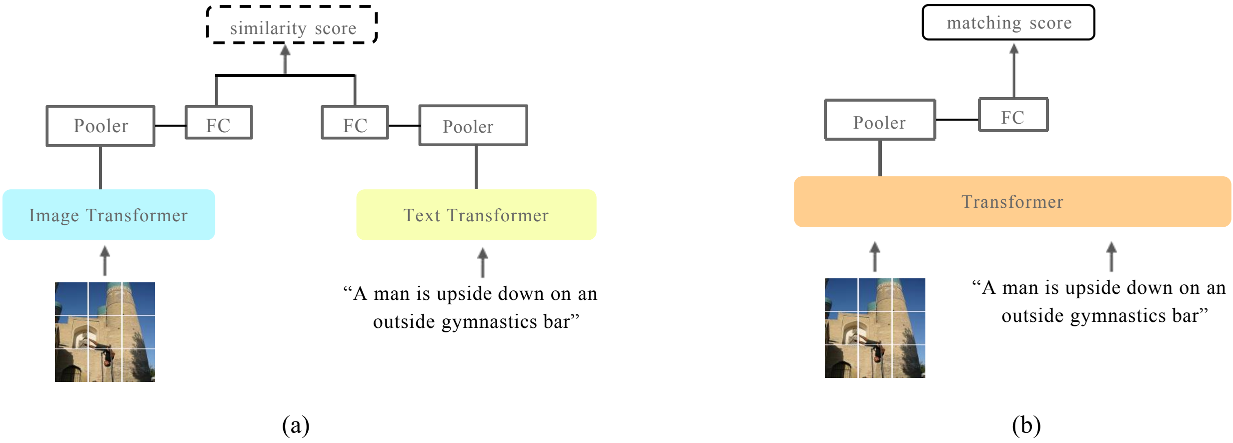

Fig. 1. Illustration of the comparison of the dual-stream and single-stream model pipelines. (a) Dual-stream model. (b) Single-stream model.

distillation approach [12] to text-image retrieval in an online manner. It transfers the distribution of predictions from single-stream models to dual-stream models via a cross-entropy loss. However, this method primarily leverages response-based knowledge [12], [18], ignoring the importance of deeper feature-based knowledge [19], [20] and relation-based knowledge [21]–[23]. For deep multi-modal models, merely learning the distribution is too superficial and fails to fully utilize the other knowledge hidden within the teacher model. Therefore, we design a multi-view distillation method to utilize as much knowledge as possible, improving the performance of the dual-stream encoder by a large margin.

The principal contributions of this paper are as follows:

- We introduce a model that incorporates response-based, feature-based, and relation-based knowledge for distillation in the TIR task, an aspect that has been previously overlooked by other methods.
- Our multi-view knowledge distillation method (LEMKD) significantly outperforms most recent single-stream encoder models and maintains state-of-the-art performance in dual-stream encoder models, as demonstrated on the Flickr30K and MS-COCO datasets.
- We provide the evidence that our model effectively bridges the performance gap between the distilled dual-stream encoder and the original joint training of single and dual-stream encoders.

II. Related Work

A. Text-Image Retrieval

TIR is a fundamental branch of information retrieval, garnering considerable research attention. The task has made tremendous progress in the past two years due to the large-scale text-image datasets [14], [15] and Vision-Language Pre-training (VLP) models [2]–[4]. These VLP models resort to dual-stream encoder or single-stream encoder due to their interaction ways.

Early dual-stream encoder models [38], [39] employing a two-branch deep neural network with multiple layers, achieved limited performance. More recent studies (e.g., CLIP [6], ALIGN [27], FILIP [28], Florence [29], BEIT-3 [30]), have improved their performance by leveraging millions of text-image network data. These models typically employ two distinct encoders to encode image and text modalities, obtaining a joint global embedding space in a decoupled manner. They then compute their similarity score via a dot product. While this architecture allows for independent and dynamic selection of encoders, enhancing computational efficiency in large-scale retrieval tasks, it falls short in modeling fine-grained semantic alignment between image regions and text phrases.

On the other hand, single-stream encoder models [11], [31], [32] typically utilize a single cross encoder (e.g., additional co-attention mechanism [7]) to process the fusion sequence of image and text modalities in an interconnected way. They then compute matching scores via a Fully Connected (FC) layer. In earlier works, a direct concatenation was employed after mapping the extracted image and text features to the same dimension, yet the effectiveness of fusion was suboptimal. Recent works have adopted a strategy where the output from the text encoder serves as the query, while the image output is integrated as the key and value. These elements are then fed into the cross encoder [2], [3], [13]. The single-stream model permits patch/token-level interaction, facilitating fine-grained cross-modal alignment. However, the model is relatively cumbersome and less efficient, as it requires inputting all information for inference, making it impractical in time-sensitive real-world text-image retrieval scenarios.

