CL2CM: Improving Cross-Lingual Cross-Modal Retrieval via Cross-Lingual Knowledge Transfer

Yabing Wang[123*], **Fan Wang**[3], **Jianfeng Dong**[15†], **Hao Luo**[34†]

[1] Zhejiang Gongshang University
[2] Xi'an Jiaotong University
[3] DAMO Academy, Alibaba Group
[4] Hupan Lab, Zhejiang Province
[5] Zhejiang Key Lab of E-Commerce

Abstract

Cross-lingual cross-modal retrieval has garnered increasing attention recently, which aims to achieve the alignment between vision and target language (V-T) without using any annotated V-T data pairs. Current methods employ machine translation (MT) to construct pseudo-parallel data pairs, which are then used to learn a multi-lingual and multi-modal embedding space that aligns visual and target-language representations. However, the large heterogeneous gap between vision and text, along with the noise present in target language translations, poses significant challenges in effectively aligning their representations. To address these challenges, we propose a general framework, Cross-Lingual to Cross-Modal (CL2CM), which improves the alignment between vision and target language using cross-lingual transfer. This approach allows us to fully leverage the merits of multilingual pre-trained models (*e.g.,* mBERT) and the benefits of the same modality structure, *i.e.,* smaller gap, to provide reliable and comprehensive semantic correspondence (knowledge) for the cross-modal network. We evaluate our proposed approach on two multilingual image-text datasets, Multi30K and MSCOCO, and one video-text dataset, VATEX. The results clearly demonstrate the effectiveness of our proposed method and its high potential for large-scale retrieval.

Introduction

As the internet continues to expand globally, users from diverse linguistic backgrounds increasingly access multimodal content online, such as images and videos. Nevertheless, accurately retrieving such content poses considerable challenges, particularly for non-English speakers, as the majority of human-annotated datasets are only available in English. To address this issue, cross-lingual cross-modal retrieval (CCR) has emerged as a crucial area of research, aiming to develop models applicable to non-English languages without incurring substantial manual annotation costs.

The primary challenge for CCR lies in effectively achieving cross-lingual transfer and establishing reliable correlations between V-T. A prevalent approach is to construct pseudo-parallel data pairs using MT and explicitly learn the correspondence between V-T. As shown in Figure 1(a), the straightforward baseline method directly attempts to align visual and target-language representations, similar to conventional cross-modal alignment methods (Kim et al. 2023; Liu et al. 2022; Dong et al. 2022b). But it suffers from the noisy translation issue. Previous efforts (Wang et al. 2022; Jain et al. 2021; Zhang, Hu, and Jin 2022) attempt to alleviate this issue by incorporating the matching between visual and source-language representations, and treat the source language as the focal point, as shown in Figure 1(b). However, these methods are limited to instance-level matching [1] in the visual-text domain, as they match image-text pairs relying on global feature vectors. On the other hand, some endeavors (Ni et al. 2021; Zhou et al. 2021; Zeng et al. 2022) utilize cross-modal fusion modules to model fine-grained interactions between image regions and target-language words, as depicted in Figure 1(c). Although these methods demonstrate promising performance, they entail huge computational costs, as all possible query-candidate pairs need to be fed into the fusion modules during inference. In light of the above discussions, we argue that conventional approaches still face *a dilemma between achieving reliable alignment of V-T and computational efficiency.*

In this paper, we propose a novel solution: **improving the alignment between vision and target language using cross-lingual knowledge transfer**. We note that existing multi-lingual pre-trained models, such as mBERT (Devlin et al. 2018), have demonstrated remarkable performance in cross-lingual alignment but have not been fully utilized in CCR tasks, where they are typically only used as multilingual text backbones. Furthermore, cross-lingual sentences exhibit more analogous modality structures, with a smaller gap compared to visual elements. These properties provide a unique opportunity to comprehensively explore the knowledge between different languages and transfer it to align V-T.

In specific, we propose a novel framework named Cross-Lingual to Cross-Modal (CL2CM) that aims to address two main challenges mentioned above: 1) noisy translations and 2) reliable and comprehensive correspondence learning. As shown in Figure 1(d), CL2CM consists of a cross-

*Work done during an internship at DAMO Academy.
†Corresponding authors

[1] Instance level matching refers to the matching between individual image and text instances, and the alignment loss is applied to the final embedding layer, similar to the CLIP model.

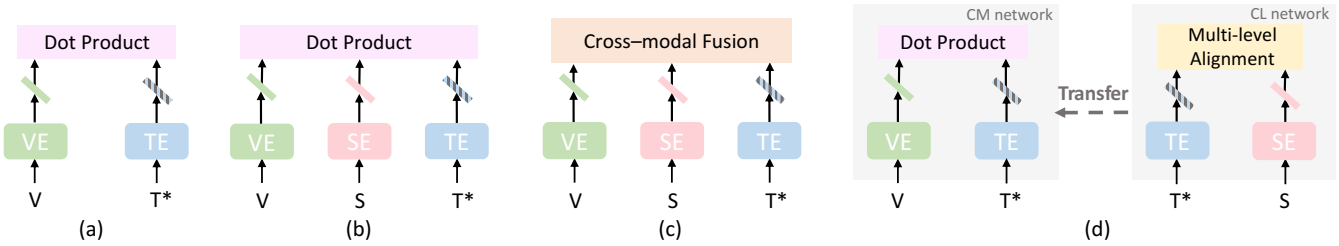


Figure 1: Comparison of different alignment methods for vision and target language: (a) Baseline method. (b) Universal CCR method, which projects the three data types into a common space by learning the alignment between them. (c) Single-stream method. (d) Proposed CL2CM method. Among these methods, (a), (b), and (d) are all dual-stream methods. * indicates data with noise. "VE", SE", and "TE" represent the image encoder, source-language encoder, and target-language encoder, respectively. "V", S", and "T" represent the visual item, source-language caption and target-language caption, respectively.

