

COVLR: COORDINATING CROSS-MODAL CONSISTENCY AND INTRA-MODAL RELATIONS FOR VISION-LANGUAGE RETRIEVAL

Fengqiang Wan¹, Xiangyu Wu¹, Zhihao Guan¹, Yang Yang^{1*}

¹Nanjing University of Science and Technology
wfq011207@163.com {xywu, zhguan, yyang}@njjust.edu.cn

ABSTRACT

Vision-language retrieval aims to perform cross-modal instances search by learning consistent vision-language representations. However, in real applications, vision-language divergence always results in strong and weak modalities and different modalities have various performances in uni-modal tasks. In this paper, we reveal that traditional vision-language hard consistency disrupts the relationships among uni-modal instances considering the weak-strong modal scenario, causing a decline in uni-modal retrieval capability. To address this issue, we propose Coordinated Vision-Language Retrieval (CoVLR), a solution that cooperatively optimizes both cross-modal consistency and intra-modal structure-preserving objectives via a meta-learning strategy. Specifically, CoVLR utilizes intra-modal structure-preserving as the meta-test task to validate the cross-modal consistency loss, which is considered the meta-train task. The effectiveness of CoVLR is validated through extensive experiments on commonly used datasets, which demonstrate superior results compared to other baselines on various retrieval scenarios.

Index Terms— Multi-Modal Learning, Visual-Language Retrieval, Meta-Optimization

1. INTRODUCTION

An important task in multi-modal machine learning is cross-modal retrieval, which involves searching for instances in one modality based on instances from another modality. In this paper, we focus primarily on two modalities, including visual signals and natural language. The main challenge of vision-language retrieval lies in the semantic gap within heterogeneous data. To address this issue, many approaches have been developed to learn consistent representations between vision and language by using well-aligned data [1, 2, 3].

Generally, the methods rely on the premise that vision-language modalities can adequately express each other so that hard consistency can be done naturally. However, due to the presence of modal sufficiency [4, 5], there exists a distinction between strong and weak modalities in terms of vision and language considering uni-modal retrieval capability. We adopt MS-COCO (1K) dataset as an example to il-

lustrate the different performance of uni-modal retrieval in Figure 1, where retrieval capacity is measured by the matrix Normalized Discounted Cumulative Gain (NDCG). This observation is consistent across different benchmarks, such as MS-COCO (5K) and VizWiz, for more details can refer to the experiment section. Considering the impact of modal divergence on modal joint optimization, there have been prior efforts to address this challenge in multi-modal tasks [6, 7], but they mainly focus on the classification task that needs shared ground-truths. In the cross-modal retrieval task, we also discover that consistent representations are beneficial for vision-language retrieval, yet have drawbacks for uni-modal retrieval. Traditional hard cross-modal consistent losses aim to learn consistent vision-language representations without considering modal discrepancy, which cannot guarantee that the relationships among strong modal instances are not affected by weak modalities. In other words, enforcing hard cross-modal consistency constraints may lead to negative bidirectional guidance between weak and strong modalities, leading to uni-modal retrieval performance degrading, and this impact on the strong modality is more noticeable.

What we need is coordinated vision-language representation learning that improves uni-modal accuracy while maintaining cross-modal retrieval capacity. The main challenge lies in studying and alleviating the inconsistency between cross-modal consistency and intra-modal structure-preserving tasks. To overcome this issue, we introduce an effective meta-optimization framework, Coordinated Vision-Language Retrieval (CoVLR). Particularly, CoVLR treats the cross-modal consistency objective as the meta-train task and the uni-modal structure-preserving objective as the meta-test task in a meta-learning scheme. The meta-train task optimizes the network for learning consistent representations, while the meta-test task validates the optimization with the uni-modal structure-preserving task, such as classification, to preserve the instances' relation. Consequently, we can ensure both cross-modal consistency and uni-modal structure simultaneously.

2. RELATED WORK

Vision-Language Retrieval aims to learn consistent representations of different modalities, thus retrieving instances of one modality based on queries from another modality. Initial

*Corresponding author

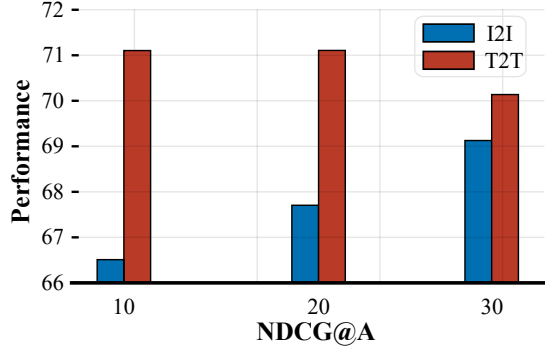


Fig. 1. Exploring modal divergence on retrieval task on MS-COCO (1K) dataset, and the uni-modal retrieval performance of image and text is summarized using Swin Transformer and Bert backbones, respectively.