B. Knowledge Distillation

Knowledge Distillation (KD), introduced by [12], is a technology initially proposed to transfer knowledge with soft targets from a teacher to a student, leading to competitive or superior performance. It has found widespread application in various domains, including computer vision [16], natural language processing [17], and multi-modal fields [2]. Knowledge manifests in diverse forms, encompassing soft target knowledge of the output layer [12], feature maps of the hidden intermediate layers [19], activation boundaries of hidden neurons [20] and instance relational feature knowledge [21]–[23]. Structurally, knowledge can be classified into three categories [18]: response-based knowledge [12], [24], feature-based knowledge [19], [20] and relation-based knowledge [21]–[23].

Recent works have applied KD to text-image retrieval tasks. [25] distill knowledge from a large dual-encoder model to a smaller dual-encoder model through a fully connected knowledge interaction graph learning method, designed to derive a series of smaller, faster, and more effective lightweight models. [3] propose Momentum Distillation (MoD), a self-distillation method generating pseudo targets to enhance the single-stream encoder's representation ability from noisy supervision signals, with the goal of bolstering the single-stream encoder's effectiveness. [13] construct score distributions for distillation between the cross-encoder and dual encoder, combining them in the same network for joint learning, aiming to enhance the dual-stream encoder's performance. Nevertheless, these methods, irrespective of their purpose, overlook or underutilize feature-based and relation-based knowledge in the distillation process. Unlike previous works concentrating on one type of KD in text-image retrieval, our objective is to leverage multiple types of knowledge to enhance the dual-stream student model's performance, enabling it to encapsulate as rich structural knowledge as the single-stream teacher model.

III. Methodology

In this section, we outline our comprehensive approach to enhance the performance of dual-stream encoders in TIR tasks through Multi-view knowledge distillation. Section III-A and III-B present the model architecture and pre-training objectives of our proposed method LEMKD. It is designed to capitalize on three distinct types of knowledge: response-based, feature-based, and relation-based during the distillation process. In Sections III-C, III-D and III-E, we delve into the three distinct types of knowledge and detail how each knowledge type is distilled and integrated into the dual-stream encoder, addressing their unique contributions and roles in enhancing model performance. Section III-F introduces the total loss computation of our method.

To provide a metaphorical description for our methodology, we draw on an ancient Chinese proverb: *”Give a man a fish, and you feed him for a day; teach a man to fish, and you feed him for a lifetime”*. In this metaphor, response-based knowledge is likened to providing the fish, offering immediate but limited nourishment. Feature-based knowledge is analogous to teaching the act and behavior of fishing, equipping the model with tools for independent learning. Relation-based knowledge imparts the methods and techniques of fishing, ensuring a deeper, more systemic understanding and application. These three types of knowledge are crucial for improving model performance.

A. Model Architecture

We show an overview of our implementation of the proposed LEMKD architecture in Figure 2. LEMKD mainly consists of two parts. The first part is the dual-stream encoder and the other part is the single-stream encoder.

1) Dual-stream Encoder Module: This module comprises two encoders: the image encoder and the text encoder. Each Transformer layer, whether in the image encoder or the text encoder, consists primarily of layers with self-attention mechanisms and feed-forward networks (FFN).

Given an image and a caption, the image is transformed into a sequence of patch embeddings $\mathbf{I} = \{\boldsymbol{i}_{cls}, \boldsymbol{i}_1, \ldots, \boldsymbol{i}_N\}$, and the text is transformed into a sequence of text embeddings $\mathbf{T} = \{\boldsymbol{t}_{cls}, \boldsymbol{t}_1, \ldots, \boldsymbol{t}_N\}$, where $\boldsymbol{i}_{cls}$ and $\boldsymbol{t}_{cls}$ represent the *[CLS]* token embedding. The image-to-text and text-to-image similarity scores are calculated through a dot product:

$$\begin{aligned} s^{i2t} &= g_i\left(\boldsymbol{i}_{cls}\right)^T g_t\left(\boldsymbol{t}_{cls}\right), \\ s^{t2i} &= g_t\left(\boldsymbol{t}_{cls}\right)^T g_i\left(\boldsymbol{i}_{cls}\right) \end{aligned} \tag{1}$$

Here, g_i and g_t represent the image and text projection heads followed by a normalization operation, which transforms the image *[CLS]* embedding and text *[CLS]* embedding to a low-dimensional (256-d) space.

