Self-Supervised Learning for Visual Relationship Detection through Masked Bounding Box Reconstruction

Zacharias Anastasakis 1,4
zaraisanastasakis@gmail.com
Dimitrios Mallis 2
dimitrios.mallis@uni.lu
Markos Diomataris 3
mdiomataris@student.ethz.ch
George Alexandridis 4
gealexandri@islab.ntua.gr
Stefanos Kollias 4
stefanos@cs.ntua.gr
Vassilis Pitsikalis 1
vpitsik@deeplab.ai
1 Deeplab, Athens 2 SnT, University of Luxembourg 3 ETH, Zürich 4 National Technical University of Athens

Abstract

We present a novel self-supervised approach for representation learning, particularly for the task of Visual Relationship Detection (VRD). Motivated by the effectiveness of Masked Image Modeling (MIM), we propose Masked Bounding Box Reconstruction (MBBR), a variation of MIM where a percentage of the entities/objects within a scene are masked and subsequently reconstructed based on the unmasked objects. The core idea is that, through object-level masked modeling, the network learns context-aware representations that capture the interaction of objects within a scene and thus are highly predictive of visual object relationships. We extensively evaluate learned representations, both qualitatively and quantitatively, in a few-shot setting and demonstrate the efficacy of MBBR for learning robust visual representations, particularly tailored for VRD. The proposed method is able to surpass state-of-the-art VRD methods on the Predicate Detection (PredDet) evaluation setting, using only a few annotated samples. We make our code available at https://github.com/deeplab-ai/SelfSupervisedVRD.

1. Introduction

Developing machines endowed with the ability to interpret and make decisions based on visual inputs is a critical Computer Vision objective. Research in the field of scene understanding aims to analyze an entire scene or image, in a way similar to that of a human observer. Visual Relationship Detection (VRD) provides an effective approach to scene understanding and constitutes an important component of larger vision pipelines for tasks such as image captioning [19] and visual question answering [26,32,34].

VRD goes beyond identifying/classifying individual components or objects, to extracting the relationships between detected entities. A relationship can be defined as a triplet in the form of $< S, P, O >$, which indicates that the subject $S$ correlates with the object $O$ through the pred-
icate $P$, which is usually selected from a finite set of possible predicate classes. In the example of Fig. 1, subject [PERSON] correlates with object [SKATEBOARD] through the predicate [RIDE]. Thus, a scene can be represented in a structured way, i.e., a directed graph where the nodes are the detected entities, while the edges represent the relations among them with a direction from subject to the object.

The emergence of large datasets providing relationship-level annotations has enabled supervised learning as the dominant approach for visual relationship detection [4, 20, 21, 37, 39, 42, 44]. However, labeled data for VRD is difficult to obtain while annotating large-size datasets is a quite time-consuming and expensive procedure. Consider that potential per-image relationships can grow exponentially to the number of objects in the scene, requiring multiple annotations to accurately capture the complex web of possible interactions and associations. Given the apparent cost of data annotation, existing datasets only capture the relationships between a few object categories mostly centered around the human-sensing domain, thus limiting a wider range of potential applications on novel object categories. Finally, defining relationships through distinct predicate classes can introduce various ambiguities (multiple predicate classes can be representative of a visual relationship), human annotators are heavily biased and predicate classes commonly demonstrate a long tail distribution on existing datasets [38]. Given the identified limitations of the supervised approach, this work focuses on the self-supervised learning (SSL) paradigm for visual relationship detection. SSL involves training models using a pretext task where supervision is provided from the data without requiring any manual annotations. Despite impressive performance in both natural language processing [2, 5] and various computer vision tasks [1, 9, 18, 23, 31], self-supervised approaches for VRD have attracted less attention.