approaches involved dual-encoder architectures, where language and vision modalities are encoded independently and embedded into a common space to maximize cross-modal similarity [1, 2, 3, 8, 9]. For example, [1] employed two independent modal encoders and incorporated a hard triplet loss function. To consider the fine-grained information, [2] further utilized the Faster R-CNN for image modality and discovered the full latent alignments using both image regions and words in a sentence as context. Considering the relationships between regions, [3] proposed a similarity graph reasoning module relying on a graph convolution network. With the developments in Transformer-based language understanding [10], large-scale vision-language transformers have inspired deeper modal interaction in retrieval models. For instance, [11] utilized the cross-attention layer to enable interaction between visual and textual inputs. The approaches above primarily focus on applying cross-modal hard consistency constraints to achieve consistent representations. However, they ignore the negative effects of modal divergence on uni-modal performance. Recently, some researchers have focused on joint training for multi-modal tasks, as demonstrated in [12]. These works typically use multi-task losses, which may be sub-optimal for balancing different objectives.

Meta-Learning refers to training a model that can extract useful information from the meta-train task and uses this knowledge to improve its performance on the meta-test tasks. This concept can be categorized into metric-based, model-based, and optimization-based techniques. In this paper, our work is mostly related to the Model-Agnostic Meta-Learning (MAML) [13], an optimization-based technique for quickly adapting to new tasks by learning a good set of initial parameters. Based on this idea, several variants have been proposed. For example, [14] considered domain alignment and classification objectives in a meta-learning scheme for unsupervised domain adaptation; [15] extended conventional few-shot meta-learning by generalizing its setup to various multi-modal tasks. In the case that multi-task losses are sub-optimal for co-optimization of cross-modal consistency constraint and

intra-modal structure-preserving loss, we attempt to learn a well-done cross-modal model that can also perform well on uni-modal objectives via a meta-learning strategy.

3. PROPOSED METHOD

3.1. Vision-Language Retrieval Model

The goal of the vision-language retrieval task is to learn consistent representations of heterogeneous modalities for conducting cross-modal searches. Besides, as mentioned above, our main focus in this paper is to preserve and potentially enhance the uni-modal structures during training. Without any loss of generality, the parallel vision-language pairs are given, i.e., $\mathcal{D} = \{(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^N$, where \mathbf{v}_i denotes the i -th image, and \mathbf{w}_i is the corresponding sentence.

Vision and Language encoders. The vision encoder utilizes Swin Transformer, a vision transformer to generate detailed visual representations. Specifically, the image \mathbf{v} is split into L_I patches, and these patches are then fed into the transformer layers, resulting in $L_I + 1$ concept representations, which include the embedding of the [CLS] token. Meanwhile, the language encoder employs BERT [10] to map the input sentence \mathbf{w} to the same dimensional subspace as the image, with $L_T + 1$ tokens.

Cross-Modal encoders. We construct two cross-attention transformers, i.e., image2text and text2image cross-attention transformers, which employ cross-modal interaction layers to process visual and textual representations. In detail, we utilize the multi-head cross-attention to learn correlated representations: $Att(\mathbf{v}) = softmax(\frac{Q_v K_w^T}{\sqrt{d_M}})V_w$, $Att(\mathbf{w}) = softmax(\frac{Q_w K_v^T}{\sqrt{d_M}})V_v$, where Q ., K ., and V represent queries, keys, and values, respectively. $d_M = d_{model}/M$ as the multi-head attention is composed of M parallel heads, where d_{model} is the dimension of common subspace. The image and text representations are fused together through cross-attention at every layer of the cross-modal encoders.

Cross-Modal Objective. The cross-modal objective comprises three elements: masked language modeling (MLM), masked patch modeling (MPM), and image-text contrastive learning (ITC).

MLM is designed for predicting the masked text tokens based on the contextualized vectors. Following [10], we randomly mask 15% of the input tokens, with 10% replaced by random tokens, 10% unchanged, and 80% set to [MASK]. The output of the image-to-text cross-modal encoder is then passed through a linear layer with softmax operation:

$$\ell_{mlm} = \mathbb{E}_{(\mathbf{w}^{msk})} CE(\mathbf{y}^{msk}, \phi_{mlm}(\hat{\mathbf{w}}^{msk})) \quad (1)$$

where \mathbf{w}^{msk} denotes the set of masked words, and $\hat{\mathbf{w}}^{msk}$ represents the corresponding output representations. \mathbf{y}^{msk} denotes the ground-truths, and $\phi_{mlm}(\hat{\mathbf{w}}^{msk})$ is the predictions of masked words, where ϕ_{mlm} denotes the classifier. CE is the cross-entropy loss.

