Electronic Letters on Computer Vision and Image Analysis 25(1):98-123, 2025

Optimizing Predicate Detection: A Novel Approach using ResNet152 Architecture and Nadam Optimizer

Meryem OUAZZANI CHAHDI*, Adnane OUAZZANI CHAHDI*, Afafe ANNICH*+, Khalid SATORI* and Abdellatif EL ABDERRAHMANI*

* *LISAC Laboratory, Computer of Science Department, Faculty of Sciences*
Dhar el Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco
+ *ISIC-Rabat, Higher Institute of Information and Communication.*
Avenue Allal El Fassi Madinat Al Irfane - Rabat Institutes - Morocco

Received 25th of March, 2025; accepted 9th of August 2025

Abstract

Visual relationship detection is crucial for semantic scene comprehension, impacting various fields, including Human Behavior Analysis, Visual Navigation, Medicine, and Security. A key challenge in this domain is managing partially visible objects and occluded objects, which complicate the accurate detection of triplet relationships. This study introduces a novel approach to address these challenges by employing the ResNet152 model as the backbone for the Faster R-CNN framework, paired with the Nadam optimizer – both of which have not been previously applied in this domain. Our methodology integrates region proposals and extracted features into a Graph Neural Network to construct scene graphs for each image. Additionally, we utilize a pre-trained Word2Vec model to encode subjects, objects, and predicate labels. To validate our approach, we conduct comparative studies evaluating the performance of ResNet152 against the widely utilized ResNet101 model and the EfficientNet-B7 model, which is known for its efficiency in image classification, but hasn't yet been explored in this context. We also compare the Nadam optimizer with the commonly adopted Adam optimizer. Our analysis focuses on the impact of these models on predicate detection and includes performance evaluations against current state-of-the-art methods. Experiments on the open-access VRD dataset demonstrate that our ResNet152-Nadam combination achieves the best performance across all recall metrics (R@50=89.66, R-ZR@50=71.83, R@100=94.96, and R-ZR@100=86.90 with a training time of 5h 15min 18s), underscoring the importance of model depth and optimizer selection in enhancing predicate detection. This approach shows significant potential for advancing applications in visual relationship detection, particularly in complex real-world scenarios where relationships may often be unseen.

Key Words: Semantic Scene Understanding, Visual Relationship Detection (VRD), ResNet152 Backbone, Graph Neural Networks (GNNs), Nadam Optimizer, Word2Vec Encoding.

Correspondence to: meryem.ouazzanichahdi@usmba.ac.ma

Recommended for acceptance by Angel D. Sappa
https://doi.org/10.5565/rev/elcvia.2131
ELCVIA ISSN:15108-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

1 Introduction

In recent years, the demand for sophisticated visual understanding systems has surged, driven by advancements in artificial intelligence and deep learning. Among these systems, visual relationship detection has emerged as a critical component for achieving semantic comprehension of images. This capability is essential for various applications, including robotics, surveillance, and human-computer interaction, where accurately interpreting the relationships between objects is paramount. Traditional predicate detection techniques often rely on convolutional neural networks (CNNs) [1],[2], graph neural networks (GNNs) [3],[4],[5] and recurrent neural networks (RNNs) [1],[2] to extract, label and fuse features; however, as the complexity of scenes increases, challenges such as handling partially visible objects and addressing occluded objects become increasingly prevalent. These issues arise due to overlapping objects, cluttered backgrounds, and varying object appearances, which complicate the accurate identification of relationships in dense visual scenes. There is a growing need to explore more efficient and effective models. In this research, we propose an innovative approach to enhance visual relationship detection by leveraging the ResNet152 model as the backbone of the Faster R-CNN framework, in combination with the Nadam optimizer – both of which have not been utilized in this area before. ResNet152 was specifically chosen for its deep architecture, which enables it to capture more abstract and hierarchical features compared to shallower models like ResNet101. The increased depth of ResNet152, supported by skip connections, allows it to effectively mitigate the vanishing gradient problem and extract richer feature representations from complex scenes. This makes it particularly well-suited for handling scenarios involving occluded or partially visible objects, where fine-grained details are essential for accurate relationship detection. Furthermore, ResNet152's proven performance in image classification tasks highlights its robustness in feature extraction, a critical requirement for constructing meaningful scene graphs in visual relationship detection. The Nadam optimizer was chosen due to its adaptive learning rate mechanism and Nesterov momentum, which enable it to efficiently handle sparse gradients and noisy updates. This is particularly beneficial in scenarios involving partially visible and occluded objects, as it helps stabilize training and ensures robust feature extraction. Nadam's ability to anticipate parameter updates and adjust learning rates dynamically allows for faster convergence and better generalization, addressing the challenges posed by complex scenes with occlusions and limited object visibility. Our methodology incorporates extracted features and region proposals into a Graph Neural Network, facilitating the construction of comprehensive scene graphs for each image. Additionally, we utilize a pre-trained Word2Vec model to effectively encode subjects, objects, and predicate labels, enriching the representation of relationships. To evaluate the effectiveness of our approach, we conduct extensive comparative studies contrasting the performance of ResNet152 with the widely adopted ResNet101 model [6],[7],[8],[9] and the EfficientNet-B7 model, which is known for its efficiency in image classification but has not yet been applied in this context. We also analyze the Nadam optimizer alongside the more commonly used Adam optimizer [10]. Our investigation emphasizes the impact of these models on predicate detection and includes performance assessments against the latest state-of-the-art methods. Our experimental results on the Visual Relationship Detection (VRD) dataset reveal that the combination of ResNet152 and the Nadam optimizer leads to significant improvements in predicate detection accuracy, outperforming existing techniques in the field. This study aims to provide valuable insights into model performance and optimization strategies, thereby contributing to ongoing advancements in visual relationship detection and paving the way for more robust and efficient applications in complex real-world scenarios.