Our proposed method comprises of a 2-stage pipeline. Initially, a model is pre-trained in a self-supervised manner on large-scale datasets and is able to learn meaningful representations which are later fine-tuned and demonstrate strong performance in few-shot relationship classification. At the core of our architecture lies a transformer encoder that learns to capture object relationships without requiring any manually annotated labels. The transformer is fed object-level visual features, extracted from bounding boxes using Faster-RCNN [29]. We propose Masked Bounding Box Reconstruction (MBBR), a pretext task particularly for VRD, where a percentage of the input features is randomly masked, and the model is charged with predicting the masked representations solely based on the context provided by the unmasked object features. The key insight is that through MBBR, the model learns representations that encapsulate the complex relationships between objects. These context-aware representations, garnered from the interaction of objects within a scene, prove to be highly predictive of predicates, thereby enabling effective VRD (framework overview in Fig. 1).

The performance of learned representations is evaluated on few-shot classification. A VRD classifier is trained for a $k$-shot setting, on top of the features derived from our trained transformer encoder. We show that a simple 2-layer Multi-Layer Perceptron (MLP) can achieve state-of-the-art results for few-shot VRD, thus demonstrating the effectiveness of the proposed SSL framework.

In summary, our contributions are: (1) We propose MBBR, a novel self-supervised pretext task particularly tailored for self-supervised VRD pre-training, that does not require manual relationship-level annotation. (2) Our proposed method can learn rich relationship-aware object representations that can be used for downstream predicate detection in a few-shot setting. (3) We extensively evaluate our proposed approach both quantitatively and qualitatively on both VRD and VG200 datasets, achieving large performance improvements for few-shot Visual Relationship Detection.

2. Related Work

2.1. Visual Relationship Detection

The VRD task was originally introduced in [30], where authors treat each triplet $< S, P, O >$ as an individual class for which a separate classifier is trained. This early formulation required training of $O(N^2K)$ classifiers (where $N$, $K$ is the number of entities and predicates respectively), which was later [22] reduced to $O(NK)$ by training a single classifier per predicate class. Lu et al. [22] proposed the addition of linguistic priors to exploit entity semantics, leading to enhanced predicate classification. In [41], authors conduct a statistical analysis on the Visual Genome (VG) dataset [38] and discover valuable motifs, i.e., repeated structures and recurring patterns across large scene graphs, which exploit later using bidirectional LSTMs.

VTransE, proposed in [43], is a visual translation embedding network which projects the subject $S_{emb}$, object $O_{emb}$ and predicate $P_{emb}$ representations into a low dimensional embedding vector space where $S_{emb} + P_{emb} = O_{emb}$. Following [43], authors in [15] introduced UVTransE based on the observation that the subtraction of $S_{emb}$ and $O_{emb}$ from the embedding of the union of subject and object, results in the predicate embedding, i.e., $U_{emb} - S_{emb} - O_{emb} = P_{emb}$. In [11] multi-head attention is employed, with a separate head for each predicate class, thus the model is able to attend on multiple visual regions of the input image. Inspired by Faster-RCNN [29] propose Graph-RCNN [40], a graph network which utilizes a relationship proposal network (RepPN) and an attentional graph convolutional network (aGCN), to al-
low contextual information sharing among objects. Compared to these methods, our proposed approach does not rely on relationship-level supervision for model training. Relationship-aware object representations are learned instead in a self-supervised manner through Masked Bounding Box Reconstruction.

Only a few works explore visual relationship detection under reduced manual supervision. Authors in [33] introduce an energy-based loss for VRD and conduct a few-shot evaluation on relationship triplets. For our few-shot formulation in contrast, a classifier is trained with \( k \) samples per predicate class (instead of relationship triplets). To enhance performance under long-tailed distribution for predicate classes of popular VRD datasets [38], authors in [8] propose a two-stage approach. Pre-training is originally performed on the 25 most frequent classes, followed by fine-tuning through few-shot VRD classification on the 25 remaining classes. Predicates are learned as functions, which are then utilized as message-passing mechanisms within a graph convolutional network (GCN). Since [8] focuses on the 25 most frequent predicates, it remains a supervised approach, in contrast to our self-supervised pre-training task that requires no manual annotations. To the best of our knowledge, we are the first to propose a fully self-supervised pre-training pretext task based on bounding box reconstruction, specifically tailored for few-shot Visual Relationship Detection.

