ROSITA: Enhancing Vision-and-Language Semantic Alignments via Cross- and Intra-modal Knowledge Integration

Yuhao Cui[1], Zhou Yu[1*], Chunqi Wang[2], Zhongzhou Zhao[2], Ji Zhang[2], Meng Wang[3], Jun Yu[1]
[1]School of Computer Science and Technology, Hangzhou Dianzi University, China
[2]Alibaba Group, China
[3]School of Computer Science and Information Engineering, Hefei University of Technology, China
{cuiyh,yuz,yujun}@hdu.edu.cn,{shiyan.wcq,zhongzhou.zhao,zj122146}@alibaba-inc.com,eric.mengwang@gmail.com

ABSTRACT

Vision-and-language pretraining (VLP) aims to learn generic multi-modal representations from massive image-text pairs. While various successful attempts have been proposed, learning fine-grained semantic alignments between image-text pairs plays a key role in their approaches. Nevertheless, most existing VLP approaches have not fully utilized the intrinsic knowledge within the image-text pairs, which limits the effectiveness of the learned alignments and further restricts the performance of their models. To this end, we introduce a new VLP method called ROSITA, which integrates the c**ROS**s- and **I**n**T**r**A**-modal knowledge in a unified scene graph to enhance the semantic alignments. Specifically, we introduce a novel structural knowledge masking (SKM) strategy to use the scene graph structure as a priori to perform masked language (region) modeling, which enhances the semantic alignments by eliminating the interference information within and across modalities. Extensive ablation studies and comprehensive analysis verifies the effectiveness of ROSITA in semantic alignments. Pretrained with both in-domain and out-of-domain datasets, ROSITA significantly outperforms existing state-of-the-art VLP methods on three typical vision-and-language tasks over six benchmark datasets.

CCS CONCEPTS

• **Computing methodologies** → **Multi-task learning**.

KEYWORDS

vision-and-language; deep learning; knowledge-enhanced learning; multimodal pretraining

ACM Reference Format:
Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, Jun Yu. 2021. ROSITA: Enhancing Vision-and-Language Semantic Alignments via Cross- and Intra-modal Knowledge Integration. In *Proceedings of the 29th ACM International Conference on Multimedia (MM'21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475251

*Zhou Yu is the corresponding author.
Work was done when Yuhao Cui was an intern at Alibaba Group.

MM'21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475251

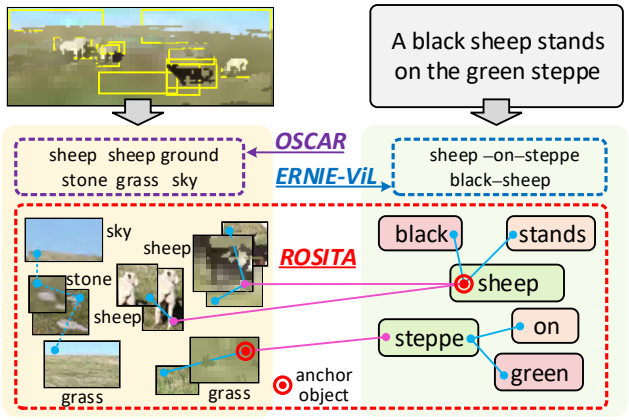

Figure 1: Schematic of the knowledge integration strategies of three VLP methods, *i.e.*, OSCAR [23], ERNIE-ViL [45], and our ROSITA. OSCAR and ERNIE-ViL only exploit the intra-modal knowledge from the image and text modalities, respectively. In contrast, ROSITA simultaneously encodes the cross-modal knowledge (pink line) and intra-modal knowledge (blue line) in a unified scene graph centered at specific anchor objects, which is used to enhance the learning of fine-grained semantic alignments across modalities.

1 INTRODUCTION

Motivated by the success of the *pretrain-then-finetune* paradigm of BERT in natural language understanding [8], there has been an increasing interest in developing vision-and-language pretraining (VLP) models [7, 23, 27, 38] to address a wide range of vision-and-language (V+L) tasks. In particular, these approaches first pretrain transformer-based models on large image-text corpus to learn task-agnostic representations, and then finetune the models on downstream V+L tasks, *e.g.*, visual question answering [50, 51], image text retrieval [18, 31], and referring expression comprehension [16, 52]. Compared to earlier methods that are only adapted to one V+L task [46, 47, 49], VLP models is generalizable across multiple tasks and also achieves significantly better performance on respective tasks.

Learning *fine-grained* semantic alignments between image regions and text words plays a key role in V+L tasks. However, manually annotating such dense alignment between regions and words is expensive and is unrealistic under the large-scale scenario. Therefore, most existing VLP approaches [7, 21, 27] use a weakly-supervised learning strategy to model the alignments implicitly. Taking the image regions and text words as inputs, they adopt multi-layer Transformers [39] as their backbones to learn *fine-grained*

semantic alignments from *coarse-grained* image-text matching supervision. Moreover, the interference within and across modalities makes the learning of semantic alignments even more challenging.

To facilitate the learning of semantic alignments, two recent VLP approaches OSCAR [23] and ERNIE-ViL [45] introduce extra knowledge in different ways. Specifically, OSCAR additionally extracts the predicted region tags from images and uses these tags as anchor points to align with text words implicitly. ERNIE-ViL explicitly constructs a scene graph from text and puts more emphasis on the keywords (*e.g.*, objects along with their attributes and relations) in the scene graph in its pretraining objectives. In terms of knowledge source, both of them use the *intra-modal* knowledge from a single modality to enhance the semantic alignments: OSCAR models the intra-modal knowledge in the image modality while ERNIE-ViL models the intra-modal knowledge in the text modality. The success of the two methods above raises a question: *Is it possible to utilize the intra-modal knowledge from both modalities along with the cross-modal knowledge to further enhance the semantic alignments?*

In this paper, we present a new VLP method called ROSITA, which encodes the c**ROS**s- and **I**n**T**r**A**-modal knowledge simultaneously in a unified scene graph. As shown in Figure 1, the graph consists of a set of knowledge entries, where each entry corresponds to an *anchor object* along with its associated cross- and intra-modal knowledge. The intra-modal knowledge refers to the relationships between the anchor object and its intra-modal contexts (*e.g.*, spatially related regions or contextually related words). The cross-modal knowledge corresponds to the relationships between the anchor object and its semantically similar objects from the opposite modality (*e.g.*, the region predicted as "*grass*" and the word "*steppe*").