2 Related Work

A large number of researchers have examined this topic with the aim of improving the quality of visual relationship detection within a scene. Since all our experiments have been conducted on the VRD dataset, we will focus on analyzing the state-of-the-art related to this dataset, as well as comparing our techniques with those already applied to it.

To address the weakness of existing approaches that expects some covert relationships between objects from

intricate scenes. Xiao et al.[8] suggested unifying knowledge graph reasoning and vision–language fusion. This involves combining external common-sense knowledge with visual feature embedding to detect the object's visual relationships. Furthermore, a subject-object proposal module was introduced to address the issue of combinatorial explosion. Experiments indicated that the suggested method surpasses the state-of-the-art techniques on the VRD open-access datasets.

As each module operates independently, the computational process encounters challenges in attaining unity and lacks logical reasonability at a higher level. In response to the aforementioned challenges, in their work, Xiao et al.[11] introduced a new technique for visual relationship detection called "affix-tuning transformers". This approach keeps the parameters of the transformer model fixed while optimizing a small, task-specific continuous vector. This reduces training costs, promotes model unity, and preserves common-sense reasoning without the need for multi-scale training. Additionally, they developed a fast framework for expressing vision-and-language sentences and trained a small number of transformer model parameters for downstream tasks.

To incorporate depth data into visual relationship detection, Gan et al.[12] introduced a VRD dataset extension named the Depth-VRD dataset and presented an adaptive depth-aware VRD module (ADVRD). They improved visual appearance by introducing a depth-aware visual combination network that integrates additional depth visual features to reinforce RGB visual features as needed. For spatial relationships, they introduced an adaptive depth spatial position technique to improve the accuracy of depth representation in object-level spatial positioning. This method utilizes the variance of regional information to assess the importance of information in every small region within the object-enclosing rectangle. The performance of the ADVRD network has been greatly enhanced by the use of depth data, as confirmed by empirical results, especially in zero-shot relationship detection tasks.

Liu et al.[13] suggested a Multimodal Similarity-Guided Relationship Interaction Network (MSGRIN), which represents visual relations as vertices to build an adaptive graph. This network improves deep message transfer through the incorporation of Multimodal Similarity Attention; Object Visual Appearance Restoration, and Object Importance Filtering, Experimental outcomes on the VRD dataset demonstrate that the MSGRIN network outperforms other approaches in terms of empirical effectiveness.

In the configurable graph reasoning (CGR) model proposed by Zhu et al.[14], the data pass through Faster R-CNN to generate region proposals and their characteristics. The subject-object bounding boxes are utilized to extract relation features. The reasoning graph operates over the knowledge graph (K.G) by matching the image's region features to entities in the knowledge graph, establishing links between knowledge from a worldwide perspective and local regions. The knowledge graph is constructed using a collection of common-sense visual relationships from image datasets containing relationship annotations, providing prior knowledge regarding semantic correlations between object and predicate nodes. Finally, the learnable configuration module determines for each relationship; whether to incorporate knowledge into the visual elements representation. Upon dynamically constructing the reasoning links, the relationships and subject-object pair labels are acquired using multilayer perceptron classifiers.

In their study, Zhou et al.[15] created a VRD-System that combines two kinds of correlation exploitation; to solve exponential growth of object pairs and the multi-label classification issues in the region pair identification and the predicate detection phases respectively. To pick the accurate and appropriate proposals; Location Rating Module (LRM) is introduced in the region-pairs identifying stage, by leveraging the spatial relationship between two objects within each pair. The Label relationship Graph Unit (LGM) was implemented during the predicate recognition stage, to assess the tacit semantic relationships between predicates. Subsequently, discrete distributed labels are assigned to predicates to enhance the top-n recall accuracy. Experiments conducted on the commonly used VRD dataset indicate that the proposed approach outperforms existing cutting-edge methods.

In [16] Chiou et al. introduced the Relational Visual-Linguistic Bidirectional Encoder Representations from Transformers (RVL-BERT) technique, which is based on executing relational reasoning by incorporating both visual and language common-sense knowledge gained through self-supervised pre-training with multimodal representations. To capture spatial details among objects, the authors incorporated in the RVL-BERT model a new masked attention module along with an effective spatial module. Additionally, the proposed model

separates relationship recognition tasks from object detection tasks by directly inputting object names, allowing it to be applied on top of any object detection system. Experiments conducted on visual relationship detection datasets validated that RVL-BERT delivered strong results compared to leading methods.

The proposed approach by Zhang et al.[9] involves extracting semantic, spatial, and visual features from image regions. Additionally, topological information between regions is integrated to improve visual relationship detection tasks. Visual features of regions are represented as vertices in a visual region graph, while the relationships between object pairs are modeled as weighted edges. Moreover, a visual relationship detection template based on regional topology construction is introduced, enabling the network to progressively combine topological structure data. The evaluation results on the VRD dataset indicate that their technique performs comparably or even surpasses the most advanced techniques according to the same evaluation criteria or state-of-the-art methods on specific evaluation metrics.

