RelCLIP: Adapting Language-Image Pretraining for Visual Relationship Detection via Relational Contrastive Learning

Yi Zhu[1]*, **Zhaoqing Zhu**[2]*, **Bingqian Lin**[3], **Xiaodan Liang**[3], **Feng Zhao**[2]†, **Jianzhuang Liu**[1]
[1]Huawei Noah's Ark Lab, [2]University of Science and Technology of China,
[3]Sun Yat-sen University

zhu.yee@outlook.com, zhaoqingzhu@mail.ustc.edu.cn, bingqianlin@126.com,
xdliang328@gmail.com, fzhao956@ustc.edu.cn, liu.jianzhuang@huawei.com

Abstract

Conventional visual relationship detection models only use the numeric ids of relation labels for training, but ignore the semantic correlation between the labels, which leads to severe training biases and harms the generalization ability of representations. In this paper, we introduce compact language information of relation labels for regularizing the representation learning of visual relations. Specifically, we propose a simple yet effective visual **Rel**ationship prediction framework that transfers natural language knowledge learned from **C**ontrastive **L**anguage-**I**mage **P**re-training (CLIP) models to enhance the relationship prediction, termed as **RelCLIP**. Benefiting from the powerful visual-semantic alignment ability of CLIP at image level, we introduce a novel Relational Contrastive Learning (RCL) approach that explores relation-level visual-semantic alignment via learning to match cross-modal relational embeddings. By collaboratively learning the semantic coherence and discrepancy from relation triplets, the model can generate more discriminative and robust representations. Experimental results on the Visual Genome dataset show that RelCLIP achieves significant improvements over strong baselines under full (providing accurate labels) and distant supervision (providing noise labels), demonstrating its powerful generalization ability in learning relationship representations. Code will be available at https://gitee.com/mindspore/models/tree/master/research/cv/RelCLIP.

1 Introduction

Visual relationship detection needs to predict the relation label between a pair of localized objects (e.g. "*man carrying bag*"). Based on such relationships we can construct a structural representation (i.e., scene graph) that regards the visual concepts within a scene as a whole and could benefit many downstream reasoning tasks, such as image retrieval (Qi et al., 2017; Johnson et al., 2015), vision question answering (Antol et al., 2015), and visual commonsense reasoning (Zellers et al., 2019).

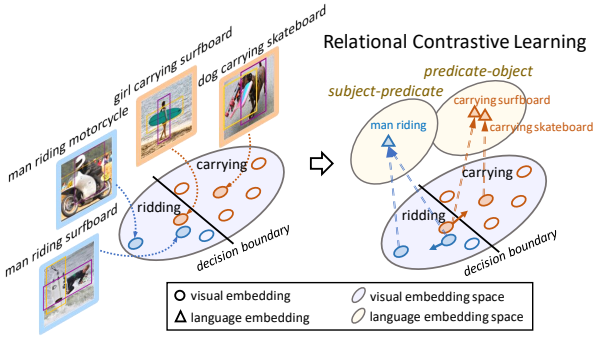

Figure 1: Overview of RelCLIP which adapts Contrastive Language-Image Pretraining (CLIP) models to enhance the visual-semantic alignment for relationship learning via a novel Relational Contrastive Learning (RCL) approach. RCL matches the cross-modal relationship embeddings in both *subject-predicate* and *predicate-object* levels and improves relationship discrimination with the help of language information of relation labels.

Existing visual relationship detection models (Yao et al., 2021; Guo et al., 2021; Chen et al., 2019; Zellers et al., 2018) learn relation representations based on the visual information of object pairs under the supervision of numeric ids of relation labels. The learned representations often suffer from the highly changeful visual appearances of the instances from the same relation category as well as the fine-detailed visual difference between relation classes. As shown in Fig. 1, the representation of the sample "girl carrying surfboard" is very close to that of "man riding surfboard" in the visual embedding space, which may cause undesirable misclassification of relations.

To alleviate this, we introduce language information of relation labels which is more compact than visual information to capture the semantic correlations among the labels, and regularize the representation learning of visual relationships via exploring visual-semantic alignment at relation level. By do-

* Equal contribution.
† Corresponding author.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4800–4810
December 7-11, 2022

ing this, we can learn more robust and discriminative relationship representations. Specifically, we propose a **Rel**ationship detection framework that adapts **C**ontrastive **L**anguage-**I**mage **P**retraining (CLIP) models (Radford et al., 2021) to enhance the visual-semantic alignment during relationship learning, termed as **RelCLIP**. The CLIP model is pre-trained on 400 million image-text pairs harvested from the Web, and is proved to be good at image-level visual-semantic alignment. During the representation learning of CLIP, the objects within an image are considered to be independent, and it is difficult to capture the interaction between objects, e.g., visual relationships, so CLIP models can not be directly applied to visual relation detection. For example, as shown in Fig. 2, CLIP features only focus on individual objects but fail to activate the area that indicates the relationship between two objects. In contrast, our RelCLIP can effectively focus on the interactive regions between objects that intuitively indicates the relations.

To achieve relation-level visual-semantic alignment, we develop a novel Relational Contrastive Learning (RCL) approach to match the visual and language embeddings of relationships, since naively using existing contrastive learning methods (Chen and He, 2021; He et al., 2020; Xie et al., 2021; Li et al., 2020) for relationship triplets (e.g. *subject-predicate-object*) may cause two severe issues. First, directly comparing the triplets may lead to trivial comparisons among all possible combinations of labels of subject, object and predicate, which is quite inefficient and greatly harms the robustness of the model. Second, the synonymous relationships (e.g., "*man riding bike*" and "*person riding bike*") and the less informative negative ones (e.g. "*man riding bike*" and "*bird sitting on branch*") may be contrasted inappropriately and inefficiently. To address these issues, our RCL 1) decouples the triplet level comparison into dual contrastive objectives which compare the same relation instance in *subject-predicate* and *predicate-object* levels to reduce the amount of comparisons; 2) adopts a new semantic-aware active sampling strategy that excludes synonymous relationships and includes informative negative samples according to their semantic meaning.