2) Single-stream Encoder Module: This module consists of a cross-encoder transformer, where each layer primarily includes a self-attention layer, a cross-attention layer, and a feed-forward neural network (FFN). Multiple layers of cross-attention are used to fuse the image patch embeddings $\mathbf{I} = \{\boldsymbol{i}_{cls}, \boldsymbol{i}_1, \ldots, \boldsymbol{i}_N\}$ and text token embeddings $\mathbf{T} = \{\boldsymbol{t}_{cls}, \boldsymbol{t}_1, \ldots, \boldsymbol{t}_N\}$. Text features are fused with image features through multi-modal fusion operations at each layer. Subsequently, after acquiring the joint representation from the last layer $\mathbf{J} = \{\boldsymbol{j}_{cls}, \boldsymbol{j}_1, \ldots, \boldsymbol{j}_N\}$ and extracting the *[CLS]* output embedding j_{cls}, we employed a linear head followed by a softmax operation to derive the matching scores:

$$m = \text{Softmax}\left(h_{\text{head}}\left(\boldsymbol{j}_{cls}\right)\right) \tag{2}$$

Here, h_{head} represents the linear head, and the softmax operation transforms the combination of the image and text *[CLS]* embedding into a two-class probability.

B. Pre-training Objectives

Due to the substantial cost of pre-training and remarkable performance, we used a pre-trained VLP model BLIP [2], which has two understanding-based objectives and one generation-based objective. In text-image retrieval tasks, we only use two primary pre-training objectives in our model fine-tuning process:

(i) Image-Text Contrastive Loss (ITC). Employed for training the dual-stream encoder, the learning objective is to ensure that the correctly positive image-text pairs are closer while pushing the negative samples farther away. It has been proven effective in many previous works [3], [6].

(ii) Image-Text Matching Loss (ITM). It is used for training the single-stream encoder. It is dedicated to learning the multimodal representation of image-text pairs and capturing the fine-grained fusion representation between the two modalities. It is a binary classification task, where the model uses a linear layer to predict whether an image and text pair is a match or not. Moreover, ITC serves to assist ITM in mining more negative

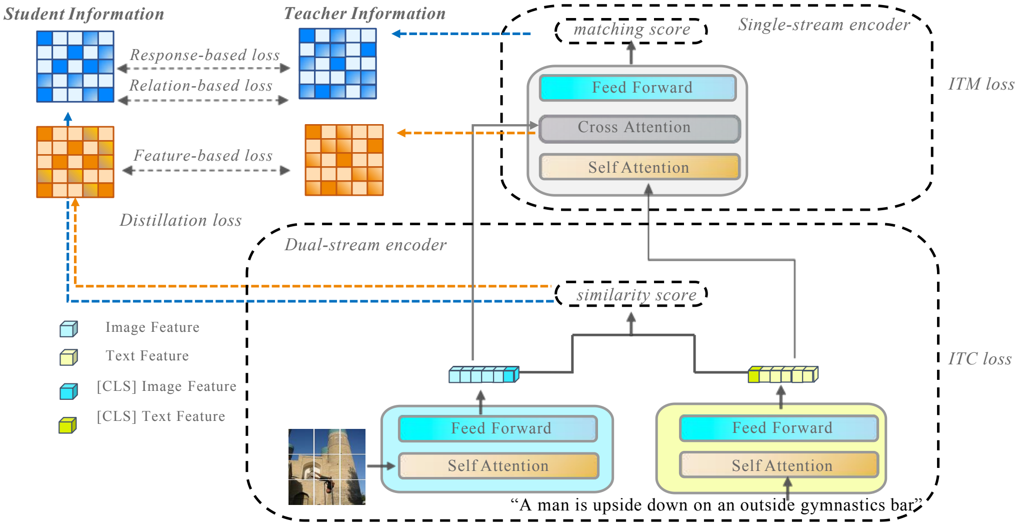

Fig. 2. Details of our method: the architecture consists of two modules, including a dual-stream student model and a single-stream teacher model. The dual-stream model includes an image encoder and a text encoder, while the single-stream model includes a cross-encoder. All encoders are implemented as transformers. In our LEMKD framework, the training objective of the student model is the Image-Text Contrast (ITC) loss, and the training objective of the teacher model is the Image-Text Matching (ITM) loss. The teacher is distilled through response-based knowledge, feature-based knowledge, and relation-based knowledge. For simplicity, the hard sample mining process is not shown in this figure.

samples. Negative pairs with high contrastive similarity in a batch are selected for the loss calculation.

C. Response-based Knowledge Distillation

Response-based knowledge refers to the neuron units noted in the final output layer of the teacher model. This simple yet highly effective form of knowledge distillation was proposed in [12] and its capabilities often match or even surpass those of the teacher model. Many previous distillation-related studies, including [3], [13], [25], primarily adopted this type of knowledge in the context of multi-modal tasks. However, they selectively opted for certain hard negative samples when distilling from matching scores, thereby disregarding similarity scores derived from the dual-stream model that envelops all text as well as image samples. Therefore, we propose that single-stream models should take all negative samples into account for a well-rounded understanding and approach. Furthermore, we prefer to directly measure the deviation between the student's and teacher's values. Hence, we have applied the Mean Squared Error (MSE) in our experiments for distillation from matching scores to similarity scores. It has been proven to yield more accurate results compared to previous methods.