lingual (CL) network with multi-level alignment and a cross-modal (CM) network for vision-target language matching. The multi-level alignment in the CL network is designed to tackle the noisy translation issue, including both sentence-level and self-supervised word-level alignment. For self-supervised word-level alignment, we generate pseudo-labels in a self-supervised manner to align cross-lingual representations. This approach mitigates misalignment caused by noisy words, and enhances the discrimination ability of the representations at the word level. To address the second challenge, we employ the knowledge distillation to transfer the CL knowledge to the CM network. The CL network, utilizing multi-level alignment, can provide more comprehensive and reliable correspondence, which can also alleviate the effect of noisy translations (*i.e.,* noise correspondence (Huang et al. 2021b)). *Moreover, during inference, only the cross-modal network is utilized in CL2CM, bringing no additional cost compared to other dual-stream methods.*

To sum up, our contributions can be summarized as follows: (i) We propose a novel framework called CL2CM that improves the alignment between vision and target language using cross-lingual knowledge transfer. To the best of our knowledge, we are the first to explore cross-lingual transfer for CCR. (ii) We introduce a multi-level alignment strategy to explore the cross-lingual knowledge and alleviate the effect of noisy translations. (iii) Extensively experiments on three CCR benchmarks, *i.e.,* Multi30K, MSCOCO, and VATEX, demonstrating the effectiveness of our proposed method and its high potential for large-scale retrieval.

Related Work

Cross-lingual Cross-modal Retrieval

Cross-lingual cross-modal retrieval has been garnering increased attention amongst researchers, as it enables the acquisition of images or videos utilizing a non-English query, without relying on human-labeled vision-target language data (Li et al. 2023; Wu et al. 2023). This method mitigates the constraints of conventional cross-modal retrieval tasks (Dong et al. 2022a; Zheng et al. 2023) centered on English, and offers a highly efficient and cost-effective solution for target-language based retrieval, greatly reducing the need for human-labeled data. In terms of the model architecture, there are mainly two broad directions to conduct CCR. The first line (Zhou et al. 2021; Ni et al. 2021; Zeng et al. 2022) utilizes a single-stream model that incorporates cross-lingual and cross-modal fusion modules to model both image regions and multilingual text word representations in a unified semantic space, capturing the fine-grained relationship between them. For example, Zhou *et al.,* (Zhou et al. 2021) use a cross-lingual cross-modal encoder to model the interaction between vision and multiple languages, employing the masked modeling loss. However, this approach has limitations in real-world scenarios, such as high computational costs and reliance on object detectors for image regions, rendering it less practical for large-scale CCR tasks.

The other line (Wang et al. 2022; Huang et al. 2021a; Jain et al. 2021; Zhang, Hu, and Jin 2022; Wang et al. 2023) involves two-stream models, where each stream is dedicated to modeling the vision or language inputs. For example, Jain et al. (Jain et al. 2021) learn V-T alignment by combining image-text matching and text-text matching tasks, using scalable dual encoder models trained with contrastive losses. Although these approaches are efficient, they only consider the instance-level alignment but ignore fine-grained correspondence between visual and textual information, resulting in suboptimal performance. In our work, we dedicate the two-stream structure, and leverage CL knowledge to improve the alignment between vision and the target language.

Knowledge Distillation

Knowledge Distillation is defined as a learning manner that extracts "dark knowledge" from a teacher network to guide the learning of a student network, encouraging the student model to imitate the teacher's behavior. Hinton et al. (Hinton, Vinyals, and Dean 2015) first proposed minimizing the KL-divergence of category distributions for classification tasks by matching logits produced by two models. Except for final layer logits distillation, some methods (Romero et al. 2015; Sun et al. 2020) use knowledge distillation to distill compact feature representations from the teacher network, such as, Sun *et al.,* (Sun et al. 2020) proposed MobileBERT to compress and accelerate the popular BERT model.

Moreover, in the multi-modal field, knowledge distillation can be applied to compress visual-and-language (VL) models (Fang et al. 2021; Rao et al. 2021; Dong et al. 2023). For instance, Fang *et al.,* (Fang et al. 2021) employ knowl-

edge distillation to effectively compress a transformer-based large VL model into a small VL model. In our work, we use knowledge distillation to transfer the CL knowledge to improve the alignment between vision and target language.

Methods

Preliminary

We consider a dataset $\mathcal{D} = (V, S)$ consisting of annotated paired images (or videos) V with source-language captions S. However, obtaining human-labeled V-T pairs during training may not always be feasible. To address this issue, we follow the approach in (Wang et al. 2022) and create an extended dataset $\hat{\mathcal{D}} = (V, S, T)$, where T indicates the translated target-language captions corresponding to S. Note that during inference, we use only human-input target-language sentences as queries for retrieval.

Figure 2 presents an overview of our approach. In what follows, we first depict the CL network with multi-level alignment, followed by the description of the CM network, and finally, introduce the CL knowledge transfer. For ease of description, we use image-text retrieval as an example to describe our framework.