Extensive experiments on the Visual Genome dataset show that RelCLIP achieves significant improvements over strong baselines trained using whether human-annotated labels from full supervision or noisy labels from distant supervision, demonstrating its powerful generalization ability in learning relationship representations.

Figure 2: Grad-CAM visualization of relational activation. Cross-modal pretraining models (e.g., CLIP) are powerful in image-level visual-semantic alignment but fail to capture the relationship between objects. In contrast, RelCLIP can activate the interaction area between objects (e.g., "riding" and "sitting on") and achieve relation-level visual-semantic alignment.

2 Related Work

Visual Relationship Detection. Visual Relationship Detection (Li et al., 2017; Zellers et al., 2018; Lu et al., 2016; Tang et al., 2019) has raised wide concern in the computer vision community for its potential benefits that would be brought to downstream visual reasoning tasks (Johnson et al., 2018; Yang et al., 2019; Shi et al., 2019). Early works tend to detect objects and pairwise relationships independently, which overlooks the rich visual context and may lead to sub-optimal performance (Lu et al., 2016; Zhuang et al., 2017; Zhang et al., 2017; Zhu and Jiang, 2018). Later, many works have explored the message passing for context propagation and feature refinement (Xu et al., 2017; Zellers et al., 2018; Dai et al., 2017). At the same time, some works also have noticed some connections between objects and pairwise relationships, and kept their cooperative relationship (Zhang et al., 2019; Li et al., 2022). More recently, Yao *et al.* (Yao et al., 2021) explores a novel visual distant supervision that retrieves possible relation labels from commonsense knowledge bases for object pairs. Nevertheless, most visual relationship detection works typically trained the relationship prediction models under the supervision of numeric ids of predicate labels, which usually suffer from the highly diverse visual object appearances of the same predicate. To break the limitation for better relation detection, we resort to the learned natural language knowledge

in cross-modal pre-training models for capturing rich semantic dependencies among predicates and objects in this paper.

Cross-modal Pre-training. Recently, there has been a surging interest in employing cross-modal pre-training (Su et al., 2020; Chen et al., 2020; Li et al., 2020, 2019; Tan and Bansal, 2019; Lu et al., 2019) for improving the performance of downstream tasks, such as Visual Question Answering (VQA) (Antol et al., 2015), Visual Commonsense Reasoning (VCR) (Zellers et al., 2019), and Referring Expression Comprehension (Hamilton et al., 2017). LXMERT (Tan and Bansal, 2019) and ViLBERT (Lu et al., 2019) are two pioneering works, which rely on two Transformers to encode image and text modalities with a third Transformer built on top for multi-modal fusion. Unlike this two-stream architecture, single-stream architectures where two modalities are directly fused in the early stage are further designed in some recent works, such as VL-BERT (Su et al., 2020), VisualBERT (Li et al., 2019), UNITER (Chen et al., 2020) and Unicoder-VL (Li et al., 2020). Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021) is a recently proposed cross-modal pretrained model that uses 400 million image-text pairs collected from the web. It is shown to have the outstanding ability for object-level visual-semantic alignment and improving numerous downstream tasks. Powered by this, our method adapts the CLIP model to enhance the relation-level visual-semantic alignment for learning more discriminative relationship representations.

3 Methodology

3.1 Preliminaries

Problem Setup. Given a scene image I, visual relationship detection models detect a set of visual relationships in the form of a triple token $\langle a_i, a^r_{ij}, a_j \rangle$, where $a_i, a_j \in \mathcal{A}$ are the class labels of two objects i and j localized by bounding boxes $\mathbf{b}_i$ and $\mathbf{b}_j$, and $a^r_{ij} \in \mathcal{A}^r$ is the class label of the predicate that connects the object pair. For each image I, we employ the Faster R-CNN framework (Ren et al., 2015) to extract a set of M object proposals $B^o = \{\mathbf{b}\}_{i=1}^M, \mathbf{b}_i = [x_i, y_i, w_i, h_i]$ with (x_i, y_i) being the coordinate of the top left corner and w_i and h_i being the width and the height of the bounding box. We take each pair of objects as a relation proposal $\mathbf{b}_{ij} = \text{Union}(\mathbf{b}_i, \mathbf{b}_j)$ if there is an overlap between the object boxes. $\text{Union}(\cdot, \cdot)$ denotes the bounding box of the relationship which is the joint box of two objects. We can obtain visual representations of object and relation proposals via performing ROI Pooling on spatial feature maps generated from the visual backbone.

Pre-training Model. RelCLIP adapts a recent successful cross-modal pretraining model called Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) to enhance visual-semantic alignment during relationship learning. CLIP consists of a text encoder $\mathcal{T}(\cdot)$ and an image encoder $\mathcal{V}(\cdot)$, which are pre-trained on 400 million image-text pairs harvested from the web. The text encoder is a Transformer as in (Vaswani et al., 2017). The base architecture of the image encoder can be either a Convolutional Neural Network (CNN) (He et al., 2016) or a Vision Transformer (ViT) (Dosovitskiy et al., 2021). Considering that relationships are usually associated with larger image regions than an object and require much more visual context, we choose the ViT version as it is good at capturing long-range dependencies within an image. The pre-trained image-text encoders are integrated in a plug-and-play manner, and their parameters are fixed during training.