After obtaining the similarity scores s^{i2t}, s^{t2i} and matching scores m. We initially perform a reshape operation on the output from the single-stream encoder and then concatenate each batch of image and text embeddings to get the representation of $m^{Response}$. As illustrated in Fig. 2, the response-based knowledge distillation loss is calculated as follows:

$$\mathcal{L}_{Response} = \frac{1}{N}\sum_{n=1}^{N}(m_n^{Response} - s_n^{i2t})^2 + \frac{1}{N}\sum_{n=1}^{N}(m_n^{Response} - s_n^{t2i})^2 \quad (3)$$

Here, N is the number of selected image and text samples, $m^{Response}$ is the response-based knowledge from the single-stream teacher model, and s^{i2t} and s^{t2i} are the image-to-text and text-to-image outputs from the dual-stream student model. Note that our objective is to maintain the parameters of the teacher model unchanged, so we halt the gradients from back-propagating when calculating $m^{Response}$.

D. Relation-based Knowledge Distillation

Relation-based knowledge, as its name suggests, highlights various levels and data samples per input's intricate relationships. Notably, it offers a relationship map to facilitate the student model in acquiring the relational knowledge ingrained in the teacher model.

Solely deploying output-to-output response-based distillation may not be optimal for fostering a comprehensive understanding of the knowledge to be transferred. As such, we also introduce relation-based knowledge for distillation to import the relational knowledge that exists between the teacher and student models. In our approach, we directly leverage the scores derived from the output layer, as outlined in Section III-C, and denote these scores as $m^{Relation}$. Additionally, we integrate two elemental relationships into our methodology for knowledge distillation: distance-wise and angle-wise relations.

We utilize the Euclidean distance metric to measure the correlation between the student and teacher model for distance-wise relation. Meanwhile, the angle-wise relation is determined by calculating the two models' inner product. To expedite the knowledge transfer process, we employ the Smooth L1 loss [36]. Here's the conduction of the necessary computations:

$$\begin{aligned}
L_{dist} &= \frac{1}{N}\sum_{(i,j)\in\mathbb{K}^2} SL1\left(\frac{1}{\psi(m)}\left\|m_i^{Relation} - m_j^{Relation}\right\|_2, \frac{1}{\psi(s)}\left\|s_i^{i2t} - s_j^{t2i}\right\|_2\right), \\
L_{angle} &= \frac{1}{N}\sum_{(i,j)\in\mathbb{N}^2} SL1\left(\frac{1}{\psi(m)}\left\|m_i^{Relation} - m_j^{Relation}\right\|_2, \frac{1}{\psi(s)}\left\|s_i^{i2t} - s_j^{t2i}\right\|_2\right), \quad (4) \\
\psi(x) &= \frac{1}{|\mathbb{N}^2|}\sum_{(i,j)\in\mathbb{N}^2}\|x_i - x_j\|_2
\end{aligned}$$

where $\mathbb{N}^2 = \{(i,j) \mid i \neq j, 1 \leqslant i,j \leqslant N\}$, m^{Relation} is the relation-based knowledge from the single-stream teacher model, $\psi(\cdot)$ acts as normalization factor for distance and angle, and *SL1* refers to Smooth L1 loss.

As demonstrated in "(4)", the method allotted more weight to the relationship between teacher and student samples in terms of relational knowledge. The expression for loss is:

$$\mathcal{L}_{\mathit{Relation}} = w_{dist} \cdot L_{dist} + w_{angle} \cdot L_{angle} \tag{5}$$

Here, w_{dist} and w_{angle} denote the weights of the distance relation and the angle relation respectively, and can be adjusted during the training process.

E. Feature-based Knowledge Distillation

Feature-based knowledge typically refers to the information or features captured in intermediary layers, such as attention maps created by the teacher model. These maps encapsulate high-dimensional and in-depth information about the input data. Recent studies have demonstrated the application of transferring cross-attention maps in several works within both NLP and CV fields. The potential application of these maps also extends to multi-modal tasks.

Attention mechanisms serve as a valuable component of neural networks due to their ability to mediate computation between elements. The attention matrix is computed as follows:

$$\text{attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

Where Q and K represent the query and key in the attention layer of the transformer block, while V indicates this layer's value. d_k serves as a scaling factor for the key's dimension.

However, due to the inherent structural differences between single-stream and dual-stream models, transferring feature-based knowledge on a layer-by-layer basis poses a significant challenge. After extensive trials, we decided to employ an average aggregate strategy and use the final transformer layer data from the teacher model for knowledge distillation. The advantage of this approach is that the attention maps of the final transformer block provide the student model with comprehensive and detailed knowledge. Simultaneously, the complexity of the task is significantly reduced by eliminating the need to identify the optimal layer mapping.