The method proposed by Wang et al. [17] employed both negative sampling and recurrent attention mechanisms to devise a rapid visual relationship detection model. This model integrates three components: visual features extracted by VGGNet, which serves as the backbone for the Faster-RCNN model; semantic embedding of object categories obtained from the Word2Vec model; and spatial features. These combined features are then processed by the Gated Recurrent Unit (GRU) to produce a weight mask. The mask determines which regions of the feature maps are output by the object detector to focus on, guiding the network to examine designated areas for prediction. The offered model is straightforward, yet experiments show its effectiveness and its capability to outperform leading results on the standard VRD dataset in most cases.

He et al. [18] introduced a novel VRD model comprising three modules: knowledge, visual, and mapping models. The translation embedding network used by the knowledge model incorporates a knowledge graph embedded in a reduced-dimensionality vector space. The visual model employs a Faster R-CNN to identify objects in images and derives their visual features. The mapping model is employed to link the knowledge space and the visual features for the final detection of the visual relationships triplets. Additionally, the network complexity was decreased by bypass-ing the redundant feature learning. The proposed VRD-Model was evaluated on the challenging VRD dataset.

With the intention of improving the zero-shot performance, Hung et al. [7] proposed a novel visual relationship detector model based on the architecture of the VTransE framework [19] and named it UVTransE, using the object pairs' union box as an additional feature to both subject and object bounding box features previously extracted by the Faster R-CNN model, for learning the predicate embedding. Extensive tests on the VRD dataset have shown that the UVTransE framework is capable of achieving state-of-the-art results in multiple challenging scenarios, including small-scale and large-scale problems, using both the full test set and zero-shot settings.

Xu et al. [20] present a framework for visual relationship detection using a novel feature representation called the Contextual Coefficients Excitation Feature (CCEF). Their approach includes extracting detected object regions from an input image and passing them through a CNN model to obtain visual features. These features are then recalibrated using contextual coefficients, which include both spatial and generated-label coefficients, to enhance their relevance. The CCEF is designed to emphasize key aspects by modeling the interdependencies between features and their context, leading to more accurate detection of relationships between objects. Additionally, the authors enhance the conditional Wasserstein Generative Adversarial Network (WGAN) by incorporating a relationship classification loss to address the long-tail distribution issue in relationship data. This improvement significantly boosts the detection of un-seen relationships, as evidenced by an increase in recall from 8.5% to 23.2% in zero-shot settings.

In their proposed model, Li et al. [21] first used the Faster R-CNN as an object detector model to detect the object's features in the image. Second, the contextual information was extracted for each object pair by fusing visual and spatial features to generate the object's enhanced features. Further, the Gated Recurrent Unit(GRU) was employed instead of Long Short-Term Memory (LSTM) to aggregate a word with contextual information and encode the long sentence. Finally, these features were fed into the relationship inference network, to generate the triplet relationship (subject, predicate, object) by choosing the highest-scoring category. The

model's efficiency was evaluated on the commonly used VRD dataset.

Liu et al. [22] submitted the RGB-Depth-Balanced Network (RDBN), which com-bines three models: the Appearance feature module based on the Faster R-CNN as an object detector that uses the VGG16 as a back-bone to extract region proposals, the objects' flat RGB appearance features, and the depth feature maps. The spatial feature module produces flat spatial features and depth spatial features, respectively by incorporating the original image and depth image previously provided in the first stage of the RDBN. Finally, the appearance and spatial features were concatenated by the Fusion-Balance Module for the final classification of the relation-ship triplets. The RDBN demonstrates that it can surpass state-of-the-art VRD techniques on the VRD dataset, particularly in zero-shot relationship detection and image retrieval tasks.

Zhan et al. [23] suggested a novel multi-task compositional network (MCN) meth-odology based on the fusion of multimodal attributes, encompassing visual, semantic, relative-spatial, and absolute-spatial features to detect significant relationships instead of the dominant insignificant relationships. The MCN subtask is divided into three models. The first one is VGG-16, which was used in Faster R-CNN as an object de-tector network with the RPN to perform 300 region proposals per image. The second is a relationship prediction network that was adopted to generate an initial relationship set that includes human-labeled and unlabeled relations using a new correlated loss function. The final Relationship Predictor network fuses different features previously extracted from the first and the second steps using a multimodal factorized bilinear pooling approach (MFB) to detect the significant relationships.

Leveraging a modified Bidirectional Recurrent Neural Network (BRNN); Dai et al. [24] presented a new approach incorporating four steps to concurrently classify ob-jects while detecting the predicates. In the object detector step, the author used the Faster R-CNN with the modified VGGNet as the backbone to detect objects, bound-ing boxes and visual features. These features served as input to the Region Proposal Networks (RPN) to extract the region proposals. In the Pair filtering step, the Paired non-maximum suppression technique is applied to address the object-pairs combina-torial explosion problem by selecting appropriate regions and eliminating redundant ones. The Joint Recognition stage fuses multiple previously extracted and filtered cues including visual, spatial, and classification features. Finally, a Bidirectional RNN is used to predict and classify the triplet visual relationships. The experimental results conducted on Visual relationship Dataset (VRD) demonstrate that the proposed meth-od yields competitive results compared to recent works in the visual relationship de-tection tasks.

In their article, Liang et al. [25] presented a novel Deep Structural Ranking Framework which combines VGG16 as a deep convolutional network for object ex-traction, and a structural ranking loss approach for visual relationship classification to improve the relevance scores of annotated relationships. They also combined multi-ple cues such as the union bounding boxes, visual appearance attributes, and spatial position indication, to enhance the representational capacity of visual appearance features and predict relationships depicted within the image. Empirical outcomes demonstrate that the proposed framework outperforms state-of-the-art methods on the commonly used VRD dataset, particularly in zero-shot relationship detection tasks.