3.2 Cross-modal Relation Embedding

Visual Embedding. We reshape the final transformer states as a 2D spatial feature map $\mathcal{V}(I)$ for image I. We denote the ROI Pooling layer as $f_{ROI}(\cdot)$ which takes the image feature maps and the bounding boxes as the inputs and outputs a feature vector. Besides, we also introduce object and relation feature adaptors (denoted as $h_{\text{obj}}(\cdot)$ and $h_{\text{rel}}(\cdot)$) to project the pre-trained visual embeddings to the domain of visual relationship detection. Each adaptor consists of a fully connected layer and two multi-head attention layers. Given a pair of objects i and j, we calculate the relation representation $\mathbf{v}^r_{ij} \in \mathbb{R}^{D_v}$ as:

$$\mathbf{v}^r_{ij} = h_{\text{rel}}([f_{\text{ROI}}(\mathcal{V}(I), \mathbf{b}_{ij}); \mathbf{p}_{ij}]), \quad (1)$$

where D_v is the dimension of the visual embeddings, $\mathbf{p}_{ij}$ is position embedding. $[\cdot;\cdot]$ denote feature concatenation. Similarly, we can get the object features $\mathbf{v}_i$ and $\mathbf{v}_j$ based on boxes $\mathbf{b}_i$ and $\mathbf{b}_j$.

Language Embedding. Existing visual relationship prediction models only use numeric ids of relation labels, causing severe biases toward the noise and harming the generalization ability. Here we introduce compact language information of relation

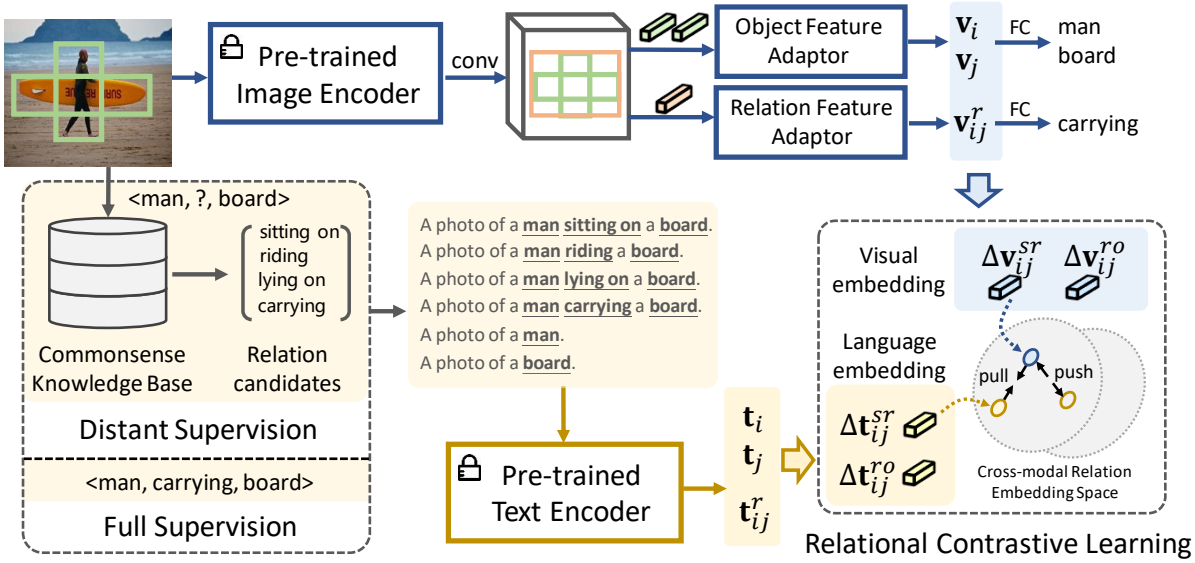

Figure 3: The model architecture of RelCLIP. Given an image, RelCLIP first extracts vision and language embeddings for relationships via adapting fixed cross-modal pre-trained models. Then, the Relational Contrastive Learning (RCL) decouples each relationship triplet into *subject-predicate* and *predicate-object* levels and integrates two contrastive objectives for comparing the relation samples at the two levels. By doing this, our approach improves relationship discrimination and model robustness.

labels for regularizing the representation learning of visual relationships. We adopt the prompt engineering (Radford et al., 2021) to extract language embeddings of each relationship token $\langle a_i, \hat{a}^r_{ij}, a_j \rangle$, $\hat{a}^r_{ij} \in \mathcal{A}^r_{ij}$, as shown in Fig. 3. The resulting text embedding of the relationship is denoted as $\mathbf{t}^r_{ij} \in \mathbb{R}^{D_t}$. Similarly, we can get object text embeddings $\mathbf{t}_i, \mathbf{t}_j \in \mathbb{R}^{D_t}$ by filling the sentence templates using object labels. D_t is the dimension of the text embedding.

3.3 Relational Contrastive Learning

Conventional contrastive learning methods may encounter two serious problems when directly applied to compare relationship instances. First, the large quantity of *subject-predicate-object* combinations will lead to extremely trivial comparisons thus reducing the efficiency of contrastive learning. Second, the undesirable comparisons among synonymous relationships will introduce semantic ambiguity and harm the robustness of relationship discrimination. To address these issues, we propose Relational Contrastive Learning (RCL), which 1) decouples the triplet level comparison into two tuples for reducing the number of comparisons and 2) introduces a new semantic-aware active sampling strategy for excluding synonymous relations and the less informative ones.