Although we have obtained a set of knowledge entries, how to effectively use them to enhance semantic alignments is still nontrivial. We propose a novel *structural knowledge masking* (SKM) strategy that can be seamlessly integrated with the masked language (region) modeling tasks, which are commonly used in existing VLP methods [7, 27]. In principle, SKM determinately masks the anchor object while selectively masking its cross- and intra-modal contexts in a knowledge entry. This strategy effectively eliminates the interference information within and across modalities and enhances the semantic alignments by enforcing the model to acquire accurate information from the *opposite* modality.

The contributions of this work are three-fold:

(1) We present a new VLP method ROSITA, which incorporates cross- and intra-modal knowledge simultaneously to enhance the semantic alignments across different modalities.
(2) We introduce a novel structural knowledge masking strategy to use the scene graph structure as a priori to be integrated with the commonly used masked language (region) modeling tasks in existing VLP methods.
(3) We achieve the best results on three typical V+L tasks over six benchmark datasets, outperforming existing state-of-the-art VLP methods.

2 RELATED WORK

We briefly review previous studies on unimodal pretraining and vision-and-language pretraining, especially those studies on knowledge enhanced pretraining.

Unimodal Pretraining. The pretraining technique has been widely used in computer vision (CV) tasks. Deep convolutional neural networks like VGGNet [35] or ResNet [12] pretrained on ImageNet can well generalize to various downstream tasks [11, 26, 33]. In contrast to CV tasks, the popularization of pretraining in the natural language processing (NLP) community is relatively late. Based on the multi-layer Transformer architecture [39], many famous pretraining approaches (*e.g.*, BERT [8], GPT [32], and XLNet [42]) have been put forward. Different from the supervised pretraining paradigm in CV tasks, the pretraining paradigm in NLP tasks is *self-supervised* that aims to train a model to predict words based on their contexts without introducing human annotations. In particular, BERT introduces a novel masking language modeling (MLM) task that randomly masks the input words and predicts these masked words based on their contexts. This MLM strategy is naturally inherited by the VLP methods.

Vision-and-Language Pretraining (VLP). Different from the purely self-supervised paradigm in NLP tasks, VLP models are pretrained on large-scale paired image-text corpus, *e.g.*, image captioning datasets like [6, 29, 34]. Mirroring the success of BERT, recent studies naturally extend its framework to the vision-and-language domain to pretrain VLP models for a wide range of V+L tasks [7, 13, 23, 27, 38, 45, 53]. ViLBERT [27] and LXMERT [38] are two pioneering works in this field, where the two-stream architectures are adopted to encode the image features and textual features with two separate Transformers and then perform multimodal fusion via a third Transformer. Recent works tend to use the single-stream architectures, where the multimodal features are directly fused using one Transformer [7, 19, 21, 36]. Moreover, other techniques like knowledge integration [23, 45], multilingual enhancement [55], contrastive learning [22], and adversarial training [9] are introduced to further improve the performance of the pretrained models.

Knowledge-Enhanced Pretraining. Incorporating prior knowledge (*e.g.*, external knowledge graph) to enhance model pretraining has been investigated earlier by two ERNIE methods [37, 54] and widely explored in recent years [25, 40, 41]. The introduced prior knowledge enables the model to better understand the syntactic and semantic structure of the text, thus facilitating model pretraining by an improved structural MLM task. In the VLP task, prior knowledge can be acquired from both the image and text modalities. ERNIE-ViL constructs a scene graph from text and puts more emphasis on the discovered keywords [45]. OSCAR exploits the predicted tags of image regions to enhance the semantic alignment across the two modalities [23]. A concurrent work UC2 utilizes off-the-shelf machine translation model to construct aligned multilingual dataset for texts and regard this extra information as prior knowledge to enhance the learning of cross-modal semantic alignment [55]. Despite the success of these knowledge-enhanced VLP methods, they only utilize the intra-modal knowledge from a single modality, which restricts their effectiveness in learning semantic alignments.

To the best of our knowledge, our ROSITA is the first VLP method to integrate the cross-modal and intra-modal knowledge simultaneously in order to enhance the learning of semantic alignments across different modalities.