MPM is intended to restore the representation of image tokens that have been masked in advance, by utilizing contextualized vectors. Similarly, we randomly mask 15% of the input tokens as mentioned above. The output from the text-to-image cross-modal encoder is fed to a linear layer:

$$\ell_{mpm} = \mathbb{E}_{(\mathbf{v}^{msk})} MSE(\mathbf{v}^{msk}, \phi_{mpm}(\hat{\mathbf{v}}^{msk})) \quad (2)$$

where \mathbf{v}^{msk} denotes the masked image patches, and $\hat{\mathbf{v}}^{msk}$ represents the output representations. $\phi_{mpm}(\hat{\mathbf{v}}^{msk})$ denotes the learned representations with ϕ_{mpm} head. MSE is the mean-squared loss.

ITC aims to learn better uni-modal representations before fusion. A similarity function is learned in which the parallel image-text pairs are assigned higher similarity scores. We define $s(\mathbf{v}, \mathbf{w}) = \phi_v(\mathbf{v}_{CLS})^\top \phi_w(\mathbf{w}_{CLS})$ and $s(\mathbf{w}, \mathbf{v}) = \phi_w(\mathbf{w}_{CLS})^\top \phi_v(\mathbf{v}_{CLS})$, where ϕ_v and ϕ_w are linear transformations with softmax operator, which take the [CLS] embeddings output by cross-modal encoders as joint representations for prediction. Then the cross-modal contrastive learning can be formulated as:

$$\ell_{itc} = \frac{1}{2} \mathbb{E}_{(\mathbf{v}, \mathbf{w})} [CE(\mathbf{y}^{i2t}(\mathbf{v}), \mathbf{p}^{i2t}(\mathbf{v})) + CE(\mathbf{y}^{t2i}(\mathbf{w}), \mathbf{p}^{t2i}(\mathbf{w}))]$$

$$p_b^{i2t}(\mathbf{v}) = \frac{\exp(s(\mathbf{v}, \mathbf{w}_b)/\tau)}{\sum_{b=1}^B \exp(s(\mathbf{v}, \mathbf{w}_b)/\tau)} \quad p_b^{t2i}(\mathbf{w}) = \frac{\exp(s(\mathbf{w}, \mathbf{v}_b)/\tau)}{\sum_{b=1}^B \exp(s(\mathbf{w}, \mathbf{v}_b)/\tau)} \quad (3)$$

where $p_b^{i2t}(\mathbf{v})$ and $p_b^{t2i}(\mathbf{w})$ denote softmax-normalized image-to-text and text-to-image similarity with batch size B and temperature scale parameter τ . CE denotes cross-entropy loss. Given $\mathbf{y}^{i2t}(\mathbf{v}) \in \mathcal{R}^B$ and $\mathbf{y}^{t2i}(\mathbf{w}) \in \mathcal{R}^B$ as the ground-truth similarity labels, where $y_b = 1$ if (\mathbf{v}, \mathbf{w}) is aligned and $y_b = 0$ otherwise, ℓ_{itc} pulls matched cross-modal samples closer and mismatched ones farther.

Overall, the cross-modal objective can be represented as:

$$\ell_{cro} = \ell_{itc} + \ell_{mlm} + \ell_{mpm} \quad (4)$$

Note that ℓ_{mlm} and ℓ_{mpm} can also be considered cross-modal objectives, as the masked tokens are involved in interactions with other modal queries in the cross-attention. Considering that the three losses are of equal importance and of the same magnitude, no hyperparameters are introduced.

Intra-Modal Objective. In scenarios where some uni-modal data have class-labels while others do not, we propose two optimization strategies for preserving the discriminative capability of intra-modal representations: classification and contrastive learning.

When the class label is known, such as for MS-COCO, the classification task can be directly adopted, which learns structure-aware representations with the cross-entropy loss:

$$\ell_{cls} = \mathbb{E}_{(\mathbf{w}, \mathbf{y})} CE(\mathbf{y}, \phi_{tcls}(\mathbf{w}_{CLS})) + CE(\mathbf{y}, \phi_{icls}(\mathbf{v}_{CLS})) \quad (5)$$

where $\mathbf{y} \in \mathcal{R}^C$, and C denotes the number of class. ϕ_{tcls} and ϕ_{icls} denote the text classifier and the image classifier.