Firstly, the attention map was extracted from the final layer of the single-stream model. Subsequently, we calculated the average of the attention scores—averaging across both attention heads and tokens. The computation of the average attention scores is detailed as follows:

$$avg = \frac{1}{H \cdot |S|} \sum_{h=1}^{H} \sum_{s=1}^{|S|} (attention_{N,h,s}) \tag{7}$$

Where H and S represent the number of attention heads and sequence length, respectively. N is the number of layers in the teacher model, while $\mathbf{attention}_N$ refers to the attention distribution of the teacher model's N^{th} layer.

Following this, the average attention scores were used to determine the weighted token vectors $wscore$ for each text and image sample. The classification scores were then determined using the classification head h_{head} , and we further applied a softmax operation to derive the predicted probability scores m^{Feature}. As illustrated in Fig. 2, by capturing the attention scores from the teacher model's intermediate layer, we express the feature-based knowledge distillation loss thusly:

$$\begin{aligned} &\mathcal{L}_{\mathit{Feature}} = \frac{1}{N}\sum_{n=1}^{N}(m_n^{\mathit{Feature}} - s_n^{i2t})^2 + \frac{1}{N}\sum_{n=1}^{N}(m_n^{\mathit{Feature}} - s_n^{t2i})^2, \\ &m^{\mathit{Feature}} = \mathit{Softmax}\left(h_{\mathit{head}}(wscore)\right), \\ &wscore = \sum_{s=1}^{S} J_s \cdot avg_s \end{aligned} \tag{8}$$

where $wscore$ denotes the weighted token vectors, while m^{Feature} signifies feature-based knowledge from the single-stream teacher model.

F. Overall loss function

In our training process, the total loss used for training the student model is as follows:

$$\mathcal{L} = \mathcal{L}_{ITC} + \mathcal{L}_{ITM} + \lambda_1\mathcal{L}_{Response} + \lambda_2\mathcal{L}_{Feature} + \lambda_3\mathcal{L}_{Relation} \tag{9}$$

Where $\mathcal{L}_{ITC}$ and $\mathcal{L}_{ITM}$ are the ITC and ITM loss used in [2], $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters used to balance the various distillation loss in the same scale.

IV. Experiment

A. Baselines

In assessing the effectiveness of our approach, we set it alongside some of the most advanced models in recent years for comparison. Our method employs three forms of knowledge distillation from the single-stream to dual-stream encoder, with the dual-stream encoder's performance being our primary focus. Hence, we initially selected a range of strong baselines based on the single-stream encoder, including UNITER [31], VILLA [37], OSCAR [11], and VinVL [32], as their works are primarily oriented towards cross-encoder performance. Subsequently, due to LEMKD assimilates knowledge from the single-stream model, we compared our method with competitive baselines that employ a strategy of integrating both types of encoder to attain notable performance in the dual-stream encoder, such as LightningDot [34], RerankSmart [33], and LoopITR [13]. Lastly, to emphasize the performance gap and speed improvement between the distilled dual-stream encoder and the original joint training of both encoders, we conducted an experiment comparing our method with LoopITR [13], which employs just one type of knowledge distillation (response-based) in text-image retrieval.

B. Datasets and Metrics

We executed experiments on the MS-COCO and Flickr30K benchmarks to manifest the effectiveness of our approach.

(1) **MS-COCO**[1], a large-scale image-text dataset, comprises 123,287 images, each accompanied by five annotations. We

[1] https://cocodataset.org/

TABLE I
EXPERIMENTAL RESULTS FINETUNED ON MS-COCO AND FLICKR30K DATASET. THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLD, AND THE SECOND BEST IS UNDERLINED.

Model	#Trainable	MS-COCO(5K test set)								Flickr30K (1K test set)							
		Text → Image				Image → Text				Text → Image				Image → Text			
	Params	R_{mean}	R@1	R@5	R@10	R_{mean}	R@1	R@5	R@10	R_{mean}	R@1	R@5	R@10	R_{mean}	R@1	R@5	R@10
single-stream encoder																	
UNITER [31]	303M	72.0	50.3	78.5	87.2	81.6	64.4	87.4	93.1	87.0	72.5	92.4	96.1	93.9	85.9	97.1	98.8
VILLA [37]	303M	-	-	-	-	-	-	-	-	87.8	74.7	92.9	95.8	94.6	86.6	97.9	99.2
OSCAR [11]	345M	74.4	54.0	80.8	88.5	85.5	70.0	91.1	95.5	-	-	-	-	-	-	-	-
VinVL [32]	345M	**77.1**	**58.1**	**83.2**	**90.1**	87.8	74.6	92.6	96.3	-	-	-	-	-	-	-	-
dual-stream encoder																	
LightningDot [33] + UNITER	306M	72.2	50.3	78.7	87.5	81.9	64.6	87.6	93.5	87.3	72.6	93.1	96.1	94.3	86.5	97.5	98.9
RerankSmart [34] + OSCAR	345M	75.0	54.7	81.3	88.9	85.7	70.8	91.0	95.2	88.7	76.4	93.6	96.2	95.4	89.4	97.7	99.0
LoopITR [13]	233M	72.8	51.7	79.2	87.5	84.5	67.6	90.5	95.4	89.7	77.2	94.3	97.6	95.9	89.6	98.6	99.5
LEMKD	223M	75.8	57.0	81.6	88.9	**88.4**	**75.5**	**92.9**	**96.8**	**92.0**	**82.2**	**95.8**	**97.9**	**98.1**	**94.7**	**99.8**	**99.9**