Cross-modal relation embeddings. We decouple each relation instance into *subject-predicate* and *predicate-object* levels and compare the same instances at both levels, respectively. To generate embeddings at each level, we exclude the information of the subject or object by subtracting their features alternatively from the feature of the relationship triplet $\mathbf{v}^r_{ij}$ as:

$$\begin{aligned} \Delta\mathbf{v}^{sr}_{ij} &= \left\|\mathbf{v}^r_{ij} - \mathbf{v}_j\right\|_2 \mathbf{W}^{vs}, \\ \Delta\mathbf{v}^{ro}_{ij} &= \left\|\mathbf{v}^r_{ij} - \mathbf{v}_i\right\|_2 \mathbf{W}^{vo}, \end{aligned} \tag{2}$$

where $\|\cdot\|_2$ stands for the L2-normalization. $\mathbf{W}^{vs}$, $\mathbf{W}^{vo} \in \mathbb{R}^{D_v \times D_v}$ are learnable weight matrices. Similarly, the text embeddings are extracted as:

$$\begin{aligned} \Delta\mathbf{t}^{sr}_{ij} &= \left\|\mathbf{t}^r_{ij} - \mathbf{t}_j\right\|_2 \mathbf{W}^{ts}, \\ \Delta\mathbf{t}^{ro}_{ij} &= \left\|\mathbf{t}^r_{ij} - \mathbf{t}_i\right\|_2 \mathbf{W}^{to}, \end{aligned} \tag{3}$$

where $\mathbf{W}^{ts}$, $\mathbf{W}^{to} \in \mathbb{R}^{D_t \times D_t}$ are learnable weight matrices. The cross-modal relation embeddings are extracted at both the *subject-predicate* and *predicate-object* levels as $(\Delta\mathbf{v}^{sr}_{ij}, \Delta\mathbf{t}^{sr}_{ij})$ and $(\Delta\mathbf{v}^{ro}_{ij}, \Delta\mathbf{t}^{ro}_{ij})$, respectively.

Learning Objectives. CLIP is proved to be good at expressing visual concepts for benefiting object recognition tasks while still struggling to capture the semantic relations between objects, as shown in Fig. 2. RelCLIP takes the advantage of CLIP's powerful ability of image-level visual-semantic alignment and explores relation-level alignment via

RCL. RCL aims to match the cross-modal embeddings of relations and pull the samples of the same relation instance together while pushing the different ones away. It maximizes the cosine similarity of the cross-modal embeddings of N positive pairs in the batch while minimizing the cosine similarity of the embeddings of the $N^2 - N$ negative pairs. For mathematical simplicity, we here denote the cross-modal relation embeddings $(\mathbf{v}, \mathbf{t})$ as the inputs of RCL objectives. We adopt the cross-modal version of the InfoNCE loss (van den Oord et al., 2018) which is minimized to lead the encoders to maximally preserve the mutual information between the correctly matched pairs. The image-to-text and text-to-image contrastive losses are defined as:

$$
\begin{aligned}
l_k^{(v\to t)} &= -\log \frac{\exp(\langle \mathbf{v}_k, \mathbf{t}_k \rangle / \tau)}{\sum_{q=1}^{|Q_k|} \exp(\langle \mathbf{v}_k, \mathbf{t}_q \rangle / \tau)}, \\
l_k^{(t\to v)} &= -\log \frac{\exp(\langle \mathbf{t}_k, \mathbf{v}_k \rangle / \tau)}{\sum_{q=1}^{|Q_k|} \exp(\langle \mathbf{t}_k, \mathbf{v}_q \rangle / \tau)},
\end{aligned} \tag{4}
$$

where $\langle \cdot, \cdot \rangle$ represents the cosine similarity between two vectors. Q_k is the actively sampled dictionary that provides informative negative samples for the positive sample according to their semantic meaning. The cross-modal objective of RCL is then computed as a weighted combination of the two losses averaged over all possible image-text pairs in each mini-batch:

$$
\mathcal{L}(\mathbf{v}, \mathbf{t}) = \frac{1}{N^2} \sum_{k=1}^{N^2} (\gamma l_k^{(v\to t)} + (1-\gamma) l_k^{(t\to v)}), \tag{5}
$$

where $\gamma \in [0, 1]$ is a scalar weight. RCL integrates two contrastive objectives regarding a same relation instance at *subject-predicate* and *predicate-object* levels. We instantiate $\mathcal{L}(\mathbf{v}, \mathbf{t})$ with $(\Delta\mathbf{v}_{ij}^{sr}, \Delta\mathbf{t}_{ij}^{sr})$ and $(\Delta\mathbf{v}_{ij}^{ro}, \Delta\mathbf{t}_{ij}^{ro})$ to form RCL loss:

$$
\mathcal{L}_{RCL} = \underbrace{\mathcal{L}(\Delta\mathbf{v}^{sr}, \Delta\mathbf{t}^{sr})}_{\text{subject-predicate}} + \underbrace{\mathcal{L}(\Delta\mathbf{v}^{ro}, \Delta\mathbf{t}^{ro})}_{\text{predicate-object}}. \tag{6}
$$

Together with $\mathcal{L}_{RCL}$, our model is also optimized by two multi-class cross-entropy losses for object classification and predicate classification.