However, many datasets (e.g., VizWiz) do not have class ground truths. Therefore, we adopt unsupervised contrastive learning for substitution to preserve the instances' relations to some extent, which can be formulated as:

$$\ell = \ell_{ince} + \ell_{tnce}$$

$$\ell_{ince} = - \sum_{\mathbf{v}} \log \frac{\exp(s(\mathbf{v}, \mathbf{v}^+)/\tau)}{\sum_{b=1}^B \exp(s(\mathbf{v}, \mathbf{v}_b)/\tau)} \quad (6)$$

$$\ell_{tnce} = - \sum_{\mathbf{w}} \log \frac{\exp(s(\mathbf{w}, \mathbf{w}^+)/\tau)}{\sum_{b=1}^B \exp(s(\mathbf{w}, \mathbf{w}_b)/\tau)}$$

where τ is still the temperature parameter and $s(\cdot)$ measures the similarity between instances, as used in ITC. The negative anchors are sampled from the same batch, and the positive anchors \mathbf{v}^+ and \mathbf{w}^+ are generated by perturbing instances using weak augmentation techniques, which mainly include standard flip and translation strategies for images and back translation for texts. Further analyses of contrastive learning variants can be found in the supplementary.

3.2. Coordinate Optimization

The cross-modal objective aims to learn consistent representations, while uni-modal objectives are to learn structure-preserving representations. The optimization direction of consistency may be inconsistent with that of the classification task. In cases where the two modalities provide diverse information, cross-modal consistency prefers to output the pair of hypotheses, which simply minimizes the disagreement rather than the optimal classifiers. Such optimization inconsistency will lead to inferior performance. Hence, a critical challenge is to incorporate optimization consistency for both cross-modal and uni-modal objectives.

Therefore, we attempt to promote the optimization consistency between these two objectives by designing a meta-optimization strategy. Our approach treats the cross-modal objective as the meta-train task and the uni-modal objective as the meta-test tasks for the same set of image-text pairs, rather than splitting the pairs as traditional MAML [13]. We utilized shared parameters Θ for vision/language/cross-modal encoders, specific parameters Θ_{cro} for the cross-modal objective, and specific parameters Θ_{uni} for the intra-modal objective. Considering the intuition that the meta-test task (uni-modal structure-preserving) evaluates the effect of model optimization on the meta-train task (cross-modal consistency constraint), the overall objective can be formulated as:

$$\min_{\Theta, \Theta_{cro}, \Theta_{uni}} \ell_{cro}(\Theta, \Theta_{cro}) + \ell_{uni}(\Theta - \alpha \nabla_{\Theta} \ell_{cro}(\Theta, \Theta_{cro}), \Theta_{uni}) \quad (7)$$

This aims to employ the meta-learning strategy to optimize both the loss of meta-train ℓ_{cro} and that of meta-test ℓ_{uni} after updating Θ with one gradient descent step: $\Theta' \leftarrow \Theta - \alpha \nabla_{\Theta} \ell_{cro}(\Theta, \Theta_{cro})$, where α denotes the meta-learning rate. As suggested by [13], we simplify the back-propagation

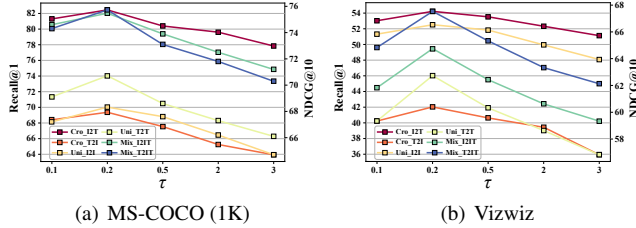


Fig. 2. Parameter analyses. We verify the influence of parameters τ for CoVLR. Cro-/Uni-/Mix- denote cross-modal, uni-modal, and mixed retrieval.

of gradients by excluding higher-order ones to reduce computational complexity. Additionally, according to [14], the Equation 7 can be approximated using the first-order Taylor expansion as:

$$\min_{\Theta, \Theta_{cro}, \Theta_{uni}} \ell_{cro}(\Theta, \Theta_{cro}) + \ell_{uni}(\Theta, \Theta_{uni}) - \alpha \nabla_{\Theta} \ell_{cro}(\Theta, \Theta_{cro}) \nabla_{\Theta} \ell_{uni}(\Theta, \Theta_{uni}) \quad (8)$$

In fact, the final term in Equation 8 maximizes the product of $\nabla_{\Theta} \ell_{cro}$ and $\nabla_{\Theta} \ell_{uni}$, which encourages the optimization directions of the two tasks to be consistent. This explicit interaction enables coordinated cross-modal and uni-modal retrieval, resulting in improvements in uni-modal and cross-modal retrieval performance.