TABLE II
COMPARISON WITH PREVIOUS DISTILLATION WORK ON THE FULL FLICKR 1K TEST SPLIT DATASET. THE COLUMN $\triangle R$ SHOWS THE GAP BETWEEN THE DISTILLED DUAL-STREAM ENCODER AND THE JOINT TRAINING OF SINGLE AND DUAL-STREAM ENCODERS. TIME REPRESENTS THE INFERENCE TIME COSTS (SECONDS).

Model	Trained Encoder	Flickr30K (1K test set)						Total	$\triangle R$	Time
		Text→ Image			Image→ Text					
		R@1	R@5	R@10	R@1	R@5	R@10	R_{mean}		
LoopITR [13]	Dual(distilled)	77.2	94.3	97.6	89.6	98.6	99.5	92.80	2.43	**11**
ALBEF [3]	Single+Dual	82.8	96.7	98.4	94.3	99.4	99.8	95.23		117
LEMKD	Dual(distilled)	82.2	95.8	97.9	94.7	99.8	99.9	95.05	**1.78**	12
BLIP [2]	Single+Dual	87.3	97.6	98.9	97.3	99.9	100.0	96.83		255

resorted to the Karpathy split method for allocation: setting aside 5,000 images for testing, 5,000 for validation, and utilizing the remaining 113,000 images for training.

(2) **Flickr30K**[2], encompasses 31,000 images and 158,915 annotations, generally denoted by five annotations per image. Following the split method as per [26], we used 1,000 images for testing, 1,000 images for validation, and the remaining 29,000 for training.

In keeping with previous studies [2], [3], [13], we adopted R@k (k = 1, 5, 10), R_{mean}, and R_{sum} as our evaluative metrics, to provide a fair and comprehensive comparison with existing cross-modal retrieval models. Specifically, R@k denotes the proportion of matching samples within the top-k retrieval results; R_{mean} signifies their average; R_{sum} represents the summation of R@1 for both t2i and i2t retrieval, which facilitates a balanced comparison across both retrieval tasks.

C. Implementation Details

Our method employs a pre-trained model as our backbone, initialized with the weights from the BLIP [2] model trained on a 129M image dataset. The image encoder involves a 12-layer Visual Transformer (ViT-B/16), and the text encoder uses a 12-layer text transformer, namely BERT$_{\text{base}}$. Additionally, the single-stream encoder comprises a 12-layer cross-encoder. We run LEMKD in the same configuration and hardware environment and the training period spanned 6 epochs on the MS-COCO and Flickr30K datasets. Table I lists the number of parameters of all baselines to show a setting with approximate trainable parameters.

We employed the AdamW optimizer [35] with a weight decay of 0.05 and utilized a cosine learning rate scheduling strategy with a peak rate of 1×10^{-5} for model parameter updates. A batch size of 16 was set and the image resolution was established at 384 × 384. Considering our method's three types of knowledge distillation loss, we ultimately adopt the hyper-parameters $\{\lambda_1 = 6\times10^{-1}, \lambda_2 = 3\times10^{-1}, \lambda_3 = 1\times10^{-1}\}$ for Flickr30K and $\{\lambda_1 = 6\times10^{-1}, \lambda_2 = 5\times10^{-1}, \lambda_3 = 1\times10^{-4}\}$ for MS-COCO. Additionally, we set w_{dist} to 25 and w_{angle} to 50 as constants for relation-based knowledge distillation. These values were diagnosed by evaluating the initial losses of various types of knowledge and ensuring that all losses were on the same order of magnitude, however, their fit for a new dataset may require some adjustments to optimize the performance of the model. Experiments were carried out on RTX 3090. All other parameters were initialized with the default values from [2], and the model was implemented via Transformers [7].

D. Experimental Results and Analysis

Table I provides the main results of two distinct models for t2i and i2t retrieval performances on the MS-COCO and Flickr30k datasets. Table II depicts the performance gap between the distilled dual-stream encoder and the original joint training of the two corresponding encoders. Through analysis of these data, we can reach the following conclusions.

Firstly, as shown in Table I, LEMKD matches the performance of advanced single-stream models in most cases, with a minor underperformance in the t2i retrieval on the MS-COCO dataset. This suggests that the multi-view distillation provides a more comprehensive view of the single-stream encoder, thereby enabling the dual-stream encoder's performance to be comparable to that of single-stream encoders.