The visual relationship extraction system proposed by Dai et al. [6]; employed Faster RCNN as an object detector and appearance feature extractor for objects. The statistical relations, the local regions, and the spatial features of object pairs were also extracted and passed, along with the appearance features, through the Deep Relational Network (DR-Net) which analyzes all extracted elements in the first step to calculate the category probabilities for every triplet relationship elements denoted as (subject, predicate, object). Ultimately, the triplet is obtained by selecting the highest probability score category for each element. These empirical results obtained on the VRD dataset indicate that the incorporation of statistical relations improves visual relation-ship detection tasks. Furthermore, the findings demonstrate a strong capability of the suggested DR-Net framework in modeling complex relations.

Zhang et al. [19] proposed the first end-to-end visual relational inference network named the Visual Trans-lation Embedding model (VTransE) inspired by contemporary advancements in CNNs and knowledge graph embedding models, which combine two models: the CNN as an object extraction network, and the Relation Prediction net-work, which positions the objects in a low-dimensional translation vector space to denote the

triplet relationship using the equation (object = subject + predicate). To support simultaneous training and inference in a single forward and backward pass; a novel feature extraction layer was introduced within the relation prediction model to transfer the object-relation knowledge in a fully convolutional manner. The VTransE efficacy is demonstrated through its superior performance compared to recent methodologies on an extensive dataset: the VRD (Visual Relationship Detection) dataset.

To enhance visual relationship detection, Plesse et al. [26] introduce learning prototypes that represent diverse relational meaning in images. This method trains a network to organize relationship triplets (subject, relation, object) within a feature space where triplets with similar spatial layouts, interactions, and relational meaning are clustered. By leveraging prototype-based classification and cosine similarity during testing, this approach enhances the diversity of detected relationships, as demonstrated on the VRD benchmark dataset, achieving notable improvements in Average Recall.

Other research studies in 3D feature extraction have demonstrated significant potential for enhancing visual relationship detection by improving spatial understanding within complex scenes. Fast depth map fusion techniques now enable real-time, high-resolution scene reconstruction using commodity depth cameras, particularly beneficial for indoor scene analysis. These advancements allow for detailed contextual modeling of both intra-object part relationships and inter-object spatial arrangements [27]. A unified framework for detecting individual objects and groups leverages contextual information, improving object segmentation and recognition even when objects are unknown. Similarly, in traffic surveillance, integrating depth information with RGB data has been shown to enhance object detection and recognition accuracy [28]. Depth information enables the separation of overlapping objects and the extraction of more discriminative spatial features, thus improving the detection of critical traffic events such as overtaking warnings and collision avoidance.

Furthermore, three-dimensional reconstruction techniques developed for 3D modeling and pose estimation also provide promising avenues for VRD by accurately positioning objects in 3D space. For example, self-calibration and bundle adjustment techniques can refine object placements, ensuring precise depth and positional accuracy [29]—essential for VRD applications in robotics, surveillance, and human-computer interaction. Together, these studies underscore the importance of integrating 3D features into VRD systems, as multi-dimensional scene analysis significantly enhances the capacity to understand and interpret complex spatial relationships.

Recent advances in video-based visual relationship detection (VidVRD) have faced challenges related to scalability and generalization due to the limitation of monolithic neural networks. Addressing these, the authors in [30] proposed MoE-VRD, a framework based on a sparsely-gated Mixture of Experts (MoE) architecture. This approach distributes the detection task across multiple specialized expert networks, enhancing both performance and computational efficiency. Each expert focuses on recognizing visual relationships expressed as subject, predicate, object triplets, enabling more robust recognition of complex spatio-temporal interactions in videos. MoE-VRD achieves superior results on benchmark datasets such as ImageNet-VRD and VidOR, surpassing many state-of-the-art approaches. The authors also discuss potential improvements, such as task-specific expert specialization and more sophisticated gating mechanisms to optimize expert selection. This work significantly contributes to ad-dressing the challenges of visual relationship detection in dynamic video environments through the use of ensemble learning strategies.

Building on recent progress in transformer-based architectures, Mingcheng Qu et al. [31] introduced the Dual Attentional Transformer Network (VRD-DAT) for video visual relation detection. VRD-DAT simultaneously models action predicates and spatial relations using two parallel transformer branches –Act-T and Spa-T- enabling a fine-grained understanding of temporal interactions and spatial configurations in videos. An attentional weighting module is then employed to fuse these complemen-tary relation cues, resulting in highly accurate visual relation predictions. Experiments conducted on the ImageNet-VidVRD and VidOR datasets show that VRD-DAT achieves state-of-the-art performance. This highlights the effectiveness of transformer-based models in capturing complex multimodal dependencies for visual relation tasks in videos, and showcases their potential to improve downstream applications that require accurate video understanding.

Drawing inspiration from cognitive neuroscience, Safaa Abdullahi et al.[32] proposed HIPA, a hierarchical

reasoning framework for visual question answering (VQA) based on the Perception-Action Cycle (PAC). In contrast to conventional attention-based models, HIPA emulates human learning by modeling the interaction between vision and language through a multi-phase reasoning process. Visual information is processed through object-level attention, organization, and interpretation, while linguistic input is handled via word-level attention, interpretation, and conditioning. These modalities are then jointly interpreted in a cyclic and interdependent manner, mimicking PAC dynamics in the human brain. Evaluations on the TDIUC, VRD, and VQA 2.0 datasets confirm the effectiveness of this biologically inspired framework, which improves both the interpretability and accuracy of VQA systems by aligning with principles of human cognition.