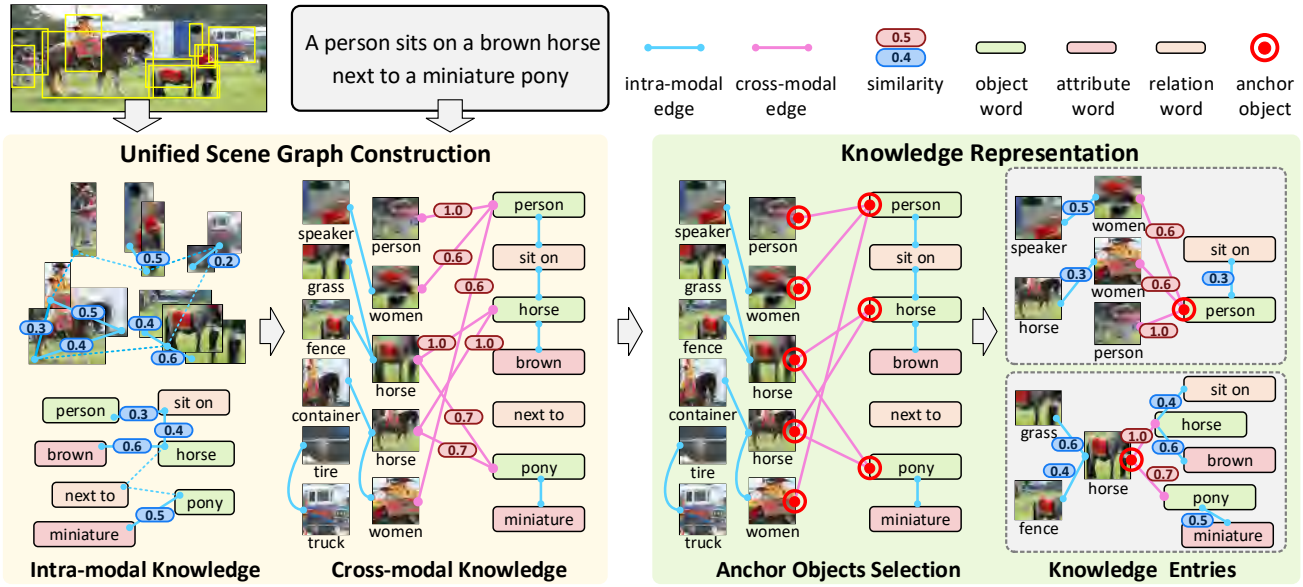

Figure 2: The flowchart of knowledge extraction given an image-text pair. It consists of two main stages, namely the unified scene graph construction and knowledge representation.

3 KNOWLEDGE EXTRACTION

In this section, we introduce the procedure of extracting knowledge entries from an image-text pair. We first construct a unified graph to model the intra- and cross-modal knowledge from an image-text pair. On top of the established graph, we select anchor objects to obtain a set of knowledge entries. The process of knowledge extraction is illustrated in Figure 2.

Unified Scene Graph Construction. Given an image-text pair, we resort to a unified scene graph structure $G =< V, E, S >$ to encode its intra- and cross-modal knowledge simultaneously [48]. The vertex set V includes the words and regions from the text and image, respectively. The edge set E and similarity set S contain pairwise relationships and their corresponding similarities between vertices (*i.e.*, edge weights), respectively.

The intra-modal knowledge within the image and text are first represented as an image scene graph and a text scene graph, respectively. For the image scene graph, regions extracted from a pretrained object detector are considered as the vertices in V. Inspired by [14, 20, 43], we calculate the similarity between each paired regions by their Intersection over Union (IoU) score. The region pairs with IoU scores larger than zero are considered to have edges in E and their IoU scores are regarded as their similarities in S. For the text graph, we use an off-the-shelf scene graph parser provided by [1] to obtain a text scene graph from a text. The text scene graph explicitly encodes the keywords of objects, attributes, and relations found in the text while discarding the rest of uninformative words. These mentioned keywords in the scene graph are regarded as the vertices in V. The word-word relationships in the scene graph (*i.e.*, object-attribute or object-relation) correspond to the edges in E. The similarity between two vertices is the co-occurrence frequencies of the referred object-attribute (or object-relation) pair calculated on the whole dataset. Since the similarity distributions of the image and text modalities may vary widely, we normalize the similarities within each modality, respectively.

As we have modeled the intra-modal knowledge in the graph, we further integrate cross-modal knowledge to align the image regions to their semantically related words. Since such cross-modal alignment supervision is not available, we establish *pseudo* semantic alignments between region-word pairs as follows. For the image regions, the predicted region tags are aligned to the object words with respect to their semantic similarities on words. We adopt a pretrained word embedding model [30] to calculate the pairwise similarities between object tags and object words[1]. We set a minimum confidence threshold of 0.5 to the similarity scores to make a trade-off between precision and recall. The region-word pairs surpass the threshold will form cross-modal edges in E and their corresponding scores represent the similarities in S.

Knowledge Representation. Based on the constructed unified scene graph G, we illustrate the procedure of extracting knowledge entries from the scene graph in detail. Note that each knowledge entry is associated with an anchor object, we first select all possible anchor objects from the graph. We define an anchor object as the vertex (an image region or a text word) in the graph that is referred to by at least one cross-modal edge. Since the attribute and relation words are not directly connected to any image region, they cannot be anchor objects according to our definition.

After obtaining the anchor objects, we integrate the intra-modal knowledge and cross-modal knowledge in G to obtain a knowledge entry. Given an anchor object $v \in V$, its corresponding knowledge entry is represented as a subgraph $g(v) \subseteq G$ and is obtained by the

[1]We have tried to establish more fine-grained alignments to include the attribute words. However, the predicted attributes from image regions are too diverse that often fail to match the attribute words in the text.

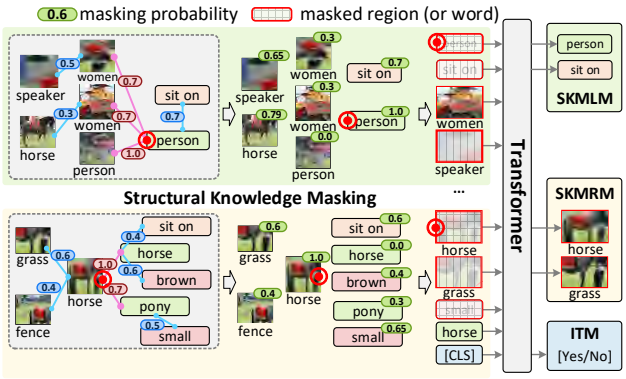

Figure 3: The flowchart of our ROSITA framework with the structural knowledge masking (SKM) strategy.

union of three subgraphs of G as follows:

$$g(v) = G_{\text{cross}}(v) \cup G_{\text{intra}}(v) \cup G_{\text{intra}}(G_{\text{cross}}(v)) \tag{1}$$

where $G_{\text{cross}}(v)$ contains the relationships between v and its directly connected contexts by cross-modal edges. $G_{\text{intra}}(v)$ models the relationships between v and its directly connected contexts by intra-modal edges. $G_{\text{intra}}(G_{\text{cross}}(v))$ includes the relationships between the vertices in $G_{\text{cross}}(v)$ and their corresponding intra-modal contexts. It is worth noting that the anchor object v can reach every vertex in $g(v)$ within two hops.