Secondly, as indicated in Table I, our model makes tangible advances in dual-stream encoder performances compared with previous works utilizing a cooperative *retrieve and rerank* strategy [33], [34] and the distillation study transferring knowledge from single-stream to dual-stream encoder [13]. Specifically, our approach achieves improvements (average in t2i and i2t task) of 5.05%, 1.75%, 3.45% (MS-COCO) and 4.25%, 3.00%, 2.25% (Flickr30k) in R_{mean}. These results highlight the impact of three typical forms of knowledge incorporation in enhancing dual-stream encoder retrieval performance.

[2] https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset

Thirdly, as presented in Table II, LEMKD clearly outperforms the previous distillation work LoopITR [13]. It not only performs a more competitive performance against the previous joint training model with a slight 1.78% gap in $\triangle R$, but also strikes a balance between efficiency and effectiveness during the inference period, leading to a 21x speed improvement. The improvements suggest that each type of knowledge does produce results and offers unique information, promoting a comprehensive and nuanced data understanding. Moreover, This equilibrium is credited to the strategic use of knowledge from the single-stream encoder and the inherent efficient feature of the dual-stream encoder.

E. Ablation Study

In this subsection, we conduct several ablation studies on the Flickr30k dataset to further explore the effect of each distillation component and the different experiment settings on the results in the proposed LEMKD framework.

*1) **Effect of Each Distillation Component:*** Table III shows the results of the ablation study of each distillation component in LEMKD. We intend to explore whether each knowledge contributes to helping text-image retrieval. As shown in Table III, we can observe that each knowledge component improves performance, especially the response-based and feature-based knowledge. Moreover, Combining the above three knowledge types leads to the best results, bringing about a 3.1 R_{sum} increase compared to the original model. Comparing the last two columns, it proves that these three types of knowledge from the single-stream model have their unique parts and overlapping parts, but we have utilized them together effectively.

TABLE III
ABLATION STUDY FOR EACH PART OF THE DISTILLATION COMPONENT.

Model	dual	dual				
	w/o distillation	distillation				
responsed-based feature-based relation-based	- - -	✓ - -	- ✓ -	- - ✓	✓ ✓ -	✓ ✓ ✓
R_{sum}	173.8	176.2	176.3	175.5	176.4	**176.9**

*2) **Analysis of Various Methods in Response-based KD and Layers in Relation-based KD:*** We conducted an analysis to investigate the specific contributions of various methods for Response-based KD and various layers for Relation-based KD. Table IV provides a comparison of these two types of knowledge, employing different distillation methods or layers for a closer analysis of their impact.

For the response-based knowledge distillation, we utilized the Mean Squared Error (MSE) and Cross-Entropy (CE) loss functions as distillation methods. It is noted in the results that MSE achieves a higher R_{sum} than CE. It might be attributed to MSE's characteristic of minimizing squared discrepancies between the student's and teacher's predictions. This aligns well with the continuous features observed in the embedding space, as reflected in the first two rows of Table IV.

On the other hand, for the relation-based knowledge distillation, we explored the implications of distilling knowledge from varying layers, contrasting between the output and intermediate layers. The results show that the use of the output layer in distillation is more effective, reflected by a superior R_{sum} value. It suggests that the output layer which measures the correlation between the student and teacher model may be more refined, as corroborated in the last two rows of Table IV.

TABLE IV
ABLATION STUDY OF RESPONSE-BASED KNOWLEDGE DISTILLATION WITH DIFFERENT METHODS AND RELATION-BASED KNOWLEDGE DISTILLATION WITH DIFFERENT LAYERS.

Model	method/layer	R_{sum}
response-based	MSE CE	**176.2** 174.9
relation-based	output layer intermediate layer	**175.5** 173.7

*3) **Exploration of Different Transformer Layers and Aggregation Methods in Feature-based KD:*** We explore the influence of employing different transformer layers and aggregation methods in the feature-based knowledge distillation process on our framework and record the results in Table V. It outlines the results corresponding to variations in the transformer layers and aggregation methods used. The results in the first row underscore the advantages of focusing on the last transformer layer and using average aggregation for distillation. It suggests that the setting of the last transformer layer, typically encapsulated highly refined feature representations, combined with average aggregation, which provides a balanced and comprehensive overview of features, enhances the effectiveness of the feature-based distillation process.

TABLE V
ABLATION STUDY OF FEATURE-BASED KNOWLEDGE DISTILLATION WITH DIFFERENT LAYERS AND POLYMERIZATION WAYS.

Model	layer	polymerization	R_{sum}
feature-based	last layer last three layers last layer last three layers	average average max max	**176.3** 175.4 175.9 175.6

V. CONCLUSION

This paper introduces LEMKD, an innovative method designed to transfer multi-view knowledge from a single-stream encoder to a dual-stream encoder for text-image retrieval tasks. By exploiting the inherent response-based, feature-based, and relation-based knowledge from the single-stream encoder, our method enhances the dual-stream encoder's performance. Experimental results on the MS-COCO and Flickr30K datasets confirm that our approach surpasses most existing single-stream encoder models. Impressively, compared with previous methods employing both single-stream and dual-stream networks, our model exhibits exceptional dual-stream encoder performance. Crucially, our method provides a more balanced approach to efficiency and effectiveness compared to previous distillation methods. We will later explore the applicability of our approach to text-video retrieval tasks.