In parallel, recent studies have explored the integration of deep learning and zero-shot learning techniques for dynamic knowledge graph construction. Lu He and Xu HU [33], for instance, proposed an approach combining semantic embedding vectors, Graph Convolutional Networks (GCNs), and multi-modal entity extraction frameworks such as Video2Entities and knowledge to recognize previously unseen object classes, actions and human-object interactions. Results demonstrate significant improvements in zero-shot tasks, with accuracy reaching 77% in object detection and 86% in action recognition. This body of work underscores the importance of semantic fusion and zero-shot generalization for enhancing the scalability and adaptability of knowledge graph, particularly in complex domain like construction, where new entities and relations constantly emerge.

To further advance knowledge graph completion (KGC), Yifan Xue et al. [34] proposed ChoicE, a novel multimodal framework that addresses two fundamental challenges: modal heterogeneity and spatial heterogeneity. ChoicE employs a Mixture of Experts within encoder-decoder architecture. The encoder uses a modality chooser to process structural, textual, and visual inputs independently, producing specialized embeddings that preserve modality-specific information. In the decoder, a geometric chooser dynamically maps the unified representations into Euclidean, complex, or hyperbolic spaces, depending on the inference patterns required by each query. Extensive experiments on benchmark datasets - including WN18RR++, FB15K237, and VTKG- demonstrate that ChoicE outperforms 13 state-of-the-art baselines. This work emphasizes the necessity of modality-aware processing and geometry-aware embedding for building scalable and accurate multimodal KGC systems, providing valuable insights into the integration of vision and language for enriched knowledge representation.

3 Approach

Visual relationship detection aims to capture and interpret the relationships or associations between objects within a scene. This process entails recognizing various connections, including visual, contextual, and positional dependencies, between subject-object pairs (Fig. 1), such as ”next to”, ”inside”, or more intricate relations like ”wearing” and ”playing”. This task classifies relationships into categories, enriching scene comprehension by forming triplets like (boy, catching, ball). Visual relationship detection involves two main stages: predicate detection and relationship prediction. The first stage identifies the specific action or connection between the subject and the object. The second stage uses the information from the first stage to categorize the relationship triplets between the subject-object pairs. The success of predicting triplets (subject, predicate, object) relies heavily on the effectiveness of predicate detection, making high accuracy in this stage essential for successful relationship prediction and comprehensive scene understanding.