4 THE ROSITA FRAMEWORK

Based on the extracted knowledge entries from image-text pairs, we introduce the ROSITA framework in this section. We first describe the image and text feature representations and the network architecture. Then, we introduce a structural knowledge masking (SKM) strategy, which takes the knowledge entries as a priori to perform the masked language (region) modeling. Finally, we describe the whole pretraining objective with multi-task learning. The overall framework is illustrated in Figure 3.

Image and Text Feature Representations. Following the commonly used strategy in existing VLP methods [6, 23], the input image is represented as a set of regional features, which are extracted from a Faster R-CNN model pretrained on Visual Genome [2]. More specifically, we extract m regions with the highest confidence probabilities from the image. For the i-th region, it is represented as a visual feature $f_i \in \mathbb{R}^{2048}$ and a positional feature $p_i \in \mathbb{R}^5$ [45]. The two features are fused into a d-dimensional image representation $x_i \in \mathbb{R}^d$ using two linear projections as follows:

$$x_i = W_f^T f_i + W_p^T p_i \tag{2}$$

where $W_f \in \mathbb{R}^{2048\times d}$ and $W_p \in \mathbb{R}^{5\times d}$. Finally, the image is represented as a feature matrix $X \in \mathbb{R}^{m\times d}$.

For its paired text, we adopt the word processing method similar to [8]. The input text is first tokenized into words and trimmed (or padded) to a maximum of n words. Each word w_i and its index i (*i.e.*, the absolute position of w_i in the text) are projected to vectors by two individual embedding layers, then added to obtain the position-aware text representation y_i as follows:

$$y_i = \text{WordEmbed}(w_i) + \text{IdxEmbed}(i) \tag{3}$$

where y_i is d-dimensional to match the image representation. The text is finally represented as a feature matrix $Y \in \mathbb{R}^{n\times d}$.

Network Architecture. The image features $X = [x_1, ..., x_m]$ and text features $Y = [y_1, ..., y_n]$ are first concatenated before feeding to the network. We insert two special tokens to the concatenated features to obtain the multimodal input features Z:

$$Z = [x_1, x_2, ..., x_m, [\text{SEP}], y_1, y_2, ..., y_n, [\text{CLS}]] \tag{4}$$

where the [SEP] token marks the boundary between the image and text features. The [CLS] token is used to predict whether the given image and text are paired or not.

The multimodal features Z are fed into a single-stream Transformer with L layers [8]. Each layer consists of a multi-head self-attention (MSA) block and a feed-forward networks (FFN) block.

$$\begin{aligned} \hat{Z}^\ell &= \text{LN}(\text{MSA}(Z^{\ell-1}) + Z^{\ell-1}), \quad && \ell = 1, 2, ..., L \\ Z^\ell &= \text{LN}(\text{FFN}(\hat{Z}^\ell) + \hat{Z}^\ell), && \ell = 1, 2, ..., L \end{aligned} \tag{5}$$

where $Z^0 = Z$. Layer normalization [4] and residual connection [12] are applied after every block, respectively.

Structural Knowledge Masking. The masked language modeling (MLM) [8] and masked region modeling (MRM) [7] tasks are commonly used in almost all the VLP methods [7, 27, 45]. They randomly mask the input tokens (*i.e.*, words or regions) and predict these masked tokens based on their contextual tokens. Since the random masking based MLM and MRM tasks are not aware of the keywords and key regions to be aligned, their efficacy in the alignment learning is weak. To this end, we present an alternative *structural knowledge masking* (SKM) strategy to selectively mask the tokens referred to by the extracted knowledge entry. Accordingly, the MLM and MRM tasks are respectively modified to the SKMLM and SKMRM tasks to adapt to the SKM strategy.

Let an image be represented as a set of regions $\mathcal{R} = \{r_1, ..., r_m\}$ and a text be represented as a sequence of words $\mathcal{W} = \{w_1, ..., w_n\}$, we construct a unified scene graph G on top of $\mathcal{R}$ and $\mathcal{W}$, and extract a set of knowledge entries from G. Let $g(v_i) = < \hat{V}, \hat{E}, \hat{S} >$ be one of the knowledge entries, where $v_i \in \hat{V}$ is the anchor object. The vertices are represented as $\hat{V} = \{v_1, ..., v_N\}$ and the similarities between the vertices are represented as $\hat{S} \in \mathbb{R}^{N\times N}$, where N is the number of vertices in this entry.

The strategy of SKM is to determinately mask the anchor object v_i while probabilistically masking its intra-modal contexts and cross-modal contexts with respect to the graph structure of the knowledge entry. Since the similarities between v_i and its contexts are different, we assign *independent* masking probabilities to each of the contexts with respect to their similarities to v_i, rather than simply using an *identical* masking probability for all the contexts. To obtain the masking probabilities for the contexts, we introduce a masking strategy that satisfies the following principles: for the intra-modal contexts, a larger similarity score refers to a higher masking probability. For the cross-modal contexts, a larger score leads to a lower masking probability. The reasons behind this masking strategy will be explained hereinafter.

Note that not all contexts have direct connections to the anchor object. Therefore, we calculate the transmission probabilities $T = [t_1, ..., t_N] \in [0, 1]^N$ from the anchor object v_i to its contexts in

$\hat{V}$ based on the normalized similarities defined in $\hat{S}$. Since v_i can reach all the vertices in $\hat{V}$ within two hops, T is defined as follows:

$$T = \frac{1}{2}\hat{S}\pi(i) + \frac{1}{2}\hat{S}\hat{S}\pi(i) \tag{6}$$

where $\pi(i) \in \{0,1\}^N$ is a one-hot vector with the i-th element to be 1. The two terms correspond to the transmission probabilities between the anchor object v_i and its one-hop and two-hop contexts, respectively. After that, we convert the transmission probabilities T to the masking probabilities $P = [p_1, ..., p_N] \in [0,1]^N$ using the following rules to satisfy our masking strategy above:

$$p_j = \begin{cases} 1, & \text{if } v_j \text{ is the anchor object} \\ \alpha t_j, & \text{if } v_j \text{ is within the intra-modal contexts} \\ (1-\alpha)(1-t_j), & \text{if } v_j \text{ is within the cross-modal contexts} \end{cases} \tag{7}$$

where α is a hyper-parameter to balance the masking probabilities of the intra-modal and cross-modal contexts. t_j and p_j denote the transmission and masking probability of the vertex v_j, respectively.