### 2.2. Self-Supervised Learning

SSL is the main learning approach used in this work for learning relationship-aware entity representations. It leverages inherent structures and patterns of the data to learn meaningful features [7, 10, 25, 27], thus circumventing the need for manual annotations. Several techniques have been proposed in the literature, with contrastive learning and Masked Image Modeling (MIM) recently attracting the most attention. Contrastive methods like SimCLR [3] or MoCo [15] can learn strong representation by comparing augmented versions of the same image (positive pairs) with distinct images (negative samples). MIM based methods learn through masking a part of an image that is subsequently reconstructed. He et al. [12] propose to mask a significant portion of an image (75% of the image patches) and use autoencoders for patch reconstruction. In BEit [1], the model is pre-trained on discrete visual tokens from a randomly masked section of the image, that are obtained from the latent representations of discrete variational autoencoders [28]. Related to ours is the pretext task proposed in [6] for image manipulation instead of predicate detection. Compared to our work, authors in [6] explore a complex network architecture that also leverages semantic features and perform reconstruction in the pixel space instead of the feature space, utilizing generative networks. In this work, motivated by the recent success of MIM for representation learning based on the transformer architecture, we propose an SSL pipeline based on Masked Bounding Box Reconstruction, particularly for VRD representation learning.

### 3. Approach

#### 3.1. Problem Definition

Let \( D = \{I_i, b^S_{(i,j)}, b^O_{(i,j)}, S_{(i,j)}, O_{(i,j)}, P_{(i,j)}\} \) be a dataset of \( N \) images with \( i \in [1, N] \), where \( b^S_{(i,j)} \) and \( b^O_{(i,j)} \) are the subject / object bounding boxes and \( S_{(i,j)}, O_{(i,j)} \),
\( P_{i,j} \) denote the subject labels, object labels and predicate labels for the \( j \)th relationship in image \( I_i \).

In this work, we focus on the Predicate Detection task (PredDet), where the goal is to learn a mapping function

\[
\Psi : \{I, b^S, b^O, S, O\} \rightarrow P_k
\]

with \( k \in [1, K] \) being the set of possible predicate classes.

We propose to address the above problem through a 2-stage architecture. Firstly (Eq. 2), a deep neural network with weights \( \theta_f \) learns to extract a representation \( z \) for each object on a scene in a self-supervised manner without using any object class or predicate class annotations.

\[
\Psi_{\theta_f} : \{I, b\} \rightarrow z
\]

Then in the second stage (Eq. 3), a small 2-layer MLP with parameters \( \theta_s \) is used to predict

\[
\Psi_{\theta_s} : \{I, z^S, z^O, S, O, b^S, b^O\} \rightarrow P_k
\]

powered by the representations extracted for \( z^S \) and \( z^O \) during stage-one pre-training. We will describe our proposed architecture for training both \( \Psi_{\theta_f} \) and \( \Psi_{\theta_s} \) in the following subsection.

### 3.2. Self-supervised representations for VRD

\( \Psi_{\theta_f} \) is trained to reconstruct the feature representation of masked objects, through the context provided by the rest of the objects in the scene (Fig. 2). Thus, learned features are highly predictive of the relationships between entities and demonstrate robust performance for few-shot predicate detection.

**Learning Target Formation:** To form the learning target that will be used for MBBR, we initially pass an input image through an off-the-shelf, pre-trained Faster-RCNN to extract visual features. We then apply multi-scale feature pooling for all features inside each entity bounding box \( b \), thus extracting a single visual representation \( f_b \in \mathbb{R}^{256} \) per entity of the scene.