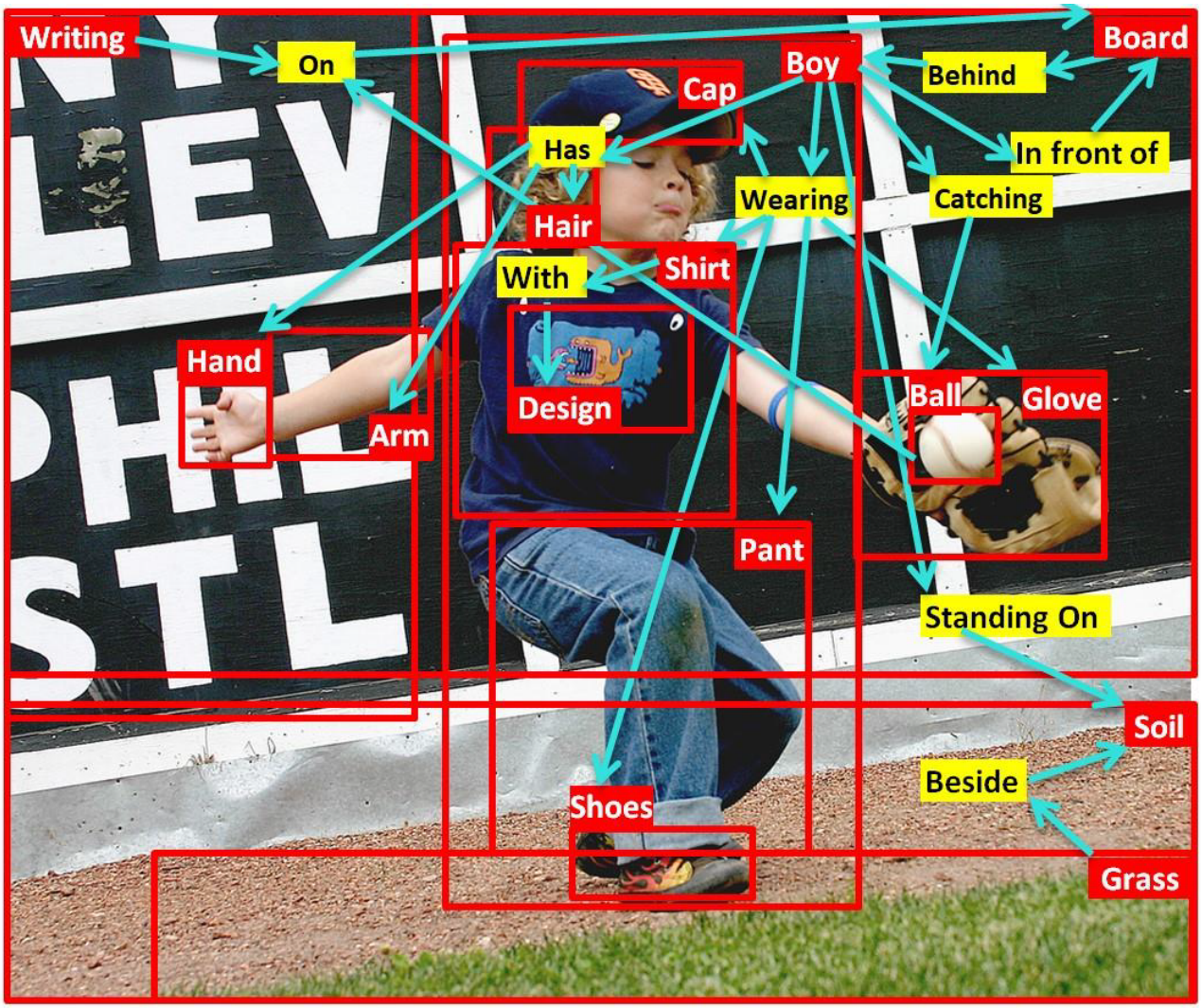

Figure 1: Illustration of visual relationships in images from the VRD dataset.

Given the critical role of predicate detection in improving visual relationship detection quality, this paper presents a novel method that leverages the ResNet152 model (Fig. 4) combined with the Nadam optimizer—both newly applied in this field. To validate our approach, we conduct comparative studies evaluating the performance of ResNet152 against the widely used ResNet101 model (Fig. 3), and the EfficientNet-B7 (Fig. 2) model, recognized for image classification efficiency but previously unexplored in this context. Additionally, we compare the Nadam optimizer with the commonly used Adam optimizer. Our analysis supported by performance comparisons with current state-of-the-art methods, highlights the impact of our approach on predicate detection, compared with these models, and the recent approaches. The method we propose, as shown in (Fig. 7) includes four main steps:

3.1 Image Pre-processing Step

- Image resizing;
- Normalization: we normalize the images using the mean and standard deviation of the dataset previously calculated using the following formulas:

$$\mu_c = \frac{1}{N_c} \sum_{i=1}^{N_c} x_{i,c} \tag{1}$$

$$\sigma_c = \sqrt{\frac{1}{N_c}\sum_{i=1}^{N_c}(x_{i,c} - \mu_c)^2} \tag{2}$$

$$\text{img}_{\text{normalized}} = \frac{\frac{\text{img}_c}{255} - \mu_c}{\sigma_c} \tag{3}$$

Where the mean and standard deviation for image data are computed across the pixel values for each channel (Red, Green, Blue) throughout the entire dataset and: μ_c denotes the mean for channel c. N_c denotes the total number of pixels in the c-th channel across all images. $x_{(i,c)}$ denotes the pixel value of the i-th pixel in the c-th channel. σ_c denotes the standard deviation for channel c. img_c Denotes the original image value ,in the channel c.

- Calculate the scaling coefficients for the image width and height separately to adjust the bounding boxes to maintain their correct positions and dimensions relative to the new image size.

$$\text{scale}_{\text{width}} = \frac{\text{new_width}}{\text{original_width}} \tag{4}$$

$$\text{scale}_{\text{height}} = \frac{\text{new_height}}{\text{original_height}} \tag{5}$$

3.2 Features extraction step:

In our approach, we focus on extracting three types of features - appearance, contextual, and relative location features- from the previously annotated objects in the Visual Relationship Detection (VRD) dataset. We utilize the pre-trained EfficientNet-B7 (Figure 1), the pre-trained ResNet101, or the pre-trained ResNet152 -each trained on the large-scale ImageNet dataset- as the backbones for Faster R-CNN. This enables us to examine the impact of different backbone architectures on network performance.

- The appearance-based subject-object features: The image data are processed through the layers of the pre-trained backbones to extract the feature maps:

$$X = \text{ReLU}\left(W_{\text{backbone}}^{\top} \cdot \text{Img}_{(w\times h\times 3)} + b\right) \tag{6}$$

Where W represents the pre-trained weights of the backbone network, which can be ResNet152, ResNet101, or EfficientNet-B7, b is the bias term, and ReLU is an activation function. The feature maps, along with the annotated object boxes, are fed into the RoI-pooling function to output fixed-size feature maps (in this case 14×14) $X_{so} = \{x_{so,1}, \ldots, x_{so,n}\} \in \mathbb{R}^{C\times 14\times 14}$ from regions of interest (RoIs), where C denotes the number of output channels for each backbone model, and $1.0/16$ is the spatial scale factor that scales the bounding boxes in the original images to $(x_{i1}/16, y_{i1}/16, x_{i2}/16, y_{i2}/16)$ with (i) denoting the box number:

$$X_{so} = \text{RoIPool}\left((14,14), \frac{1.0}{16}\right)(X, \text{boxes}_{so}) \tag{7}$$

Finally, these feature maps are passed through an AdaptiveAvgPool2d layer followed by a flattening layer, in order to capture high-level features, optimize computational efficiency, and generate reduced-dimension proposal region vectors $X_{so} \in \mathbb{R}^C$:

$$X_{so} = \text{nn.AdaptiveAvgPool2d}\left(\text{output_size} = (1,1)\right)(X_{so}) \tag{8}$$

The target output size (1×1) indicates that the dimension of each channel will be reduced to a single value instead of (14×14). For each channel c in X_{so}, the Adaptive Average Pooling computes the

average value over the entire spatial dimensions. The average for channel c is given by:

$$X_{so}[c,0,0] = \frac{1}{14 \times 14} \sum_{h=0}^{13} \sum_{w=0}^{13} X_{so}[c,h,w] \tag{9}$$

Where $X_{so}[c,h,w]$ is the value of the feature map at the channel c and the spatial location (h,w). The flattening operation aims to convert a multi-dimensional tensor into a one-dimensional tensor.

$$X_{so} = \text{torch.flatten}\,(X_{so}) \tag{10}$$

The input tensor $X_{so} \in \mathbb{R}^{C \times 1 \times 1}$ has shape $(C \times 1 \times 1)$ which means it contains C separate elements, each corresponding to a single channel. When applying torch.flatten operation, the function transforms the tensor into a single vector with C elements.