Given a knowledge entry, we use the calculated masking probabilities to obtain two groups of mask indices M_w and M_r, indicating the words and regions to be masked, respectively. The partially masked input features are passed through the network and then fed into the SKMLM and SKMRM tasks.

In particular, if the anchor point v_i refers to an object word in the text, we resort to the SKMLM task to reconstruct the masked words $\mathcal{W}_{M_w}$ as follows:

$$\mathcal{L}_{\text{SKMLM}}(\theta) = -\mathbb{E}_{(\mathcal{W},\mathcal{R})\sim D} \log P_\theta(\mathcal{W}_{M_w} | \mathcal{W}_{\setminus M_w}, \mathcal{R}_{\setminus M_r}) \tag{8}$$

where θ is the trainable parameters. Each pair $(\mathcal{W}, \mathcal{R})$ is sampled from the whole training set D. $\mathcal{W}_{\setminus M_w}$ and $\mathcal{R}_{\setminus M_r}$ refer to the remaining words in $\mathcal{W}$ and the remaining regions in $\mathcal{R}$ with excluding the masked tokens from their modalities, respectively.

Analogously, if v_i refers to a region in the image, we resort to the SKMRM task to reconstruct the masked regions $\mathcal{R}_{M_r}$ as follows:

$$\mathcal{L}_{\text{SKMRM}}(\theta) = -\mathbb{E}_{(\mathcal{W},\mathcal{R})\sim D}\, f_\theta(\mathcal{R}_{M_r} | \mathcal{W}_{\setminus M_w}, \mathcal{R}_{\setminus M_r}) \tag{9}$$

where $f_\theta(\cdot)$ refers to some loss functions. Similar to [7], we use the regression-based loss and classification-based loss jointly.

The motivations of SKM can be explained as follows: (*i*) The intra-modal contexts may contain interference information (*e.g.*, the word "*sky*" is frequently associated with an attribute "*blue*" and a visual object of "*wheel*" is usually within an object of "*car*"). Such interference information may leak out the semantics of the anchor object and reduce the difficulty of anchor object reconstruction, leading to a degradation of the pretrained model. Therefore, when masking an anchor object, its intra-modal contexts with high similarities will have high probabilities to be masked simultaneously. This operation reduces the risk of information leakage and enforces the model to acquire precise information from the opposite modality, which *implicitly* enhance the semantic alignments; (*ii*) The cross-modal contexts with low similarities may contain irrelevant or noisy information. Therefore, when masking an anchor object, its cross-modal contexts with high similarities will have low probabilities to be masked at the same time, which *explicitly* excludes potential noise thus benefiting the semantic alignments. As a result, the synergy of the masking operations above significantly facilitates the semantic alignments.

Table 1: The detailed statistics of the used datasets. Following the strategies in [6], we split them into in-domain and out-of-domain splits based on the image sources. Each cell shows the number of image-text pairs.

	in-domain		out-of-domain		total
	COCO[6]	VG[17]	CC[34]	SBU [29]	
train	533K	5.1M	3.0M	869K	9.5M
val	25K	106K	14K	10K	155K

Multi-task Learning. Similar to [7], we adopt a multi-task learning objective to pretrain our model. Besides the proposed SKMLM and SKMRM tasks, we also include the image-text matching (ITM) task. Moreover, since the SKMLM and SKMRM tasks only focus on the key tokens included in the knowledge entry, we still retain the original random masking-based MLM and MRM tasks to guarantee a good coverage of the remaining tokens in the image and text[2].

5 EXPERIMENTS

We evaluate ROSITA on three V+L tasks and perform thorough comparative analysis to the state-of-the-art VLP methods on six datasets. Furthermore, we conduct comprehensive ablation experiments to explore its effectiveness in learning fine-grained semantic alignments.

5.1 Pretraining Setup

Datasets. Following the strategy in [7], we construct the pretraining dataset consisting of 9.5M train and 155K validation image-text pairs from four existing datasets, namely the COCO Captions [6], Visual Genome Captions[17], Conceptual Captions [34], and SBU Captions [29]. The four datasets are categorized into the *in-domain* and *out-of-domain* datasets based on whether they share the same images with the downstream tasks. The statistics of the pretraining dataset are shown in Table 1.

Implementation Details. For the input image-text pairs, we extract a fixed number of 36 region features from a pre-trained Faster R-CNN model [2] and adopt the BPE strategy to tokenize the sentence into a maximum of 50 words following [8]. Our ROSITA model adopts a 12-layer Transformer encoder architecture with 768 hidden units and 12 attention heads. The hyper-parameter α in Eq.(7) is set to 0.9. The masking probabilities in the original MRM and MLM tasks are set to 15% [45]. The model is initialized with the parameters from a pretrained BERT-base model [8], and then trained up to 40 epochs with a batch size of 512.

5.2 Downstream Tasks

After obtaining the pretrained ROSITA model, we finetune it on three downstream V+L tasks as follows.

Visual Question Answering (VQA) is a task that requires the model to answer natural language questions about an image. We adopt the widely used VQAv2 dataset [3, 10], which is manually built on the images from the MSCOCO dataset [24]. The dataset is split

[2]We have made such an experiment that removes the MRM & MLM tasks. The resulting model reports slight performance drop (~0.3 points) on the downstream tasks.