Cross-lingual Network

Given a source-language caption s consisting of M words and its corresponding target-language caption t consisting of N words, we utilize a pre-trained multilingual encoder $G(\cdot)$ to extract the contextualized word embeddings $h^s = \{\overline{h^s}; h^s_1, ..., h^s_M\}$ and $h^t = \{\overline{h^t}; h^t_1, ..., h^t_N\}$, respectively. Here, $\overline{h^s}$ and $\overline{h^t}$ are [CLS] tokens, and we take them as sentence-level representations. It can be defined as:

$$h^s = G^s(s), h^t = G^t(t) \tag{1}$$

To capture the complex correspondence and align the presentations between cross-lingual data pairs accurately, we develop a multi-level alignment strategy that establishes correlations between captions in different languages at various levels of granularity.

Instance-level alignment. Given a batch of B pseudo-parallel sentence pairs, we first compute the instance-level similarity scores $\mathcal{S}^{cl}$ using sentence-level representations. Then, we utilize a symmetric InfoNCE loss over the similarity matrix to optimize the text encoder and learn the discriminative cross-lingual features. It can be formulated as:

$$\begin{aligned}\mathcal{L}^{cl}_{instance} = -\tfrac{1}{2} \times \tfrac{1}{B} \textstyle\sum_{i=1}^{B} [log \frac{exp(\mathcal{S}^{ins}_{i,i})}{\sum_{j=1}^{B} exp(\mathcal{S}^{ins}_{j,i})} \\ + log \frac{exp(\mathcal{S}^{ins}_{i,i})}{\sum_{j=1}^{B} exp(\mathcal{S}^{ins}_{i,j})}]\end{aligned} \tag{2}$$

where $\mathcal{S}^{ins}_{i,j} = \frac{\overline{h^s}_i{}^T \overline{h^t}_j}{||\overline{h^s}_i|| \cdot ||\overline{h^t}_j||}$, and $\overline{h^s}_i, \overline{h^t}_j \in \mathbb{R}^d$ represent the i-th source-language and j-th target-language sentence representations, respectively. The objective is to learn more distinctive cross-lingual sentence representations.

Self-supervised word-level alignment. Although instance-level alignment can promote cross-lingual matching and learn discriminative sentence features, it may cause the text encoder to only focus on abstract semantic information and lose its original word-level discriminative power. However, the word-level label is inaccessible, and while some word alignment tools are available, they may not be suitable for all languages, particularly for low-resource languages. To overcome this issue, we propose a self-supervised word alignment method that does not require additional word alignment tools to align cross-lingual captions at the word level. Specifically, we define the objective as computing an alignment matrix A between source and target language words, and then maximizing the probability distribution P of each aligned word pair. This optimization problem can be formulated as:

$$\min_A \sum_{m,n} -A_{m,n} log P(s_m, t_n) \tag{3}$$

To solve this equation, the Optimal Transport (OT) theory can be incorporated into our word alignment approach. OT has been applied in domain adaptation (Xu et al. 2020; Chang et al. 2022) to align the representations in the source and target domains with associated theoretical guarantees. It encourages a global mapping to mine domain statistics property for discovering intrinsic differences among clean and noisy sample pairs. We treat the source and target sentences as two probability distributions and employ OT to find the optimal alignment between them.

This can be formally defined as:

$$\begin{aligned}&\min_A \textstyle\sum_{m,n} A_{m,n} C(s_m, t_n) \\ &\text{s.t. } A\mathbb{1}_M = \tfrac{1}{N}\mathbb{1}_N,\ A^T\mathbb{1}_N = \tfrac{1}{M}\mathbb{1}_M\end{aligned} \tag{4}$$

where $\mathbb{1}_D$ represents a D-dimensional vector whose elements are all 1. The cost matrix $C(s_m, t_n) \in \mathbb{R}^{m \times n}$ represents the cost of aligning the m-th word in S with the n-th word in T, which can be formulated as:

$$C(s_m, t_n) = \frac{h^s_m{}^T h^t_n}{||h^s_m|| \cdot ||h^t_n||} \tag{5}$$

To approximate OT efficiently, we add an entropic regularizer $E(A)$ to the optimization problem, as shown below:

$$\begin{aligned}&\min_A \textstyle\sum_{m,n} A_{m,n} C(s_m, t_n) + \lambda E(A) \\ &\text{s.t. } A\mathbb{1}_M = \tfrac{1}{n}\mathbb{1}_N,\ A^T\mathbb{1}_N = \tfrac{1}{M}\mathbb{1}_M\end{aligned} \tag{6}$$

where $E(A) = \mu A log A$. The Eq. (6) has a unique solution A^* such that:

$$\begin{aligned}A^* &= diag(\mu) K diag(v) \\ K_{m,n} &= e^{C(s_m, t_n)/\mu}\end{aligned} \tag{7}$$

where $\mu \in \mathbb{R}^M_+$, $v \in \mathbb{R}^N_+$, $K \in \mathbb{R}^{N \times M}_+$, which solved with Sinkhorn's fixed point iteration:

$$u^{(t+1)} = \frac{1_N}{K v^{(t)}},\ v^{(t+1)} = \frac{1_M}{K^T u^{(t+1)}} \tag{8}$$