- Appearance relation features: In line with the previous steps, the X feature maps, along with the relation boxes, are fed into the RoI-pooling layer to extract relation feature maps with unified dimensions.

$$X_{\text{rel}} = \text{RoIPool}\left((14,14), \frac{1.0}{16}\right)(X, \text{boxes}_{\text{rel}}) \tag{11}$$

With $X_{\text{rel}} \in \mathbb{R}^{C \times 14 \times 14}$.

The relation boxes (14) are determined by calculating the union box of the subject and the object bounding boxes using the formulas (12), and (13):

$$\text{corner}_{\text{Top-left}}(ux_1, uy_1) = (\min(sx_1, ox_1), \min(sy_1, oy_1)) \tag{12}$$

$$\text{corner}_{\text{Bottom-right}}(ux_2, uy_2) = (\max(sx_2, ox_2), \max(sy_2, oy_2)) \tag{13}$$

The relation box can then be defined as:

$$\text{box}_{\text{rel}} = (ux_1, uy_1, ux_2, uy_2) \tag{14}$$

With $S_{\text{box}} = [sx_1, sy_1, sx_2, sy_2]$ and $O_{\text{box}} = [ox_1, oy_1, ox_2, oy_2]$ representing the coordinates of the subject and object boxes, respectively. The relation feature maps are then pooled and flattened similarly using the following formulas:

$$X_{\text{rel}} = \text{nn.AdaptiveAvgPool2d}\,(\text{output_size} = (1,1))\,(X_{\text{rel}}) \tag{15}$$

$$X_{\text{rel}} = \text{torch.flatten}\,(X_{\text{rel}}) \tag{16}$$

This produces the final relation feature vectors $X_{\text{rel}} \in \mathbb{R}^C$.

- Combining Features: Both X_{so} and X_{rel} are passed through a fully connected layer followed by a ReLU function, to produce low-dimensional visual feature vectors $X'_{so} \in \mathbb{R}^{256}$ and $X'_{\text{rel}} \in \mathbb{R}^{256}$. These are then concatenated along the batch dimension (dim 0) to produce a 512-dimensional feature vector:

$$V_f = \begin{bmatrix} X'_{so} \\ X'_{\text{rel}} \end{bmatrix} \in \mathbb{R}^{512} \tag{17}$$