Table 2: Results on downstream V+L tasks to compare with the state-of-the-art VLP methods. For a fair comparison, all the results are archived by the base models. Most of the models are trained on the *in-domain+out-of-domain* datasets, except for those models marked with † are trained on the *out-of-domain* datasets. IR and TR denote the image retrieval and text Retrieval, respectively. For the REC task, all the results are achieved based on the detected region features from images. Dark and light grey colors highlight the top and second best results on each evaluation metric.

task	dataset		ViLBERT† [27]	VLBERT† [36]	Unicoder-VL [19]	LXMERT [38]	UNITER [7]	ERNIE-ViL† [45]	VILLA [9]	OSCAR [23]	ROSITA (ours)
VQA	VQAv2	test-dev	70.55	71.16	-	72.42	72.70	72.62	73.59	73.16	73.91
		test-std	70.92	-	-	72.54	72.91	72.85	73.67	73.44	73.97
REC	Ref-COCO	val^d	-	-	-	-	81.24	-	81.65	-	84.33
		testA^d	-	-	-	-	86.48	-	87.40	-	87.52
		testB^d	-	-	-	-	73.94	-	74.48	-	77.98
	Ref-COCO+	val^d	72.34	71.60	-	-	75.31	74.02	76.05	-	76.06
		testA^d	78.52	77.72	-	-	81.30	80.33	81.65	-	82.01
		testB^d	62.61	60.99	-	-	65.68	64.74	65.70	-	67.40
	Ref-COCOg	val^d	-	-	-	-	74.31	-	75.90	-	77.82
		test^d	-	-	-	-	74.51	-	75.93	-	77.64
ITR	IR-COCO	R@1	-	-	46.70	-	50.33	-	-	54.00	54.40
		R@5	-	-	76.00	-	78.52	-	-	80.80	80.92
		R@10	-	-	85.30	-	87.16	-	-	88.50	88.60
	TR-COCO	R@1	-	-	62.30	-	64.40	-	-	70.00	71.26
		R@5	-	-	87.10	-	87.40	-	-	91.10	91.62
		R@10	-	-	92.80	-	93.08	-	-	95.50	95.58
	IR-Flickr	R@1	58.20	-	71.50	-	72.52	74.44	74.74	-	74.08
		R@5	84.90	-	90.90	-	92.36	92.72	92.86	-	92.44
		R@10	91.52	-	94.90	-	96.08	95.94	95.82	-	96.08
	TR-Flickr	R@1	-	-	86.20	-	85.90	86.70	86.60	-	88.90
		R@5	-	-	96.30	-	97.10	97.80	97.90	-	98.10
		R@10	-	-	99.00	-	98.80	99.00	99.20	-	99.30

into train (83k images and 444k questions), validation (41k images and 214k questions), and test (81k images and 448k questions) sets. Following the strategy in [7], we feed the representation of the [CLS] token to a linear classifier to predict the corresponding answer from a vocabulary of size 3129 [49].

Referring Expression Comprehension (REC) is a task that requires to localize an image region referred to by a natural language query. We evaluate the performance on RefCOCO [16], RefCOCO+ [16] and RefCOCOg [28] datasets. All the three datasets are collected from COCO images [31]. RefCOCO and RefCOCO+ are split into four subsets, including train (120k queries), validation (11k queries), testA (6k queries about people), and testB (6k queries about objects), while RefCOCOg is split into three subsets, including train (81k queries), validation (5k queries), and test (10k queries). The representation for each image region is used to predict a ranking score and a refined bounding box.

Image-Text Retrieval (ITR) is a task that requires the model to calculate a similarity score between an image and a sentence and then perform cross-modal retrieval. We conduct experiments on the COCO Captions [6] and Flickr30K [44] datasets, respectively. Following the partition strategy by [15], the COCO dataset is split into 82k/5k/5k train/validation/test images, while the Flickr30K dataset is split into 29k/1k/1k train/validation/test images. Similar to [7], we use an offline hard sample mining strategy to obtain 128 negative samples per each positive sample, and use the representation of the [CLS] token to predict a matching score.

5.3 Main Results

We compare the proposed ROSITA model against existing state-of-the-art VLP methods. As shown in Table 2, ROSITA achieves the overall best performance on all downstream tasks, which verifies the effectiveness of the integrated cross- and intra-modal knowledge and the corresponding SKM strategy[3].

It is worth noting that some methods like ViLBERT, LXMERT, and ERNIE-ViL adopt the two-stream architecture, which have much more parameters (ROSITA: 116M, VilBERT: 221M, LXMERT: 183M, ERNIE-ViL: 228M). Some methods like UNITER and VILLA use a larger number of image features (up to 100 regions), which has been verified to benefit the performance at the expense of much higher computational cost. In contrast, ROSITA uses a fixed number

[3]We have conduct such an experiment that pretrains ROSITA on the *out-of-domain* datasets only. The resulting model consistently outperforms the counterparts [27, 36, 45], verifying the generalization capability of our approach.

Table 3: Ablations of ROSITA variants without the cross- and intra-modal knowledge. All models are pretrained on the *in-domain* datasets and then finetuned on specific downstream tasks. For each model, we report the accuracies on the pretraining tasks an downstream tasks, respectively. As we only have positive image-text pairs in the pretraining datasets, we use the offline hard sample mining strategy to generate an equal number of negative samples for the evaluation of the ITM task.

#	model	pretraining tasks			downstream tasks			
		ITM	SKMLM	SKMRM	VQAv2 (dev)	RefCOCO (val)	IR-Flickr (test)	TR-Flickr (test)
1	ROSITA (full)	**84.34**	**67.16**	**76.50**	**73.19**	**84.22**	**85.09**	**94.33**
2	-w/o cross-modal knowledge	83.54	63.69	72.56	72.86	83.85	84.23	93.63
3	-w/o intra-modal knowledge	83.30	63.75	73.90	72.98	83.31	84.79	93.90
4	-w/o both types of knowledge	82.22	61.19	68.58	72.47	82.12	82.11	92.57

Table 4: Ablations of four ROSITA variants with two alternative masking strategy in SKM (*i.e.*, independent probabilities and identical probability). All models are pretrained on the *in-domain* datasets and finetuned on the downstream tasks.

masking prob.	VQAv2 (dev)	RefCOCO (val)	IR-Flickr (test)	TR-Flickr (test)
independent	**73.19**	**84.22**	**85.09**	**94.33**
identical (p=45%)	72.79	83.18	83.70	93.20
identical (p=30%)	72.93	83.29	84.36	93.63
identical (p=15%)	72.75	82.96	83.75	93.53

of 36 image features. We believe the performance of ROSITA can be further improved by taking these advanced strategies above.