**Visual geometry embeddings:** Visual geometry embeddings are also used to capture the arrangements of visual embedding within the scene. Following Hu et al. [14], we represent the position of each image entity by a 4-d vector \((X_{LT}, Y_{LT}, X_{RB}, Y_{RB})\), where \((X_{LT}, Y_{LT})\) and \((X_{RB}, Y_{RB})\) are the coordinates of the top left and right bottom corners of the bounding box, respectively, of the entity and \( W, H \) are the width and height of the input image. This vector is then projected into a high-dimensional space \( f_{pos} \in \mathbb{R}^{256} \) by computing sine and cosine functions of different wavelengths.

**Masked Bounding Box Reconstruction:** We randomly mask each entity feature \( f_b \) with a probability of 50\%. Features are then concatenated with geometry embeddings \( f_{pos} \) and projected through a linear layer to an entity embedding \( f_e \in \mathbb{R}^{256} \). For this work, \( \Psi_{\theta_f} \) (Eq. 2) is modelled as a standard feed-forward transformer encoder of [35]. The output of the transformer is a representation per object \( z_i \) that is then projected through a linear layer to form the reconstructed entity embeddings \( \hat{y}_{rec,i} \). Our model is trained through a mean square error (MSE) loss between reconstructed entity embeddings and the input embeddings \( f_{b,i} \) from the pre-trained Faster-RCNN, for all \( N \) entities in a scene. The embeddings \( z_i \) can be used later for training a simple MLP in a few-show setting, as discussed next.
3.3. Few-shot Classification

After self-supervised pre-training, we keep only the encoder without the projection head, and treat the derived representations $z_i$ as visual features for training a classifier on few-shot VRD (Fig. 3). In addition to the visual features, we employ linguistic features, i.e., the word2Vec [24] embeddings of each entity’s label, and spatial features [11] for modeling the respective location of the subject-object pairs. The few-shot classifier takes as input the concatenation of the above features and is trained for predicate class prediction using a standard cross-entropy loss.

4. Experiments

4.1. Datasets and Metrics

Evaluation is performed on two commonly used publicly available datasets, namely VRD [22] and VG [38]. The first is a widely used for visual relationship detection. It contains 5,000 images with 100 object and 70 relationship categories. We use the same split as [22], i.e., 4,000 training images and 1,000 test images. The total number of annotated triplets is 203,284 in the training set and 7,624 in the test set.

Visual Genome is one of the largest datasets in visual relationship detection. It contains 108,077 images, 3.8 million annotated objects and 2.8 million annotated triplets. We follow the same train/test split as in [38], i.e., 75,651 training images and 32,422 testing images with 150 object classes and 50 relation classes.

In this paper, we focus on the PredDet task. Our proposed SSL pipeline is evaluated in a few-shot setting. Note that in our formulation, $k$-shot refers to $k$ samples per predicate class. Thus, for 10-shot evaluation, in VRD we will use $70 \times 10$ relationships, where 70 is the number of predicate classes. We use $\text{Recall}_k@N \ (R_k@N)$ as our evaluation metric. Given an input image with $x$ subject-object pairs, $R_k@N$ considers only the top-$k$ predictions for each pair and then selects the $N$ most confident out of a total of $x \cdot k$ predictions. Following [11], we refer to evaluation with $k = 1$ as graph constraints, indicating that only one edge between entities is allowed. Larger values of $k$ are signified as no graph constraints, allowing multiple edges between entities. In this work, when referring to no graph constraints, $k$ is set to 50 and 70 for the VG200 and VRD, respectively.