With the solved stochastic matrix A^*, we can produce the alignment labels $\widetilde{A^*}$ by applying a threshold γ and a function $h(,)$ to A^*:

$$\begin{aligned}\widetilde{A^*}_{m,n} &= \frac{h(A^*_{m,n}, \gamma) \times A^*_{m,n}}{\sum_{n=1}^{N} h(A^*_{m,n}, \gamma)} \\ h(A^*_{m,n}, \gamma) &= 1 - \tfrac{1}{2}(1 + sgn(A^*_{m,n} - \gamma))\end{aligned} \tag{9}$$

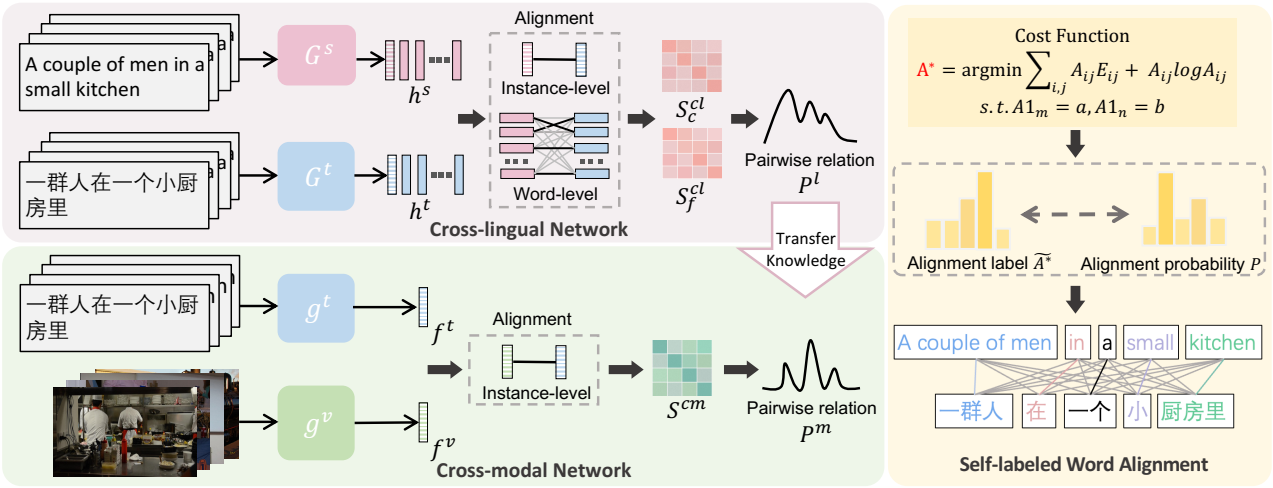


Figure 2: Overview of the proposed CL2CM framework. It consists of a cross-lingual (CL) with multi-level alignment and a cross-modal (CM) network with instance-level alignment. We aim to improve the alignment quality between vision and target-language using cross-lingual knowledge transfer.

where γ is a threshold of alignment, which we set to the mean value of A*, and $sgn(x) = 1$ when $x > 0$, and equals -1 otherwise. The training objective is to maximize the alignment probabilities between aligned words:

$$\mathcal{L}_{word} = \sum_{m,n} -\widetilde{A^*}_{m,n} log P(s_m, t_n)$$
$$P(s_m, t_n) = \frac{exp(C(s_m, t_n))}{\sum_{n=1}^{n} exp(C(s_m, t_n))} \quad (10)$$

The objective function described above is designed to minimize the distance between the best-matched word pairs while increasing the distance between poorly matched pairs. This helps alleviate the problem of misalignment caused by noisy words to some extent, as shown in Figure 3. Finally, we can align words in different languages and identify the best matching result from a candidate word set. To train the cross-lingual network, we define the objective function as $\mathcal{L}^{cl} = \mathcal{L}^{cl}_{instance} + \mathcal{L}_{word}$. Besides, other OT variants may exist, but this does not affect the starting point of this paper.

Cross-modal Network

The cross-modal network comprises an image encoder $g^v(\cdot)$ and a target-language encoder $g^t(\cdot)$, which can produce an image feature vector and a target-language caption feature vector, respectively. We denote them as f^v and $f^t \in \mathbb{R}^d$, respectively. To enhance the target-language representation, we share $g^t(\cdot)$ with $G^t(\cdot)$ in the CL network.

$$f^v = g^v(V)$$
$$f^t = g^t(T) \quad (11)$$

The goal of the cross-modal network is to accurately learn the alignment between visual and target-language representations. To achieve this, we define the objective function as:

$$\mathcal{L}^{cm} = \alpha \cdot \mathcal{L}^{cm}_{instance} + (1 - \alpha) \cdot \mathcal{L}_{ckt} \quad (12)$$

where α is a weight hyper-parameter, $\mathcal{L}^{cm}_{instance}$ is the instance-level alignment objective, which aligns image and target-language features at the instance level using the global vectors, and $\mathcal{L}_{ckt}$ represents the cross-lingual knowledge transfer loss which will be introduced in next Section. The $\mathcal{L}^{cm}_{instance}$ can be computed as:

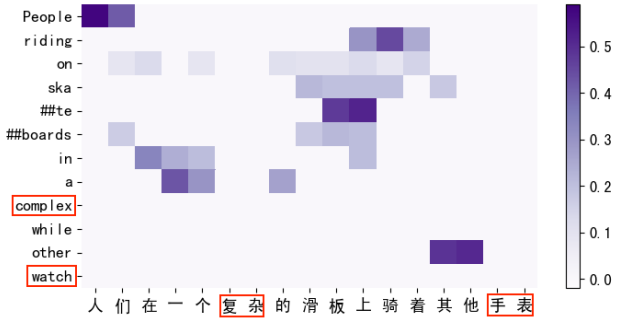


Figure 3: Visualization of the generated pseudo-label in self-supervised word-level alignment. The red box represents the incorrect translated word of the corresponding source-language word.