Semantic-aware Active Sampling. Given a query sample k, our method actively constructs its negative sampling dictionary Q_k by ranking the rest samples according to their semantic meaning (e.g., word2vec embeddings (Pennington et al., 2014)). To exclude the synonymous relationships from the negative samples, we remove from Q_k the samples whose similarity scores are higher than a threshold ϵ. To pay more attention to the most informative samples, we assign each of the remaining samples with a weight which is the normalized similarity score with the query sample. We construct the sampling dictionaries Q_k^{sr} and Q_k^{or} for the dual contrastive objectives of RCL as:

$$
\begin{aligned}
Q_k^{sr} &= \{(\Delta\mathbf{v}_n^{sr}, \Delta\mathbf{t}_n^{sr}; w) | \eta \le w\}, w = \langle \mathbf{e}_n^{sr}, \mathbf{e}_k^{sr} \rangle, \\
Q_k^{or} &= \{(\Delta\mathbf{v}_n^{or}, \Delta\mathbf{t}_n^{or}; w) | \eta \le w\}, w = \langle \mathbf{e}_n^{or}, \mathbf{e}_k^{or} \rangle,
\end{aligned} \tag{7}
$$

where η is the threshold for excluding the less informative samples, $\mathbf{e}_n^{sr}$ and $\mathbf{e}_n^{or}$ are word embeddings for the word tokens of *subject-predicate* and *predicate-object* for the n-th relation sample.

Our semantic-aware active sampling strategy constructs dynamic sampling dictionaries for the same relationship in each of the contrastive objectives, which further improves the diversity of negative samples. As is shown in Fig. 4, for the positive sample "*person lying on bed*", the sample "*man lying on grass*" will not be selected as a negative sample at the subject-predicate level. While in the predicate-object level, it is selected as an informative negative sample since it is closer to the positive sample than the other samples in the semantic space.

3.4 Visual Relationship Detection

Prediction. As is shown in Fig. 3, we regard $\mathbf{v}_i$ and $\mathbf{v}_j$ as the learned object features and classify them to predict the final labels of the object i and j. Similarly, the learned relation feature $\mathbf{v}_{ij}^r$ is used to predict final relation label between object i and j. And the corresponding object proposals detected by RPN are employed as our bounding box prediction.

Supervision. RelCLIP is compatible and effective when trained with different types of supervision signals, e.g., full supervision and distant supervision (Yao et al., 2021). Conventional visual relationship detection models require full supervision where all training samples are elaborately annotated. To alleviate the burden of manual annotations, distant supervision is proposed to automatically generate relation labels from the commonsense knowledge base. The knowledge contains a huge amount of relationship triplets parsed from the Conceptual Caption dataset (Sharma et al., 2018). Given an object pair $\langle a_i, a_j \rangle$, we can retrieve possible relations from the knowledge base and get a multi-hot label for training, as shown in Fig. 3. In RCL, all the

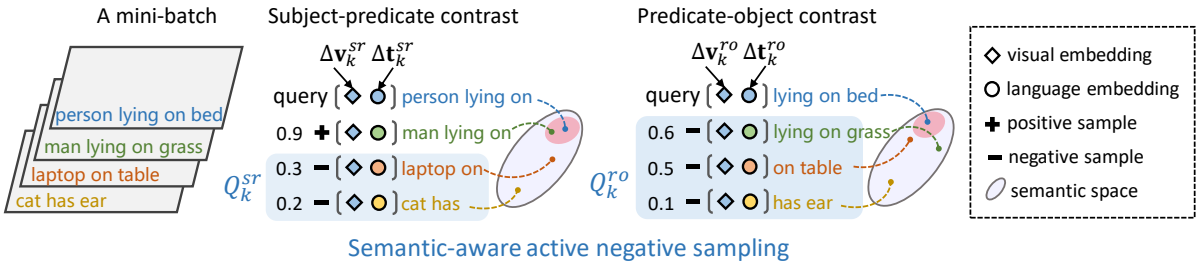

Figure 4: Illustration of the semantic-aware active sampling strategy of our Relational Contrastive Learning. Given a relation instance as a query sample, the negative sampling dictionary excludes the synonymous relation samples and the less informative ones based on their semantic distance to the query samples. Our sampling strategy also builds different dictionaries for each of the contrastive objectives at *subject-predicate* and *predicate-object*.

Models	Predicate Classification				Scene Graph Classification				Scene Graph Detection			
	R@50	R@100	mR@50	mR@100	R@50	R@100	mR@50	mR@100	R@50	R@100	mR@50	mR@100
Distant supervision												
EXT	6.64	9.74	10.66	15.16	3.96	4.82	4.25	4.92	1.93	3.06	1.66	2.49
KB	30.61	33.48	20.98	23.25	15.69	16.99	11.06	12.53	9.36	10.26	6.56	7.13
KB + EXT	38.21	40.90	24.94	27.45	17.52	18.85	11.66	12.56	15.84	18.31	9.49	11.23
Motif (Zellers et al., 2018)	50.23	53.18	33.99	40.62	24.90	26.00	16.50	18.03	20.09	22.74	12.21	14.42
BA-SGG (Guo et al., 2021)	51.48	55.19	38.68	**45.98**	23.72	25.56	17.29	18.55	22.78	25.50	13.38	15.05
VisualDS (KB) (Yao et al., 2021)	51.54	54.53	36.93	41.97	24.81	26.08	16.13	17.56	22.83	24.36	13.48	14.45
VisualDS (KB+EXT) (Yao et al., 2021)	53.40	56.54	37.68	41.98	26.12	27.46	17.20	18.39	22.10	24.24	**13.84**	15.23
RelCLIP (KB)	54.15	57.06	37.57	41.98	26.99	27.89	21.44	**24.92**	22.80	25.53	13.03	15.11
RelCLIP (KB+EXT)	**55.06**	**58.46**	**39.12**	43.47	**27.57**	**28.41**	**23.88**	24.75	**24.06**	**26.70**	13.75	**15.45**
Full supervision												
BA-SGG (Guo et al., 2021)	68.04	70.08	47.19	50.52	30.48	31.49	16.75	17.71	28.40	33.23	13.35	15.67
VisualDS (Yao et al., 2021)	67.93	70.20	52.65	55.41	31.14	31.92	23.53	25.27	28.90	31.25	18.26	20.63
RelCLIP (Ours)	**69.95**	**72.55**	**56.99**	**61.21**	**34.79**	**35.60**	**26.90**	**27.89**	**33.74**	**36.51**	**26.64**	**29.52**

Table 1: Comparison with semi, distant and fully supervised visual relationship detection methods on VG dataset.