4. EXPERIMENTS

In this section, we focus on retrieval results and ablation study due to page limitations. More experiments, e.g., mixed retrieval, model convergence, meta-optimization analysis, and case study can be found in the supplementary material.

4.1. Experimental Setups

Datasets and Baselines. We conduct the experiments on the following three datasets: MS-COCO (1k/5k) and VizWiz. For comparative analysis, we evaluated four types of state-of-the-art approaches: 1) dual-encoder retrieval methods, including VSE [1], SCAN [2], IMRAM, SGRAF [3], GSMN, VSRN, and NAAF[16]. 2) transformer based retrieval methods, including ALBEF, X-VLM, and BLIP. 3) uni-modal models, e.g., Swin Transformer and BERT. 4) coordinate optimization methods, i.e., CYCLIP and UMT. Note that CYCLIP incorporates regularizers about uni-modal and cross-modal structural information into the cross-modal contrastive item in the form of a multi-task loss, and UMT optimizes the cross-modal model by distilling the instance’ relations learned by the uni-modal model. More details about datasets and baselines are in the supplementary material. To validate the retrieval performance, we focus on three tasks: 1) cross-modal retrieval. 2) uni-modal retrieval. 3) mixed retrieval (detailed results available in the supplementary material).

Evaluation Metrics. The performance of CoVLR is evaluated using the recall at A (R@A) metric, which is commonly used in most cross-modal retrieval methods for both

image2text (I2T) and text2image (T2I) retrieval tasks [1]. During cross-modal retrieval, the corresponding captions of the given image or corresponding images of the given caption are expected, while in uni-modal retrieval, what we need are relatively similar instances. As a result, R@A metric as a binary correlation that only examines whether the retrieval results are relevant is not suitable for uni-modal (I2I and T2T) and mixed retrieval (I2IT and T2IT). A more comprehensive metric NDCG@A is adopted instead to promote the items with higher relevance scores to appear in better ranking positions. Note that the ranking is computed using text similarity scores (i.e., ROUGE-L) between a sentence and the sentences associated with a certain image.

4.2. Retrieval Results

Evaluation on Cross-Modal Retrieval. Firstly, the cross-modal retrieval performance is evaluated through two tasks: text-to-image (T2I) and image-to-text (I2T) retrievals. Table 1 shows the comparison of CoVLR with state-of-the-art methods. To ensure experimental fairness, we exclusively utilized the given datasets for model training, rather than pre-training additional data as the large-scale models (i.e., ALBEF, X-VLM, CYCLIP, and BLIP) do. So we retrained the ALBEF and other large models from scratch (marked with ‘*’), thus causing different results compared to the original paper. The results reveal that: 1) CoVLR outperforms the best cross-modal retrieval method, BLIP*, on I2T Recall@1 and T2I Recall@1 by 1.0/0.9/7.7 and 1.7/2.2/4.4 respectively, on the listed datasets. This phenomenon reveals that CoVLR with structure preservation can also enhance the learning of cross-modal consistent representations by bringing visual (i.e., weak modality) representations closer to the language (i.e., strong modality) ones without compromising the original uni-modal structure. 2) CoVLR performs better than CYCLIP* (multi-task optimization) and UMT (uni-modal distillation learning) across most settings, which emphasizes that meta-optimization emerges as a relatively superior strategy because direct multi-task loss and distillation learning prove challenging to balance the learning of two terms, which may cause sub-optimal bias during training. 3) the performance of large-scale pre-trained models is limited when data is limited, e.g., ALBEF* and X-VLM* exhibit competitive performance with dual-encoder models like SGRAF and NAAF.

Evaluation on Uni-Modal Retrieval. Then, the performance preservation on uni-modal retrieval is focused on, including image-to-image (I2I) and text-to-text (T2T) retrieval, as is illustrated in Table 2. Experiment results indicate that: 1) CoVLR performs better than all cross-modal retrieval comparison methods in both I2I and T2T retrievals and even outperforms the best uni-modal retrieval methods, especially in the visual modality. E.g., CoVLR outperforms the Swin Transformer on I2I NDCG@10 by 1.8/3.3/3.3 respectively on three datasets. On the other hand, CoVLR still remains a performance gap for the strong modality (i.e., text modal-

Table 1. Cross-modal retrieval performance comparison. Evaluation criteria are R@A.