4.2. Implementation details

The transformer encoder we use is comprised of 8 attention heads, 6 layers and a feature dimension of 256. We use the Adam optimizer [17] with a base learning rate $2 \times 10^{-3}$ and weight decay of $10^{-4}$. The model is trained for 30 epochs with batch size 16. After pre-training, we further fine-tune our model on a $k$-shot setting for 20 more epochs. For all reported experiments, self-supervised pre-training is performed on VG200 [38] as in VRD there are not enough images for our model to learn useful representations. All of our models are implemented in PyTorch.
Figure 4. Directed graphs extracted from 10-shot classifiers on the VRD dataset [22]. We provide a qualitative comparison of our method and UVTransE [15]. Even in cases where our method does not recover the ground truth triplet, predicted predicates remain semantically accurate. For example see groundtruth relationship \(<\text{BUILDING}, \text{BELOW}, \text{SKY}>\) (second column). Our method recovers a plausible triplet \(<\text{BUILDING}, \text{UNDER}, \text{SKY}>\) compared to \(<\text{BUILDING}, \text{HAS}, \text{SKY}>\) for UVTransE.

<table>
<thead>
<tr>
<th>Method</th>
<th>Graph Constraints 1-shot</th>
<th>Graph Constraints 2-shot</th>
<th>Graph Constraints 5-shot</th>
<th>No Graph Constraints 1-shot</th>
<th>No Graph Constraints 2-shot</th>
<th>No Graph Constraints 5-shot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster-RCNN [29]</td>
<td>4.23±2.87</td>
<td>4.24±2.4</td>
<td>5.2±5.34</td>
<td>9.65±2.36</td>
<td>10.52±2.12</td>
<td>12.6±4.55</td>
</tr>
<tr>
<td>Motifs [41]</td>
<td>0.1±0.1</td>
<td>1.15±2.35</td>
<td>9.62±7.97</td>
<td>2.28±2.54</td>
<td>2.30±2.55</td>
<td>11.76±8.17</td>
</tr>
<tr>
<td>VTransE [43]</td>
<td>7.49±1.55</td>
<td>9.43±2.69</td>
<td>12.10±1.45</td>
<td>14.43±1.05</td>
<td>16.6±3.90</td>
<td>23.4±1.63</td>
</tr>
<tr>
<td>UVTransE [15]</td>
<td>4.47±2.84</td>
<td>7.52±2.88</td>
<td>12.6±0.98</td>
<td>10.95±5.95</td>
<td>14.88±4.59</td>
<td>23.6±1.37</td>
</tr>
<tr>
<td>ATR-Net [11]</td>
<td>1.43±1.30</td>
<td>1.17±0.62</td>
<td>2.67±0.84</td>
<td>3.96±1.92</td>
<td>5.28±2.56</td>
<td>8.75±0.69</td>
</tr>
<tr>
<td>Our method</td>
<td><strong>9.53±1.88</strong></td>
<td><strong>11.98±2.52</strong></td>
<td><strong>19.90±1.00</strong></td>
<td><strong>16.92±3.14</strong></td>
<td><strong>19.22±3.18</strong></td>
<td><strong>29.92±2.83</strong></td>
</tr>
</tbody>
</table>

Table 3. Comparison of our method with state-of-the-art on \{1, 2, 5\}-shot predicate detection on VRD dataset [22]. For this experiment, samples used for few-shot learning are selected manually to ensure effective training.
4.3. Results

In this section, we analyze the effectiveness of our proposed approach, both quantitatively and qualitatively. Since self-supervised representation learning for predicate detection is a previously unexplored area, comparisons are performed w.r.t. recent supervised methods, trained on a few-shot setting [11, 15, 41, 43]. We also include a Faster-RCNN baseline where instead of using $z_1$ derived from our pre-trained SSL model, we use the object-level visual features that are extracted from the Faster-RCNN.

Tables 1 and 2 summarize the obtained results on both datasets. Evaluation for PredDet is performed in a 10 and 20-shot setting. We observe that through self-supervised training, our encoder can learn robust and generalizable representations that surpass both recent supervised methods and the Faster-RCNN baseline in the few-shot setting, thus demonstrating the effectiveness of MBBR.