Method	P@10	P@20	R@50	R@100
RelCLIP (No RCL)	39.42	37.14	54.37	57.11
RelCLIP (Ours)	**49.41**	**46.10**	**55.06**	**58.46**

Table 2: Human evaluation results (precision @K) on the PredCls task.

possible labels are fused according to their weights to form the language embedding. Training models under distant supervision is more challenging than full supervision since the retrieved labels may introduce noise.

4 Experiments

4.1 Settings

Datasets. We evaluate our method on the popular large-scale Visual Genome (VG) (Krishna et al., 2017) benchmark including approximately 108k images. Following previous works (Chen et al., 2019; Yao et al., 2021), we employ the data split which contains the most frequent 150 object categories and 20 well-defined predicates. The refined predicate schemes defined by Chen *et al.* (Chen et al., 2019) remove synonyms and super-sets from the 50 predicates in the Visual Genome dataset.

Tasks. We follow three conventional tasks (Zellers et al., 2018; Chen et al., 2019; Chen et al., 2019; Yao et al., 2021) to evaluate the proposed SGG model: 1) Predicate classification (PredCls), which predicts the predicate labels given a ground truth set of object boxes and object labels, 2) Scene graph classification (SGCls), where both the object classes and the relation type of each object pair are predicted given the ground-truth object bounding boxes. 3) Scene graph detection (SGDet), which only takes the original image as input and sequentially predicts the object bounding boxes, the object labels and the relationships between object pairs.

Metrics. We use the following metrics which are under graph constraint to evaluate the performance of the proposed approach: (1) Recall@K (R@K) (Lu et al., 2016), which computes the fraction of relationship hits in the top K confident relationship predictions, (2) mean Recall@K (mR@K) (Tang et al., 2019; Chen et al., 2019), which takes the average R@K of all predicate classes to give a fair performance measure for both head and tail classes.

Implementation Details. We implement our

Method	Relational Contrast Level	R@50	R@100	mR@50	mR@100
Unary Contrast	predicate	54.14	57.50	37.27	41.75
Triple Contrast	subject-predicate-object	54.59	57.63	37.15	40.09
Dual Contrast (Ours)	subject-predicate, predicate-object	**55.06**	**58.46**	**39.12**	**43.47**

Table 3: Ablation studies of the relational contrastive learning on the PredCls task. The results are with graph constraint on the VG dataset.

Method	Pretrain	R@50	R@100
VisualDS	ImageNet1K	53.40	56.54
RelCLIP	ImageNet1K	54.08	57.47
RelCLIP (Ours)	CLIP	**55.06**	**58.46**

Table 4: Ablation of pretraining data.

model using the MindSpore Lite tool (MindSpore). We train our model with the SGD optimizer for 20,000 iterations with image batch size 12. The initial learning rate is 0.001. The dimension values D_t and D_v of the cross-modal embeddings are both set to 512. We use the pre-trained CLIP model that is publicly available [1]. The dimension of the visual and language embeddings D_t and D_v are set to 512. The sampling threshold ϵ for splitting the positive and negative samples is set to 0.7.

4.2 Visual Relationship Detection

Compared Baselines. We compare our method with strong baselines under distant, and full supervision. The **distant supervised** models are trained using a set of relation candidates retrieved from the commonsense knowledge collected on Conceptual Caption (Sharma et al., 2018). The candidates are image-agnostic and each of them has equal probability. To further enhance the distant supervision (KB), VisualDS (Yao et al., 2021) also introduces an external semantic signal (EXT) which leverages CLIP models to assign probability scores to the relation candidate sets regarding the image content. We adopt two **fully supervised** methods as strong baselines, i.e., *Motif* (Zellers et al., 2018) which is a widely used scene graph model and *BA-SGG* (Guo et al., 2021) is one of the most recent models.

Numerical Results. In Tab. 1, we compare our approach with existing baselines of visual relationship detection under different types of supervision signals. Our RelCLIP achieves consistent improvements over prior arts (e.g., VisualDS (Yao et al., 2021) and BA-SGG (Guo et al., 2021)) on all sub-tasks under both distant and full supervision, demonstrating that RelCLIP can learn more robust and discriminative representations for visual relationships. Though distant supervision provides many noise labels, RelCLIP can still learns discriminative representations from a small amount of correctly labeled samples with the help of the language knowledge of relation labels.

Human Evaluation. Since there are amounts of unsatisfactory annotations in the VG dataset as shown in Fig. 5, the Recall metric is insufficient to comprehensively evaluate the accuracy of predictions. We ask 10 people to evaluate the precision of our predicted relations. Predictions from 200 images are selected and the results are reported as mean Precision@10 and 20, see Tab. 2. When comparing with RelCLIP (No RCL), the precision (P@10/20) of improves nearly 10%, which further demonstrates the excellent ability of RelCLIP in recognizing visual relationships.

4.3 Ablation Study

In this section, we evaluate the effectiveness of the training objectives, sampling strategies, and pretraining knowledge under the most challenging setting of distant supervision.