$$\mathcal{L}^{cm}_{instance} = \frac{1}{2} \times \frac{1}{B} \sum_{i=1}^{B} [log \frac{exp(\mathcal{S}^{cm}_{i,i})}{\sum_{j=1}^{B} exp(\mathcal{S}^{cm}_{j,i})}$$
$$+ log \frac{exp(\mathcal{S}^{cm}_{i,i})}{\sum_{j=1}^{B} exp(\mathcal{S}^{cm}_{i,j})}] \quad (13)$$

where $\mathcal{S}^{cm}_{i,j} = \frac{f_i^{v\,T} f_j^t}{||f_i^v|| \cdot ||f_j^t||}$ represents the instance-level similarity of i-th image and j-th target language caption.

The final training objective can be formulated as $\mathcal{L} = \mathcal{L}^{cm} + \mathcal{L}^{cl}$. It's worth noting that we only use the CM network to calculate the instance-level similarity matrix $\mathcal{S}^{cm}$ for image-target language retrieval during inference.

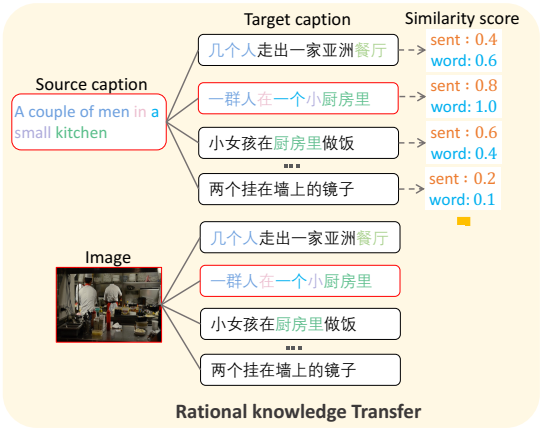


Figure 4: An illustration of the CL knowledge transfer.

Cross-lingual Knowledge Transfer

The significant heterogeneity gap between images and target languages may pose challenges in establishing their correlations. However, the reliable semantic relations are crucial for cross-modal retrieval, as the essence of retrieval is ranking. Considering the CL network uses a multi-lingual pre-trained model and training with multi-level alignment, enabling it to model more reliable and comprehensive correspondence. Therefore, we introduce Relational Knowledge Distillation (Park et al. 2019) to transfer the relational knowledge learned by the CL network to the CM network to provide the extra discriminative information, as shown in Figure 4. Specifically, we propose a multi-level cross-lingual knowledge transfer that involves a voting ensemble of two-level structural relational similarities (i.e., sentence and word levels). A high similarity score at both levels indicates a strong correlation between the sample pairs. The CL knowledge can be formulated as follows:

$$\begin{aligned} \mathcal{S}^{cl}(s_i, t_j) &= \lambda \cdot \mathcal{S}^{cl}_{sent}(s_i, t_j) + (1-\lambda) \cdot \mathcal{S}^{cl}_{word}(s_i, t_j) \\ \mathcal{S}^{cl}_{word}(s_i, t_j) &= \frac{1}{M} \sum_{m=1}^{M} \max_{0 \le n \le N} \frac{{h^s_m}^\top h^t_n}{||h^s_m|| \cdot ||h^t_n||} \end{aligned} \tag{14}$$

where λ is a weight hyper-parameter, $\mathcal{S}^{cl}_{sent}$ denotes the sentence-level similarity (similar to $\mathcal{S}^{ins}$) and $\mathcal{S}^{cl}_{word}$ is calculated from the word-level cross-lingual representations that can capture the subtle semantic difference between data pairs. By combining the two-level similarities, CL knowledge accurately depicts mutual relations beyond what each individual level can achieve alone. The CL knowledge transfer loss is then formulated as:

$$\begin{aligned} p^{cl} &= softmax(\mathcal{S}^{cl}/\tau) \\ p^{cm} &= softmax(\mathcal{S}^{cm}/\tau) \\ \mathcal{L}_{rkt} &= \frac{1}{B} \sum_{i=1}^{B} \sum KL(p^{cm}_i \,||\, p^{cl}_i) \end{aligned} \tag{15}$$

where KL denotes the Kullback-Leiber divergence, and τ is a temperature parameter. By transferring the CL knowledge to the CM network, we can improve the quality of alignment between visual and target-language representations, ultimately enhancing the retrieval performance.

Experiment

Experimental Settings

Datasets. We conduct experiments on two public multilingual image-text retrieval datasets (Multi30K and MSCOCO), as well as a video-text retrieval dataset (VATEX). Notably, we only use the annotated vision-source language data pair in the training process, while using the annotated vision-target language data pairs during inference.