Methods	MS-COCO (1K)						MS-COCO (5K)						Vizwiz					
	I2T			T2I			I2T			T2I			I2T			T2I		
	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10
VSE	49.5	81.0	90.0	38.1	73.3	85.1	39.0	67.9	79.5	29.3	59.1	72.4	35.1	58.1	65.4	25.3	48.1	58.4
SCAN	72.7	94.8	98.4	58.8	88.4	94.8	50.4	82.2	90.0	38.6	69.3	80.4	42.6	65.8	74.8	26.0	48.7	59.0
IMRAM	76.7	95.6	98.5	61.7	89.1	95.0	53.7	83.2	91.0	39.7	69.1	79.8	42.5	67.5	78.6	27.6	52.5	63.6
SGRAF	79.6	96.2	98.5	63.2	90.7	96.1	57.8	83.5	91.6	41.9	71.3	81.3	43.9	73.4	80.1	28.8	54.4	64.2
GSMN	78.4	96.4	98.6	63.3	90.1	95.7	55.2	81.3	86.2	37.2	68.3	77.3	43.3	72.4	79.3	26.9	53.2	63.8
VSRN	76.2	94.8	98.2	62.8	89.7	95.1	53.0	81.1	89.4	40.5	70.6	81.1	39.0	64.1	71.9	21.7	51.6	62.3
NAAF	78.1	96.1	98.6	63.5	89.6	95.3	58.9	85.2	92.0	42.5	70.9	81.4	44.1	69.9	78.0	31.0	54.8	64.2
ALBEF*	72.5	94.4	97.2	57.6	88.4	94.2	52.3	80.4	88.1	39.8	67.2	82.6	42.2	69.0	80.3	30.6	61.4	73.3
X-VLM*	78.2	96.4	98.5	66.7	91.6	95.1	58.6	80.4	90.7	43.3	74.5	82.7	53.8	80.6	88.1	40.7	71.3	80.9
BLIP*	81.4	96.7	99.2	67.7	92.1	96.3	57.8	84.1	91.2	43.4	72.8	82.7	46.2	73.2	82.2	37.2	64.0	74.5
CYCLIP*	78.4	96.0	98.9	65.7	91.5	96.6	53.0	81.0	89.3	41.0	71.7	82.7	52.4	79.7	87.2	39.2	70.5	79.6
UMT	81.5	97.3	99.1	69.2	93.2	97.2	58.4	85.0	92.7	43.6	74.5	84.3	52.7	81.2	87.6	41.8	71.7	80.4
CoVLR	82.4	97.5	99.3	69.4	93.7	97.4	58.7	85.8	92.9	45.6	76.5	85.3	53.9	81.8	88.1	41.6	72.3	81.2

Table 2. Uni-modal retrieval performance comparison. Evaluation criteria are NDCG@A (@A for simplicity).

Methods	MS-COCO (1K)						MS-COCO (5K)						Vizwiz					
	I2I			T2T			I2I			T2T			I2I			T2T		
	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50
VSE	62.5	64.2	66.5	67.9	69.4	70.2	43.2	45.2	48.5	43.8	45.5	48.6	54.2	55.6	56.7	57.8	58.6	59.3
SCAN	64.1	65.1	66.7	69.9	70.3	69.5	59.7	61.2	63.4	63.5	64.8	65.8	59.7	61.6	65.1	62.5	62.6	62.8
IMRAM	65.1	66.6	68.6	70.2	70.5	70.1	60.4	62.0	64.5	64.4	62.1	65.9	59.5	61.5	65.0	61.2	62.4	63.1
GSMN	61.6	63.3	65.9	59.7	61.1	61.5	55.1	56.7	59.1	49.1	50.7	53.1	56.4	57.3	58.6	52.5	56.3	55.9
SGRAF	62.5	64.0	66.3	61.3	63.4	65.0	56.7	58.3	60.8	53.8	56.1	58.4	59.1	61.1	64.7	56.3	59.3	62.9
VSRN	66.4	66.7	67.5	70.6	70.6	70.5	64.0	65.2	66.3	62.1	63.4	65.2	60.8	62.8	65.9	61.9	61.6	62.1
NAAF	63.7	64.8	66.8	70.7	70.9	72.1	59.5	60.8	62.8	65.3	66.6	68.6	59.1	61.2	64.7	61.7	61.8	62.9
ALBEF*	67.0	67.1	69.0	66.8	66.5	66.1	60.0	62.2	63.9	58.0	61.2	62.9	63.4	65.1	68.3	54.3	55.6	57.2
X-VLM*	67.1	67.8	69.0	52.9	54.1	55.5	64.8	65.3	68.5	49.8	51.6	53.5	64.5	66.1	69.1	49.4	51.4	53.8
BLIP*	68.1	68.9	70.1	64.4	63.5	62.4	64.0	65.5	67.6	59.0	59.3	59.5	63.0	64.9	68.0	58.3	58.8	59.6
CYCLIP*	67.3	68.2	69.0	68.9	68.4	67.3	62.7	64.0	66.3	63.4	64.5	66.8	63.0	64.5	67.4	56.0	58.2	59.9
UMT	67.5	68.8	70.7	68.1	65.9	64.4	64.8	65.8	67.7	57.3	57.2	66.3	64.0	65.5	67.4	59.0	60.2	61.9
CoVLR	68.3	69.5	71.6	70.7	70.9	72.4	65.2	68.0	70.4	65.3	66.9	69.3	65.8	66.4	69.1	62.6	62.7	63.2
Swin Transformer	66.5	67.7	69.1	-	-	-	61.5	63.2	65.7	-	-	-	62.5	64.3	66.6	-	-	-
BERT	-	-	-	71.1	71.1	70.2	-	-	-	63.4	64.8	66.1	-	-	-	66.8	66.2	65.5