Effect of relational contrastive objectives. We develop variants of RCL to verify the effectiveness of dual contrastive objectives. The experimental results on PredCls are presented in Tab. 3. First, we replace our *Dual Contrast* objectives with *Unary Contrast* which solely contrasts relationship instances at predicate level and find that the performance drops. This is because the Unary Contrast fails to capture the semantic dependencies between predicates and objects. Then, we use *Triple Contrast* where relationship instances are contrasted at subject-predicate-object level. The results on Recall and mean Recall are lower than dual contrast, which indicates that our dual contrast can better capture the complicated and cluttered dependencies between the predicate and objects.

Effect of pretraining models. We replace RelCLIP's vision encoder with an ImageNet1K pretrained backbone. As shown in Tab. 4, RelCLIP achieves higher performance when using the same pretrained backbone as VisualDS, and the perfor-

[1] https://github.com/openai/CLIP, ViT-B/32

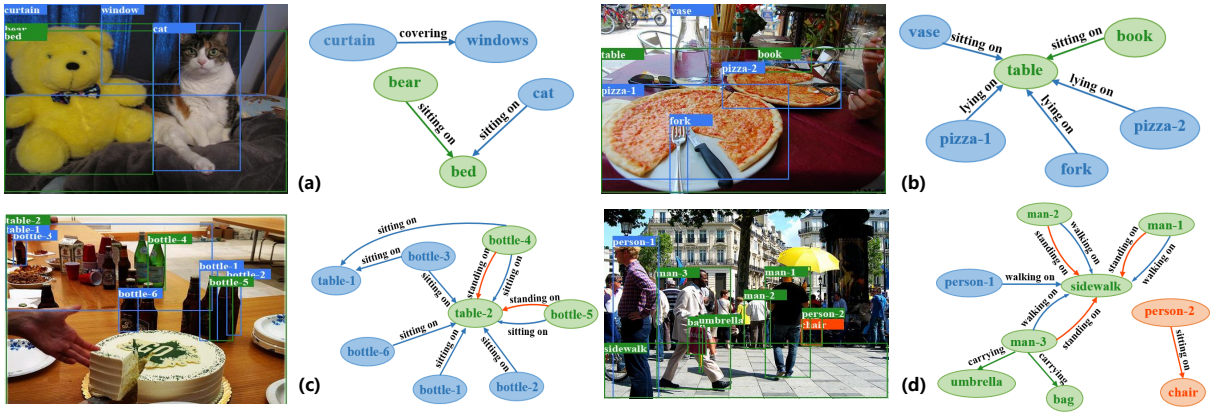

Figure 5: Visual relationship detection examples on VG Dataset. Green boxes and edges are correct predictions, Orange boxes and edges are missed in our prediction. Blue boxes and edges indicate reasonable predictions from our model but are not annotated in GT. Best viewed in digital version.

Method	R@50	R@100	mR@50	mR@100
No Sampling	52.86	56.19	34.27	36.99
No Del. Syn.	53.58	56.71	35.05	37.64
No Weight	53.67	56.38	35.82	41.19
Ours	**54.15**	**57.06**	**37.57**	**41.98**

Table 5: Ablation of our sampling strategy.

Threshold	R@50	R@100	mR@50	mR@100
0.3	52.98	55.80	35.19	37.91
0.5	53.81	56.69	33.84	38.86
0.7	**54.15**	**57.06**	**37.57**	**41.98**
0.9	52.23	54.98	34.04	36.63

Table 6: Ablation of the threshold ϵ of our sampling.

mance can be further improved when using CLIP models, which demonstrates the ability of RelCLIP in adapting cross-modal pretraining knowledge.

Effect of semantic-aware active sampling. We build three baselines for validating the effect of our sampling strategy to construct the effect negative sampling dictionary. 1) The *No Sampling* baseline disables the whole sampling functionality during model training. 2) The *No Del. Syn.* baseline preserved the synonymous relation instances. 3) The *No weight* baseline removes the synonymous relationships but disables the sample weighting which suppresses the effect of less informative samples. Compared with *Ours* that excludes both the synonymous instances and the less informative ones, the performances of the ablation baselines above drop, indicating the effectiveness of our semantic-aware active sampling strategy, see Tab. 5.

Effect of the threshold ϵ of sampling. We set the threshold value ϵ to 0.3, 0.5, 0.7, and 0.9 for training RelCLIP models and the results on the PredCls task are shown in Tab. 6. When the threshold is set to 0.7, we can obtain the highest performance. If we use a larger threshold (e.g., 0.9), more synonymous negative samples will be included in the sampling dictionary and the performance drops. If we use a smaller threshold (e.g., 0.3 and 0.5), some informative samples will be removed undesirably and thus lead to performance drop.

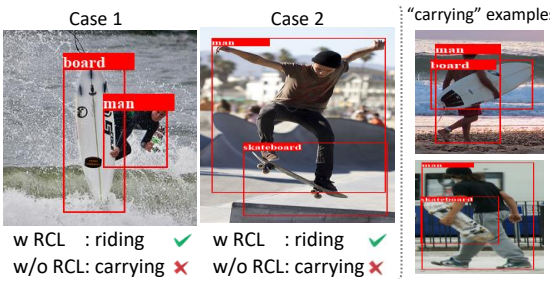

Figure 6: Case study of "riding" and "carrying". RCL can help the model correctly capture the semantic difference between visual relationships of "man riding board" and "man carrying board".