5.4 Ablation Studies

We run a number of ablations to investigate the reasons of ROSITA's effectiveness. The results show in Table 3-4 and Figure 4-5 are discussed in detail below.

Cross- and Intra-modal Knowledge. In Table 3, we show the effects of the intra-modal knowledge and cross-modal knowledge based on the performance on the pretraining and downstream tasks. Taking the full ROSITA as the reference model (Line #1), we obtain the different variants by removing the cross-modal knowledge or the intra-modal knowledge. The variant without cross-modal knowledge (Line #2) indicates that the model is not aware of the anchor objects and the SKM strategy is performed only on a single modality using the intra-modal knowledge. In contrast, the variant without intra-modal knowledge (Line #3) indicates that the model is aware of the anchor objects but is not aware of the intra-modal contexts. Finally, by removing both the cross- and intra-modal knowledge, we obtain a baseline variant nearly identical to UNITER [7] (Line #4)[4].

Given the pretrained models of the four variants above (*i.e.*, without finetuning on downstream tasks), we evaluate their performance on three pretraining tasks. The ITM task examines the ability of semantic alignment between image-text pairs. From the results, we can see that both types of knowledge bring performance improvement to the ITM task (#4 *vs.* #3 and #2). Moreover, the two types of knowledge are complementary that their synergy brings 2.1 points improvement compared to the baseline model without any knowledge (#4 *vs.* #1). Although the ITM task is the most straightforward metric for semantic alignment, it only measures the *coarse-grained* alignments on the image-text level, thus cannot fully reveal the capability of ROSITA. As a complement, we resort to the SKMLM and SKMRM tasks to evaluate the *fine-grained* alignments on the region-word level. Compared with the baseline model in #4, the full ROSITA model improves the accuracies by 7.0 and 7.9 points on the SKMLM and SKMRM tasks, respectively.

Next, we report the performance of these variants on different downstream tasks. From the demonstrated results, we obtain similar observations to those on the pretraining tasks. The full ROSITA model consistently outperforms all the counterparts, verifying the effectiveness of the cross- and intra-modal knowledge.

SKM Strategy. After extracting knowledge entries from image-text pairs, we have two alternative masking strategies in SKM, *i.e.*, the independent probabilities and the identical probability. For the masking strategy with identical probability, we evaluate the choices of different probabilities within {15%, 30%, 45%}. The results in Table 4 show that the model pretrained with independent probabilities steadily outperforms all the counterparts with the identical probability. For the models pretrained with the identical probability strategy, their performance is sensitive to the choices of the predefined probability. A small masking probability (*e.g.*, 15%) may degrade the model towards the baseline without any knowledge. A large masking probability (*e.g.*, 45%) may shield the essential information that is necessary to learn the semantic alignments. In comparison, the masking strategy with independent probabilities provides a more fine-grained understanding of the knowledge structure, leading to a more robust pretrained model.

Cross-modal Semantic Alignments. The effect of *fine-grained* semantic alignments across modalities can be inferred from the attention maps of the learned Transformer model [5]. We visualize the learned *cross-modal attentions* (*i.e.*, region-to-words and word-to-regions attentions) from the pretrained UNITER [7] and our ROSITA models, as shown in Figure 4. Taking the image-text pair as inputs with exactly one token (a region or a word) being masked at a time, we pass the multimodal features through the pretrained model and extract the attention map from the last MSA block[5]. The region-to-words and word-to-regions attentions of the masked

[4]Our model has slight performance deviations compared with the original UNITER model since we use different visual features and pretraining hyper-parameters.

[5]We perform element-wise addition over the attention maps from different heads followed by row-wise softmax normalization to obtain one aggregated attention map.

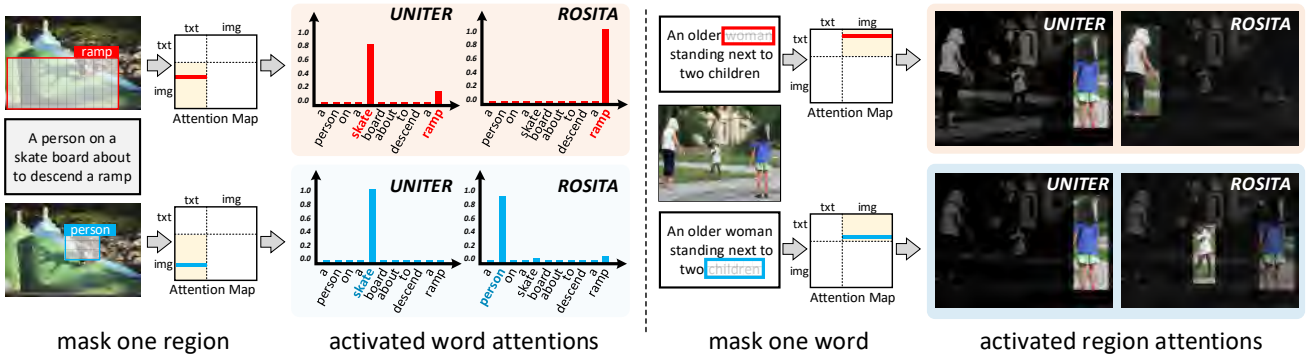

Figure 4: Visualizations of the learned cross-modal attentions (*i.e.*, region-to-words attentions on the left and word-to-regions on the right) from UNITER [7] and ROSITA. Taking the image-text pair as inputs with exactly one region (or word) being masked at a time, we extract the attention map from the last MSA block of the pretrained model. The region-to-words (word-to-regions) attentions correspond to one specific row in the bottom-left (top-right) area of the attention map, respectively.