Table 3. Performance comparison with different variants. Evaluation criteria are R@A and NDCG@A.

	MS-COCO (1K)						MS-COCO (5K)						Vizwiz					
	I2T			T2I			I2T			T2I			I2T			T2I		
	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10
CoVLR	82.4	97.5	99.3	69.4	93.7	97.4	58.7	85.8	92.9	45.6	76.5	85.3	53.9	81.8	88.1	41.6	72.3	81.2
w/o ℓ_{mlm}	77.7	94.2	97.4	63.6	89.0	93.5	55.1	80.9	89.6	43.0	72.5	83.7	50.5	78.9	86.7	39.8	68.5	77.6
w/o ℓ_{mpm}	80.2	95.3	97.9	67.1	90.1	95.8	56.4	82.7	91.0	43.5	74.3	84.2	51.7	80.4	87.5	40.2	70.4	79.3
w/o ℓ_{incke}/ℓ_{icl}	81.7	96.8	99.1	68.9	93.4	97.3	57.3	84.6	91.3	44.7	75.9	84.4	53.4	81.7	87.8	41.3	71.5	80.9
w/o ℓ_{tncke}/ℓ_{tcls}	80.8	96.5	99.0	68.0	93.0	97.1	56.8	82.5	90.4	43.5	73.7	83.9	52.5	80.9	87.5	40.2	70.9	80.2
	I2I			T2T			I2I			T2T			I2I			T2T		
	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50	@10	@20	@50
CoVLR	68.3	69.5	71.6	70.7	70.9	72.4	65.2	68.0	70.4	65.3	66.9	69.3	65.8	66.4	69.1	62.6	62.7	63.2
w/o ℓ_{mlm}	66.3	67.6	69.2	62.9	61.7	62.2	61.9	63.5	66.8	54.1	56.4	57.2	59.5	78.9	86.7	39.8	68.5	77.6
w/o ℓ_{mpm}	67.4	68.8	70.6	66.9	65.9	64.8	63.0	64.2	68.9	64.1	65.0	66.3	64.0	65.5	67.4	59.0	60.2	61.9
w/o ℓ_{incke}/ℓ_{icl}	67.8	68.7	70.5	69.5	69.8	71.3	64.3	65.4	69.2	65.1	66.5	68.2	65.9	66.8	68.1	61.3	62.4	62.9
w/o ℓ_{tncke}/ℓ_{tcls}	67.6	68.6	70.2	69.3	69.5	70.6	63.6	64.9	67.3	64.8	65.8	67.6	64.7	65.1	67.7	59.1	60.6	61.5

ity), which can be attributed to the degree of influence from the weak modality. 2) Although several cross-modal retrieval methods can improve the I2I retrieval, e.g., X-VLM* increases the NDCG@10 compared with Swin Transformer

(0.6/3.3/2.0 on MS-COCO (1K)/MS-COCO (5K)/Vizwiz), almost all methods have varying degrees of decline in performance in T2T retrieval, meanwhile, the strong modality decreases more than the weak modality increases even if

the improvement exists. For instance, X-VLM* decreases 18.2/13.6/17.4 of T2T NDCG@10 on the listed datasets compared with BERT, with only 0.6/3.3/2.0 promotion of I2I NDCG@10 over Swin Transformer. 3) CoVLR performs better than CYCLIP* and UMT, which reveals that meta-optimization is superior to learning structure-preserving representations by better incorporating optimization consistency.