5 Visualization

Case Study. In Fig. 6, we compare the accuracy of "riding" of RelCLIP with and without RCL on Recall@100. When RCL is disabled, the performance drops by a significant margin of 5%, and a large proportion of the "riding" failures are misclassified as "carrying". As shown in the two failure cases, RelCLIP with RCL can well capture the minor differences of interactions between man and board via comparing the similarity and discrepancy across relation instances, thus correctly distinguish "riding" and "carrying".

Prediction Examples. We show some visualization examples in Fig. 5. Though trained with distant supervision, RelCLIP can correctly detect relation-

ships (in green). Surprisingly, numerous reasonable relationships (in blue) are also predicted by RelCLIP, even though they are missed in the human annotations, e.g., in Fig. 5(a), "cat sitting on bed" can be obviously observed by humans, but is not labeled in GT. Furthermore, we observed that our failure cases (in orange) are more precise than GTs from the human perspective, e.g., in Fig. 5(d), the relationship between "man-3" and "sidewalk" is more like "walking on" rather than "standing on".

6 Conclusion

In this work, we develop RelCILP, a simple yet effective visual relationship detection framework that successfully adapts cross-modal pre-training for improving relationship discrimination. To achieve this, we propose Relational Contrastive Learning (RCL) that not only enables efficient comparisons via decoupling triplet-level contrast into *subject-predicate* and *predicate-object* levels, but also ensures effective comparisons via a novel semantic-aware sampling strategy. By collaboratively learning the semantic coherence and discrepancy from relationship instances, the model can generate more discriminative and robust relationship representations. Our RelCLIP significantly outperforms existing methods under both full and distant supervision, demonstrating the effectiveness and generalization ability of relational contrast powered by cross-modal pretraining models.

Limitations

The limitations of RelCLIP can be summarized as follows: First, Relational Contrastive Learning (RCL) contrasts the visual and language embeddings of relationships within each mini-batch, which makes RCL sensitive to batch size and may introduce bias into the training of RelCLIP. In our future work, we will try to introduce a global memory bank for achieving cross-batch comparison. Second, RelCLIP can learn visual relationships under distant supervision without any annotations. However, the performance gap between distant and full supervision is still large, because distant supervision may inevitably introduce noisy labels. The quality of distant supervision depends on the commonsense knowledge base, so the quality and capacity of the knowledge base may affect the performance of RelCLIP to some extent. Third, RelCLIP is implemented following the two-stage framework where the visual relationship detection problem is decoupled into object detection and relationship classification. The powerful knowledge of CLIP is integrated to enhance the discrimination of relationships, but it is not used to improve the object detector. So the performance of RelCLIP is still limited by the accuracy of object detection. In future works, we will explore how CLIP models can be used to simultaneously improve both the object detector and the relation classifier.

Acknowledgement

We gratefully acknowledge the support of MindSpore, CANN (Compute Architecture for Neural Networks), and Ascend AI Processor used for this research. This work was supported in part by National Key R&D Program of China under Grant 2020AAA0109700, Guangdong Outstanding Youth Fund under Grant 2021B1515020061, GuangDong Basic and Applied Basic Research Foundation under Grant 2022A1515011835, and Anhui Provincial Natural Science Foundation under Grant 2108085UD12. We also acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

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 (ICCV)*, pages 2425–2433.

Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. 2019. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6163–6171.

Vincent S Chen, Paroma Varma, Ranjay Krishna, Michael Bernstein, Christopher Re, and Li Fei-Fei. 2019. Scene graph prediction with limited labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2580–2590.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 104–120.

Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3298–3308.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yuyu Guo, Lianli Gao, Xuanhan Wang, Yuxuan Hu, Xing Xu, Xu Lu, Heng Tao Shen, and Jingkuan Song. 2021. From general to specific: Informative scene graph generation via balance adjustment. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 16383–16392.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 1024–1034.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738.

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 (CVPR)*, pages 770–778.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1219–1228.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678.

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

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 Association for the Advance of Artificial Intelligence (AAAI)*, volume 34, pages 11336–11344.

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*.

Rongjie Li, Songyang Zhang, and Xuming He. 2022. SGTR: End-to-end scene graph generation with transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19486–19496.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2020. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 2592–2607.

Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. 2017. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1270–1279.

Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 852–869.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 32, pages 13–23.

MindSpore. Mindspore. https://www.mindspore.cn/.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Mengshi Qi, Yunhong Wang, and Annan Li. 2017. Online cross-modal scene retrieval by binary representation and semantic graph. In *Proceedings of the ACM Multimedia (ACM MM)*, pages 744–752.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pages 91–99.

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 Association for Computational Linguistics (ACL)*, pages 2556–2565.

Jiaxin Shi, Hanwang Zhang, and Juanzi Li. 2019. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8376–8384.

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 *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5099–5110.

Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. 2019. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6619–6628.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. 2021. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16684–16693.

Danfei Xu, Yuke Zhu, Christopher Bongsoo Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3097–3106.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10685–10694.

Yuan Yao, Ao Zhang, Xu Han, Mengdi Li, Cornelius Weber, Zhiyuan Liu, Stefan Wermter, and Maosong Sun. 2021. Visual distant supervision for scene graph generation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 15816–15826.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6720–6731.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural Motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5831–5840.

Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed Elgammal. 2017. Relationship proposal networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5226–5234.

Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. 2019. Large-scale visual relationship understanding. In *Proceedings of the Association for the Advance of Artificial Intelligence (AAAI)*, volume 33, pages 9185–9194.

Yaohui Zhu and Shuqiang Jiang. 2018. Deep structured learning for visual relationship detection. In *Proceedings of the Association for the Advance of Artificial Intelligence (AAAI)*, pages 7623–7630.

Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. 2017. Towards context-aware interaction recognition for visual relationship detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 589–598.