REFERENCES

[1] W. Chen, L. Yao, and Q. Jin, "Rethinking benchmarks for cross-modal image-text retrieval," *arXiv preprint arXiv:2304.10824*, 2023.

[2] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.

[3] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[4] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *International Conference on Machine Learning*. PMLR, 2022, pp. 23 318–23 340.

[5] X. Wang, L. Li, Z. Li, X. Wang, X. Zhu, C. Wang, J. Huang, and Y. Xiao, "Agree: Aligning cross-modal entities for image-text retrieval upon vision-language pre-trained models," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 2023, pp. 456–464.

[6] 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.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[8] H. Bao, W. Wang, L. Dong, Q. Liu, O. K. Mohammed, K. Aggarwal, S. Som, S. Piao, and F. Wei, "Vlmo: Unified vision-language pre-training with mixture-of-modality-experts," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 897–32 912, 2022.

[9] Y. Tay, D. Bahri, D. Metzler, D.-C. Juan, Z. Zhao, and C. Zheng, "Synthesizer: Rethinking self-attention for transformer models," in *International conference on machine learning*. PMLR, 2021, pp. 10 183–10 192.

[10] T. Yu, R. Khalitov, L. Cheng, and Z. Yang, "Paramixer: Parameterizing mixing links in sparse factors works better than dot-product self-attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 691–700.

[11] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 2020, pp. 121–137.

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

[13] J. Lei, X. Chen, N. Zhang, M. Wang, M. Bansal, T. L. Berg, and L. Yu, "Loopitr: Combining dual and cross encoder architectures for image-text retrieval," *arXiv preprint arXiv:2203.05465*, 2022.

[14] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 104–12 113.

[15] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 278–25 294, 2022.

[16] X. Dai, Z. Jiang, Z. Wu, Y. Bao, Z. Wang, S. Liu, and E. Zhou, "General instance distillation for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7842–7851.

[17] J. FitzGerald, S. Ananthakrishnan, K. Arkoudas, D. Bernardi, A. Bhagia, C. Delli Bovi, J. Cao, R. Chada, A. Chauhan, L. Chen *et al.*, "Alexa teacher model: Pretraining and distilling multi-billion-parameter encoders for natural language understanding systems," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 2893–2902.

[18] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021.

[19] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[20] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge transfer via distillation of activation boundaries formed by hidden neurons," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3779–3787.

[21] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3967–3976.

[22] Y. Liu, J. Cao, B. Li, C. Yuan, W. Hu, Y. Li, and Y. Duan, "Knowledge distillation via instance relationship graph," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7096–7104.

[23] X. Li, J. Wu, H. Fang, Y. Liao, F. Wang, and C. Qian, "Local correlation consistency for knowledge distillation," in *European Conference on Computer Vision*. Springer, 2020, pp. 18–33.

[24] S. W. Kim and H.-E. Kim, "Transferring knowledge to smaller network with class-distance loss," 2017.

[25] J. Wang, C. Wang, X. Wang, J. Huang, and L. Jin, "Conaclip: Exploring distillation of fully-connected knowledge interaction graph for lightweight text-image retrieval," *arXiv preprint arXiv:2305.17652*, 2023.

[26] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[27] 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.

[28] L. Yao, R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, and C. Xu, "Filip: Fine-grained interactive language-image pre-training," *arXiv preprint arXiv:2111.07783*, 2021.

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

[30] W. Wang, H. Bao, L. Dong, J. Bjorck, Z. Peng, Q. Liu, K. Aggarwal, O. K. Mohammed, S. Singhal, S. Som *et al.*, "Image as a foreign language: Beit pretraining for all vision and vision-language tasks," *arXiv preprint arXiv:2208.10442*, 2022.

[31] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *European conference on computer vision*. Springer, 2020, pp. 104–120.

[32] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Making visual representations matter in vision-language models," *arXiv preprint arXiv:2101.00529*, vol. 1, no. 6, p. 8, 2021.

[33] S. Sun, Y.-C. Chen, L. Li, S. Wang, Y. Fang, and J. Liu, "Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 982–997.

[34] G. Geigle, J. Pfeiffer, N. Reimers, I. Vulić, and I. Gurevych, "Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 503–521, 2022.

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

[36] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[37] Z. Gan, Y.-C. Chen, L. Li, C. Zhu, Y. Cheng, and J. Liu, "Large-scale adversarial training for vision-and-language representation learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6616–6628, 2020.

[38] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5005–5013.

[39] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3441–3450.