4.3. Ablation Study

Ablation studies are conducted to confirm the effectiveness of each module, including the masked language modeling ℓ_{mlm} , masked patch modeling ℓ_{mpm} , and uni-modal objectives such as classification for MS-COCO (1K/5K) and contrastive learning for other datasets. The results are recorded in Table 3 and reveal the following observations: 1) The removal of ℓ_{mlm} results in the worst performance, emphasizing the significance of the text as a strong modality. 2) The fusion of all modules yields the best performance, suggesting that each task contributes to improving performance.

To investigate the impact of the temperature parameter, we tune τ values to $\{0.1, 0.2, 0.5, 2, 3\}$. The results are illustrated in Figure 2, which indicates that the best retrieval performance is achieved when τ is set to 0.2 for the datasets we tested. This finding suggests that the target point has a small number of similar neighbors, which can serve the learning of structure-aware representations.

5. CONCLUSION

In this paper, to address the decrease in uni-modal retrieval performance caused by hard cross-modal consistency constraints in traditional cross-modal retrieval tasks, we proposed a novel vision-language retrieval approach called Coordinated Vision-Language Retrieval (CoVLR), which balanced cross-modal and uni-modal retrieval. Based on a meta-optimization strategy, CoVLR treated cross-modal consistency as the meta-train task and intra-modal structure preservation as the meta-test task to optimize them in a coordinated manner. Experimental results showed that CoVLR outperforms baselines.

6. ACKNOWLEDGMENTS

National Key RD Program of China (2022YFF0712100), NSFC (62276131), the Fundamental Research Funds for the Central Universities (No.30922010317)

7. REFERENCES

- [1] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler, “VSE++: improving visual-semantic embeddings with hard negatives,” in *BMVC*, 2018.
- [2] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He, “Stacked cross attention for image-text matching,” in *ECCV*, 2018, pp. 201–216.
- [3] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu, “Similarity reasoning and filtration for image-text matching,” in *AAAI*, 2021, pp. 1218–1226.
- [4] Yang Yang, Yi-Feng Wu, De-Chuan Zhan, Zhi-Bin Liu, and Yuan Jiang, “Complex object classification: A multi-modal multi-instance multi-label deep network with optimal transport,” in *SIGKDD*, 2018, pp. 2594–2603.
- [5] Yang Yang, Zhao-Yang Fu, De-Chuan Zhan, Zhi-Bin Liu, and Yuan Jiang, “Semi-supervised multi-modal multi-instance multi-label deep network with optimal transport,” *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 2, pp. 696–709, 2021.
- [6] Weiyao Wang, Du Tran, and Matt Feiszli, “What makes training multi-modal classification networks hard?,” in *CVPR*, 2020, pp. 12692–12702.
- [7] Yang Yang, Jia-Qi Yang, Ran Bao, De-Chuan Zhan, Hengshu Zhu, Xiaoru Gao, Hui Xiong, and Jian Yang, “Corporate relative valuation using heterogeneous multi-modal graph neural network,” *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 211–224, 2023.
- [8] Yang Yang, Chubing Zhang, Yi-Chu Xu, Dianhai Yu, De-Chuan Zhan, and Jian Yang, “Rethinking label-wise cross-modal retrieval from a semantic sharing perspective,” in *IJCAI*, 2021, pp. 3300–3306.
- [9] Yang Yang, Ran Bao, Weili Guo, De-Chuan Zhan, Yilong Yin, and Jian Yang, “Deep visual-linguistic fusion network considering cross-modal inconsistency for rumor detection,” *Sci. China Inf. Sci.*, vol. 66, no. 12, 2023.
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *NAACL-HLT*, 2019, pp. 4171–4186.
- [11] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee, “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” in *NeurIPS*, 2019, pp. 13–23.
- [12] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover, “Cyclip: Cyclic contrastive language-image pretraining,” in *NeurIPS*, 2022, pp. 6704–6719.
- [13] Chelsea Finn, Pieter Abbeel, and Sergey Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *ICML*, 2017, pp. 1126–1135.
- [14] Guoqiang Wei, Cuiling Lan, Wenjun Zeng, and Zhibo Chen, “Metaalign: Coordinating domain alignment and classification for unsupervised domain adaptation,” in *CVPR*, 2021, pp. 16643–16653.
- [15] Risto Vuorio, Shao-Hua Sun, Hexiang Hu, and Joseph J. Lim, “Multimodal model-agnostic meta-learning via task-aware modulation,” in *NeurIPS*, 2019, pp. 1–12.
- [16] Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang, “Negative-aware attention framework for image-text matching,” in *CVPR*, 2022, pp. 15640–15649.