- **Multi30K** (Elliott et al. 2016): This dataset consists of 31,000 images and is a multi-lingual version of Flickr30K (Young et al. 2014). It involves four languages, *i.e.,* English(en), German(de), French(fr), and Czech(cs). We adopt a similar data partition as (Young et al. 2014).
- **MSCOCO** (Chen et al. 2015): This dataset consists of 123,287 images, and each image has 5 captions. We translate the training set from English into Chinese(zh) and Japanese(ja) by resorting to MT, and using the test sets from the (Li et al. 2019) and (Yoshikawa, Shigeto, and Takeuchi 2017), respectively. We follow the data split as in (Zhou et al. 2021).
- **VATEX** (Wang et al. 2019): This is a bilingual video-text retrieval dataset with over 41,250 videos, each paired with 10 English and 10 Chinese sentences. We use only the annotated English captions from the training set and generate the corresponding Chinese translations using MT. We adopt a similar data partition as (Chen et al. 2020).

Evaluation metrics. Following previous work (Wang et al. 2022), for video-text retrieval, we use rank-based metrics, namely $R@K$, mean Average Precision (mAP), and sum of all Recalls (SumR) to evaluate the performance. For image-text retrieval, we only report the SumR.

Implementation Details. We apply the CLIP (ViT-B/32) (Radford et al. 2021) and mBERT-base (Devlin et al. 2018) as the image and text encoder, respectively. For video encoder, we adopt I3D (Carreira and Zisserman 2017) video features and use multi-layer perceptron followed by mean-pooling. Besides, we set $\lambda = 0.6$, $\alpha = 0.4$, and $\tau = 0.07$ in our experiments. The batch size is 128, and an Adam optimizer with an initial learning rate 2.5e-5 and adjustment schedule similar to (Luo et al. 2022) is utilized.

Ablation Studies

Impact of the component modules. We analyze the effect of each component using the control variable method in Table 1. The first line shows the performance of the Baseline method without the CL network. Based on the Baseline, we further add multi-level alignment and CL knowledge transfer. Both methods result in significant improvements, demonstrating the effectiveness of our approach in enhancing the alignment between V-T from two perspectives. Moreover, we observe a steady improvement when gradually adding our two components. This means that our two ideas are complementary and effective.

MLA		CLKT		en2de	en2fr	en2cs
$L^{cl}_{instance}$	L_{word}	S^{cl}_{sent}	S^{cl}_{word}			
				473.2	482.3	472.6
✓				488.1	488.7	476.0
✓	✓			490.0	493.0	480.4
✓		✓		491.5	492.9	480.7
✓	✓	✓		493.0	494.4	481.2
✓	✓		✓	494.3	496.1	483.6
✓	✓	✓	✓	498.0	499.7	485.8

Table 1: Ablation study of each model component on Multi30K. "MLA" and "CLKT" indicate the multi-level alignment and CL knowledge transfer, respectively.

Method	en2de	en2fr	en2cs
Cross-attention	495.7	496.1	480.4
Ours	**498.0**	**499.7**	**485.3**

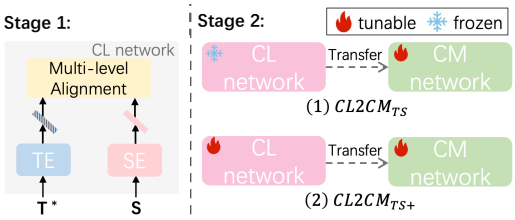
Table 2: Ablation study of generating pseudo-labels in self-supervised word-level alignment on Multi30K. "Cross-attention" indicate that generating pseudo word alignment label relies on the cross-attention module.

The exploration of self-supervised word-level alignment. In table Table 2, we attempt to replace the OT-based method with the cross-attention module. In specific, we use source-language word representations as query, target-language ones as key and value, and apply generated attention as the pseudo-label. The results show that the OT-based method yields significantly better results than the cross-attention counterpart. We speculate this can be attributed to the greater robustness of the OT-based method in handling noisy translations compared to cross-attention.

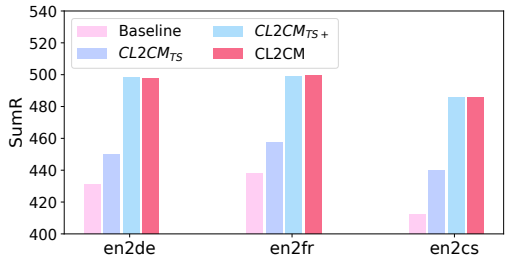
The exploration of different model variants. Figure 5 illustrates the performance of different model variants. We find that incorporating CL knowledge transfer led to better performance compared to the Baseline method. However, the CL2CM$_{TS}$ method, which only provides CL knowledge without representation improvement, resulted in sub-optimal results. Additionally, the use of the two-stage training did not lead to significant performance gains. Therefore, we used the end-to-end training strategy in subsequent experiments. We suspect that the pre-trained language models used in our experiments already have strong cross-lingual alignment capabilities, and the dataset we used is relatively small. Moreover, even without the first-stage training, cross-lingual alignment can still achieve rapid convergence. While using additional parallel corpora in the first stage may lead to further performance improvements, it is not the focus of this paper, we leave further exploration to future research.