Additional few-shot results are provided on Table 3 for {1, 2, 5}-shots. For this evaluation, we opt to manually select the few accurate relationships that are used to train our classifiers. The reason is that relationship tuples in both VRD and VG200 can be highly noisy [43], e.g., <SKY, HAS, SKY> and learned classifiers might fail to generalize when trained on a very small number of noisy examples. As previously mentioned, we find that our pre-trained model surpasses all related methods by a large margin. Interestingly, the 5-shot setting in Table 3 results in a similar performance to 20-shots in Table 1, thus further demonstrating the importance of selecting accurate relationships for few-shot classification.

5. Ablation Studies

We conduct an ablation study to further investigate MBBR as a pretext task for VRD representation learning.
Reconstruction vs Classification Loss. Our model is trained to reconstruct inputs through an MSE loss on feature space (similar to the commonly used perceptual loss [16]), achieving a $R@20$ score of $20.87 \pm 2.46$ for 10-shot predicate detection with graph constraints on the VRD dataset [22] (Table 1 for 5 random initializations). An alternative approach could be learning representations through the prediction of the masked object’s class, as in [31]. We find that a classification loss degrades representation learning performance, with a $R@20$ score of $16.7 \pm 1.51$ on the same task.

Impact of linguistic / spatial features. For few-shot predicate classification, our method utilises not only visual representations $V$ (learned through MBBR), but also linguistic $L$ and spatial $S$ features (as in [11]). To quantify the impact of linguistic and spatial features in overall model performance, we compare our complete model $(L+S+V)$ to a variant that only utilises language / spatial features (denoted as $L+S$). In Table 4, we show that the addition of visual representation leads to superior performance, with an approximate 8% increase in 10-shot and 5.5% increase in 20-shot in $R@20$ compared to the variant that does not perceive visual information.

Self-Attention Scores. Through MBBR we learn object representations that are highly predictive of object relationships. To gain further insight into the entity relationships discovered by our self-supervised approach, we provide self-attention visualizations. In Fig. 6 we observe entities tend to attend to other subjects/objects in the scene on which they are connected with a visual relationship. For example in row 3 of Fig. 6, the [SKI]’s attend more to the [PERSON] and other [SKI]’s for representation prediction and less to the person’s [HEAD] or [HAT].

Qualitative Results. We also include a qualitative comparison with the UVTransE method of [15] of Fig. 4 on 10-shot setting. Even in cases where our model’s predicate prediction differs from ground-truth, we can see that the detected predicate is semantically more accurate compared to the corresponding predictions of UVTransE. For example see $<\text{HAT}, \text{SLEEP ON}, \text{PERSON}>$ for UVTransE compared to our prediction of $<\text{HAT}, \text{ABOVE}, \text{PERSON}>$ for a ground-truth relationship of $<\text{HAT}, \text{ON}, \text{PERSON}>$. Overall, our proposed method can produce accurate predicate predictions even when fine-tuned with a very small number of relationship samples.

6. Conclusion

We have introduced a self-supervised learning method for visual relationship detection based on Masked Bounding Box Reconstruction (MBBR). MBBR efficiently learns rich context-aware representations that are highly predictive of object visual relationships. Our experiments demonstrate that learned representations can surpass existing methods for downstream VRD in a few-shot setting.

Acknowledgements This project has been funded by deeplab.ai, as part of its research activities, i.e., funding of student research-training and collaborations with academic institutions. Work was conducted in part, while Z.Anastasakis was an intern with deeplab.ai.
References


[38] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In CVPR, 2017. 2, 3, 5, 7


[40] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph R-CNN for scene graph generation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, ECCV, 2018. 2


[42] Ao Zhang, Yuan Yao, Qianyu Chen, Wei Ji, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. Fine-grained scene graph generation with data transfer. In ECCV, 2022. 2