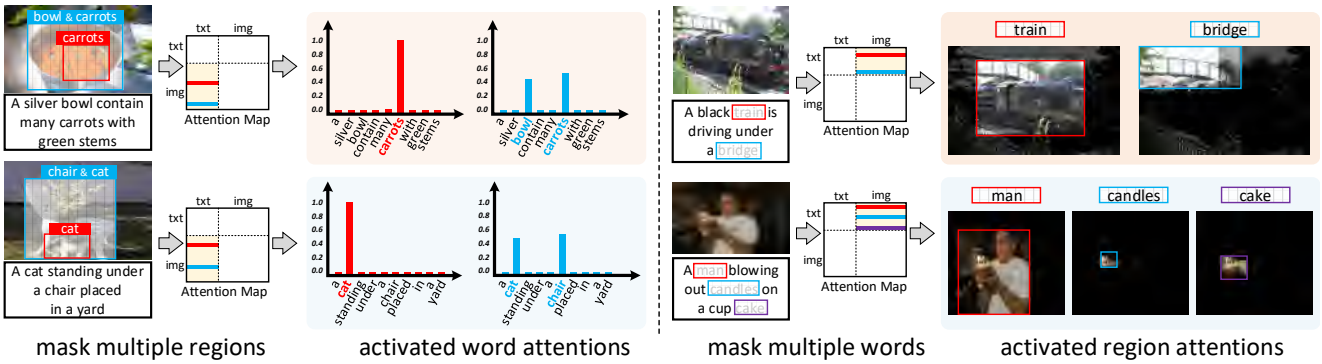

Figure 5: Visualizations of the region-to-words attentions (left) and word-to-regions attentions (right) from a pretrained ROSITA model with masking multiple regions (or words) at one time.

token correspond to one specific row in the bottom-left and top-right area of the attention map, respectively.

From the visualized cross-modal attentions, we can see that ROSITA learns significantly better semantic alignments than UNITER. ROSITA can precisely align the masked object to its reference object in the opposite modality while UNITER fails to establish such cross-modal alignments. For example, when the region of "*ramp*" is masked, ROSITA activates the word "*ramp*" precisely while UNITER obtains the largest attention value on the word "*skate*". When another region of "*person*" is masked, ROSITA precisely activates the word "*person*" while UNITER still activates the incorrect word "*skate*". Similar phenomena are observed in the opposite direction. ROSITA activates the accurate regions to the masked words while UNITER fails to do it.

To step further, we conduct a more challenging task as follows. We mask *multiple* regions (or words) at the same time to examine whether the semantic alignments can still be achieved. The visualized results in Figure 5 show that ROSITA works surprisingly well to establish accurate semantic alignment for each masked token. For example, when the regions of "*bowl*" and "*bowl & carrots*" are masked simultaneously, the region of "*carrot*" is precisely aligned to the word "*carrots*", and the region of "*bowl & carrots*" is aligned to the two words "*bowl*" and "*carrots*" uniformly. In the opposite direction, when the words "*man*", "*candles*", and "*cake*" are masked at the same time, their corresponding regions are highlighted in the learned attentions, respectively.

6 CONCLUSION

In this paper, we present a new VLP method called ROSITA, which integrates the cross- and intra-modal knowledge in a unified scene graph to enhance the learning of cross-modal semantic alignment. We introduce a novel structural knowledge masking (SKM) strategy to perform masked language (region) modeling with respect to the knowledge entries extracted from the unified scene graph. Extensive ablations, comparative experiments, and comprehensive analysis show that ROSITA significantly outperforms existing state-of-the-art VLP approaches on three typical V+L tasks over six benchmark datasets. We hope our study will be helpful to inspire future research in the vision-and-language community and beyond.

ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0100603, and in part by National Natural Science Foundation of China under Grant 62072147, Grant 61836002, and Grant 62020106007.

REFERENCES

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*. Springer, 382–398.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6077–6086.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[5] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*. Springer, 565–580.

[6] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*. Springer, 104–120.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. 4171–4186.

[9] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *Advances in Neural Information Processing Systems*.

[10] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6904–6913.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*. 2961–2969.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[13] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849* (2020).

[14] Yash Kant, Dhruv Batra, Peter Anderson, Alexander Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. 2020. Spatially aware multimodal transformers for textvqa. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 715–732.

[15] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 3128–3137.

[16] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 787–798.

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[18] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.

[19] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11336–11344.

[20] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 10313–10322.

[21] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[22] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *ACL* (2021).

[23] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*. Springer, 121–137.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, 740–755.

[25] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 2901–2908.

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3431–3440.

[27] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*. 13–23.

[28] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11–20.

[29] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in Neural Information Processing Systems* 24 (2011), 1143–1151.

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[31] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*. 2641–2649.

[32] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).

[33] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2017), 1137–1149.

[34] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2556–2565.

[35] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

[36] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

[37] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223* (2019).

[38] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5103–5114.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6000–6010.

[40] Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Cuihong Cao, Daxin Jiang, Ming Zhou, et al. 2020. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808* (2020).

[41] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation. *Transactions of the Association for Computational Linguistics* 9 (2021), 176–194.

[42] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing System*. 5754–5764.

[43] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*. 684–699.

[44] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.

[45] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[46] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1307–1315.

[47] Zhou Yu, Yuhao Cui, Jun Yu, Meng Wang, Dacheng Tao, and Qi Tian. 2020. Deep Multimodal Neural Architecture Search. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3743–3752.

[48] Zhou Yu, Fei Wu, Yi Yang, Qi Tian, Jiebo Luo, and Yueting Zhuang. 2014. Discriminative coupled dictionary hashing for fast cross-media retrieval. In *Proceedings of the 37th international ACM SIGIR Conference on Research & development in Information Retrieval*. 395–404.

[49] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6281–6290.

[50] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering. *IEEE International Conference on Computer Vision (ICCV)* (2017), 1839–1848.

[51] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems* 29, 12 (2018), 5947–5959.

[52] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. 2018. Rethinking Diversified and Discriminative Proposal Generation for Visual Grounding. *International Joint Conference on Artificial Intelligence (IJCAI)* (2018).

[53] Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. 2020. DeVLBert: Learning Deconfounded Visio-Linguistic Representations. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4373–4382.

[54] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1441–1451.

[55] Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. UC2: Universal Cross-lingual Cross-modal Vision-and-Language Pre-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4155–4165.