Evaluation on Cross-lingual Image-Text Retrieval

We evaluate the performance of our CL2CM method against state-of-the-art approaches on two widely used image-text retrieval datasets, *i.e.,* Multi30K and MSCOCO. Notably, M^{3}P (Ni et al. 2021), UC^2 (Zhou et al. 2021), CCLM (Zeng et al. 2022), MURAL (Jain et al. 2021), and MLA (Zhang, Hu, and Jin 2022) were pre-trained on large-scale multi-lingual vision-language datasets, while NRCCR (Wang et al. 2022) and our method do not require additional pre-training data. As shown in Table 3, our CL2CM outperforms all two-stream methods, achieving significant performance gains. When equipped with a powerful backbone (*i.e.,* SwinTransformer for image encoding and XLM-R for text encoding), CL2CM† achieved significant performance gains, demonstrating the generalizability of our approach. Besides, the single-stream CCLM, which achieves better performance by adopting the cross-modal fusion module additionally en-



(a) Two-stage CL2CM variants training pipeline



(b) Performance comparison of different variants

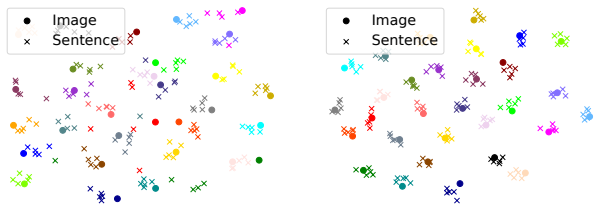
Figure 5: Ablation study to investigate the impact of different CL knowledge transfer approaches on Multi30K. Two-stage CL2CM variants initially train a CL network, followed by knowledge transfer to the CM network.

Method	Multi30K			MSCOCO	
	en2de	en2fr	en2cs	en2zh	en2ja
Single-Stream:					
M^{3}P*	351.0	276.0	220.8	332.8	336.0
UC^2*	449.4	444.0	407.4	492.0	430.2
CCLM*	**540.0**	**545.4**	**536.4**	**546.0**	**532.8**
Two-Stream:					
MURAL*	456.0	454.2	409.2	-	435.0
MLA*	495.6	510.0	457.2	-	482.4
NRCCR	480.6	482.1	467.1	512.4	507.0
CL2CM	498.0	499.7	485.3	522.0	515.9
CL2CM†	**530.4**	**536.0**	**526.3**	**544.3**	**546.2**

Table 3: Cross-lingual image-text retrieval results on Multi30K and MSCOCO (the source language is English and the target language is non-English). *: Models pre-trained on large-scale datasets, e.g., CC3M and its MT version. †: Model uses the same initialization parameters in the backbone with CCLM. The metric is the sum of all Recalls (sumR) following previous works. During inference, CL2CM† is 11x faster than CCLM.

Method	Text-to-Video Retrieval				Video-to-Text Retrieval				SumR
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	
MMP (Huang et al. 2021a)	23.9	55.1	67.8	-	-	-	-	-	-
MMP (Huang et al. 2021a)*	29.7	63.2	75.5	-	-	-	-	-	-
NRCCR (Wang et al. 2022)	30.4	65.0	75.1	45.64	40.6	72.7	80.9	32.40	364.8
CL2CM	**32.1**	**66.7**	**77.3**	**47.49**	**48.2**	**77.1**	**85.5**	**35.77**	**386.9**

Table 4: Cross-lingual video-text retrieval results on VATEX (the source language is English and the target language is Chinese). *: Model pre-trained on a large-scale dataset Multi-HowTo100M (Huang et al. 2021a).



(a) Baseline method (b) CL2CM

Figure 6: T-SNE visualization. Dots with the same color indicate representations belonging to the same class.

hances the interaction between vision and text data. However, this comes at the cost of efficiency, as all possible query-candidate pairs need to be fed into the fusion modules during inference. In contrast, our method CL2CM† is a two-stream architecture, allowing for more efficient calculation of similarity scores. Specifically, the training time of our proposed model is approximately 4% shorter than that of CCLM. In terms of inference time, CCLM takes 11x longer than ours. In short, CL2CM† achieves comparable performance with CCLM while achieving a good trade-off between performance and computational cost.

Evaluation on Cross-lingual Video-Text Retrieval

As shown in Table 4, our CL2CM outperforms all compared methods by a large margin on VATEX, and these methods all focus on instance-level alignment. Moreover, even without resorting to extra pre-training datasets, our CL2CM outperforms the MMP model that was pre-trained on large-scale multi-lingual multi-modal datasets by 4.6% in terms of sumR in text-to-video retrieval. This result demonstrates the effectiveness of our proposed CL2CM framework.

Visualization of Representations

We visualize image and target-language sentence representations using t-SNE for CL2CM and the Baseline method. Specifically, we randomly selected 20 images and their corresponding 5 German sentences from the test set of Multi30K, assigning the same color to indicate the same class. As illustrated in Figure 6, we incorporate the assistance of the CL network built upon the Baseline method, making the image and target language sentence more compact in the vector space. This visualization result confirms

Figure 7: The qualitative results of text-to-image retrieval on Multi30K. The first line is text queries in German, the second line shows top-5 retrieved results. Correct answers are highlighted in green, while wrong answers are marked in red. The text within the parentheses ( ) represents the corresponding English translation.

that our CL2CM method effectively improves the alignment quality of image and target language representations.

Qualitative Results

As shown in Figure 7, the top-5 ranking images not only contain the right objects but also accurately represent the semantic relationships among them. For instance, the retrieval results in the first example are all semantically related to the objects "a boy and a girl" and the action "look at" expressed in the text. This result demonstrates that our CL2CM method can more comprehensively capture the correspondence between images and target-language sentences.

Conclusion

This paper proposes a new framework, CL2CM, to improve the alignment between vision and target language. By transferring knowledge from the CL network to the CM network, CL2CM is capable of modeling the correspondence between data pairs accurately and alleviating the effect of noisy translations. Extensive experiments on three benchmarks demonstrate the effectiveness of CL2CM and its robustness in the presence of noisy translations.

Acknowledgments

This work was supported by the Pioneer and Leading Goose R&D Program of Zhejiang (No. 2023C01212), Young Elite Scientists Sponsorship Program by CAST (No. 2022QNRC001), and the National Natural Science Foundation of China (No. 61976188), Alibaba Group through Alibaba Research Intern Program.

References

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Chang, W.; Shi, Y.; Tuan, H.; and Wang, J. 2022. Unified optimal transport framework for universal domain adaptation. *Advances in Neural Information Processing Systems*, 35: 29512–29524.

Chen, S.; Zhao, Y.; Jin, Q.; and Wu, Q. 2020. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10638–10647.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, J.; Chen, X.; Zhang, M.; Yang, X.; Chen, S.; Li, X.; and Wang, X. 2022a. Partially Relevant Video Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, 246–257.

Dong, J.; Wang, Y.; Chen, X.; Qu, X.; Li, X.; He, Y.; and Wang, X. 2022b. Reading-strategy Inspired Visual Representation Learning for Text-to-Video Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*.

Dong, J.; Zhang, M.; Zhang, Z.; Chen, X.; Liu, D.; Qu, X.; Wang, X.; and Liu, B. 2023. Dual Learning with Dynamic Knowledge Distillation for Partially Relevant Video Retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11302–11312.

Elliott, D.; Frank, S.; Sima'an, K.; and Specia, L. 2016. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*.

Fang, Z.; Wang, J.; Hu, X.; Wang, L.; Yang, Y.; and Liu, Z. 2021. Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1428–1438.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network.

Huang, P.-Y.; Patrick, M.; Hu, J.; Neubig, G.; Metze, F.; and Hauptmann, A. 2021a. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. *arXiv preprint arXiv:2103.08849*.

Huang, Z.; Niu, G.; Liu, X.; Ding, W.; Xiao, X.; Wu, H.; and Peng, X. 2021b. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34: 29406–29419.

Jain, A.; Guo, M.; Srinivasan, K.; Chen, T.; Kudugunta, S.; Jia, C.; Yang, Y.; and Baldridge, J. 2021. Mural: multimodal, multitask retrieval across languages. *arXiv preprint arXiv:2109.05125*.

Kim, J. M.; Koepke, A.; Schmid, C.; and Akata, Z. 2023. Exposing and Mitigating Spurious Correlations for Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2584–2594.

Li, X.; Xu, C.; Wang, X.; Lan, W.; Jia, Z.; Yang, G.; and Xu, J. 2019. COCO-CN for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9): 2347–2360.

Li, Y.; Min, K.; Tripathi, S.; and Vasconcelos, N. 2023. SViTT: Temporal Learning of Sparse Video-Text Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18919–18929.

Liu, B.; Zheng, Q.; Wang, Y.; Zhang, M.; Dong, J.; and Wang, X. 2022. FeatInter: exploring fine-grained object features for video-text retrieval. *Neurocomputing*, 496: 178–191.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2022. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304.

Ni, M.; Huang, H.; Su, L.; Cui, E.; Bharti, T.; Wang, L.; Zhang, D.; and Duan, N. 2021. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3977–3986.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3967–3976.

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.

Rao, J.; Qian, T.; Qi, S.; Wu, Y.; Liao, Q.; and Wang, X. 2021. Student can also be a good teacher: Extracting knowledge from vision-and-language model for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3383–3387.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2015. FitNets: Hints for Thin Deep Nets. arXiv:1412.6550.

Sun, Z.; Yu, H.; Song, X.; Liu, R.; Yang, Y.; and Zhou, D. 2020. MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2158–2170.

Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.-F.; and Wang, W. Y. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4581–4591.

Wang, Y.; Dong, J.; Liang, T.; Zhang, M.; Cai, R.; and Wang, X. 2022. Cross-Lingual Cross-Modal Retrieval with Noise-Robust Learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, 422–433.

Wang, Y.; Wang, S.; Luo, H.; Dong, J.; Wang, F.; Han, M.; Wang, X.; and Wang, M. 2023. Dual-view Curricular Optimal Transport for Cross-lingual Cross-modal Retrieval. *arXiv preprint arXiv:2309.05451*.

Wu, W.; Luo, H.; Fang, B.; Wang, J.; and Ouyang, W. 2023. Cap4Video: What Can Auxiliary Captions Do for Text-Video Retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10704–10713.

Xu, R.; Liu, P.; Wang, L.; Chen, C.; and Wang, J. 2020. Reliable weighted optimal transport for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4394–4403.

Yoshikawa, Y.; Shigeto, Y.; and Takeuchi, A. 2017. Stair captions: Constructing a large-scale japanese image caption dataset. *arXiv preprint arXiv:1705.00823*.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Zeng, Y.; Zhou, W.; Luo, A.; and Zhang, X. 2022. Cross-View Language Modeling: Towards Unified Cross-Lingual Cross-Modal Pre-training. *arXiv preprint arXiv:2206.00621*.

Zhang, L.; Hu, A.; and Jin, Q. 2022. Generalizing Multimodal Pre-training into Multilingual via Language Acquisition.

Zheng, Q.; Dong, J.; Qu, X.; Yang, X.; Wang, Y.; Zhou, P.; Liu, B.; and Wang, X. 2023. Progressive localization networks for language-based moment localization. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2): 1–21.

Zhou, M.; Zhou, L.; Wang, S.; Cheng, Y.; Li, L.; Yu, Z.; and Liu, J. 2021. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4155–4165.