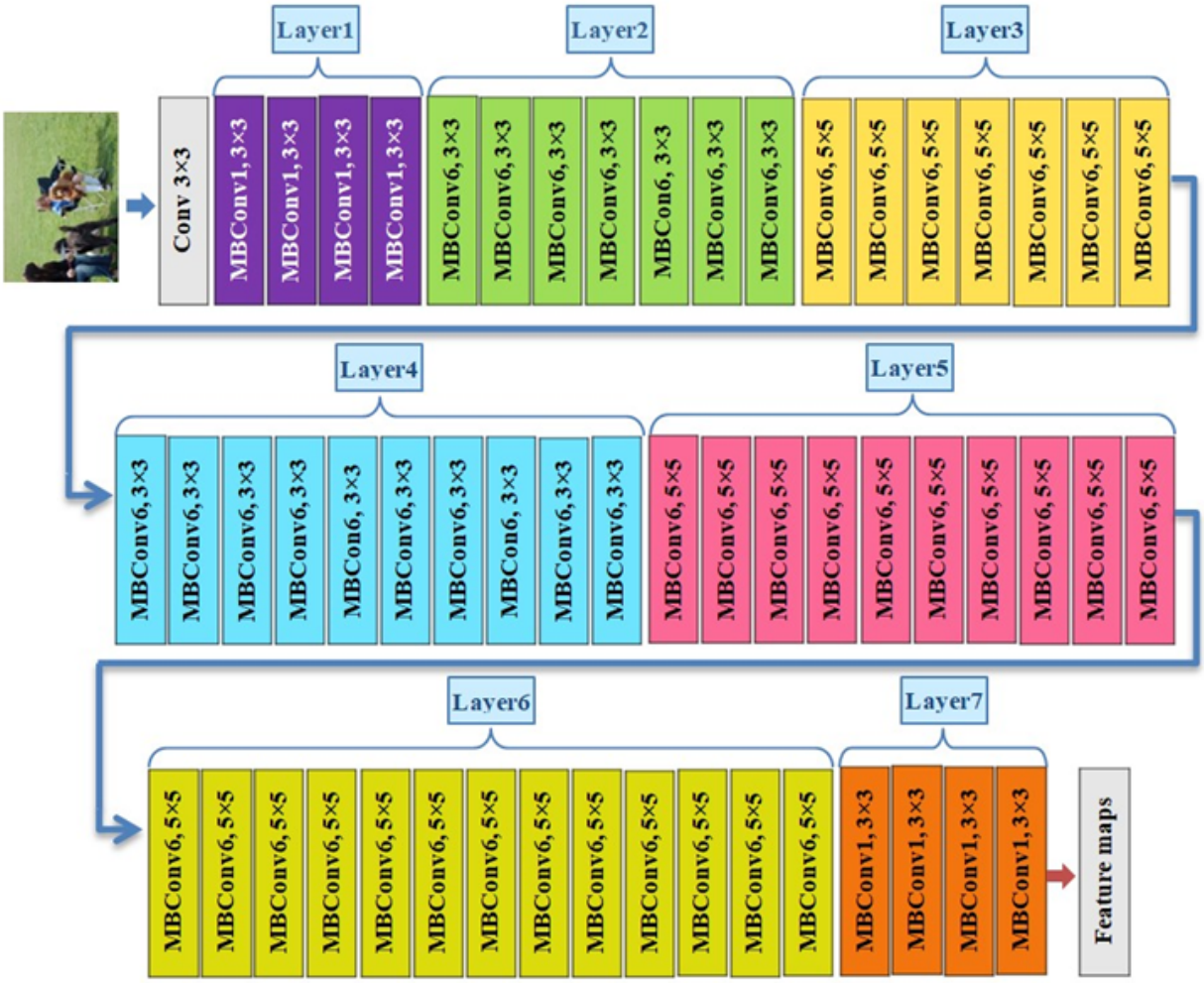

Figure 2: EfficientNet-B7 Architecture

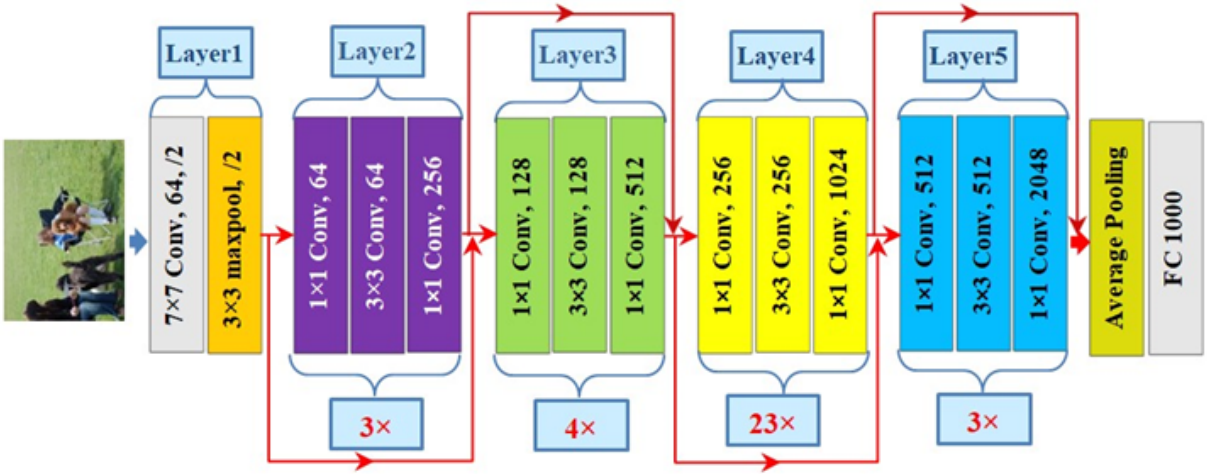

Figure 3: ResNet101 Architecture

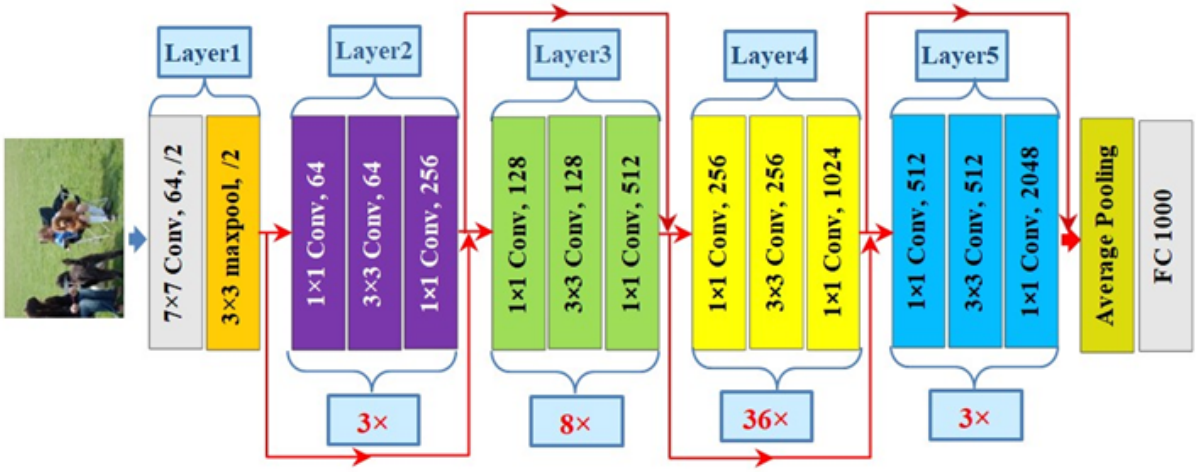

Figure 4: ResNet152 Architecture

- The contextual features: To provide word embedding vectors (Fig. 5) for subject, object and predicate labels within an image, we use in our approach the Word2Vec, a pre-trained model created by Google that offers 300-dimensional vectors for words. It is widely used in various natural language processing (NLP) tasks, including word similarity and analogies. Each word embedding can be represented as a vector $V_w \in \mathbb{R}^{300}$ (18).

$$V_w = [v_0, v_1, \ldots, v_{298}, v_{299}] \quad \text{with} \quad -1 < v_i < 1 \tag{18}$$

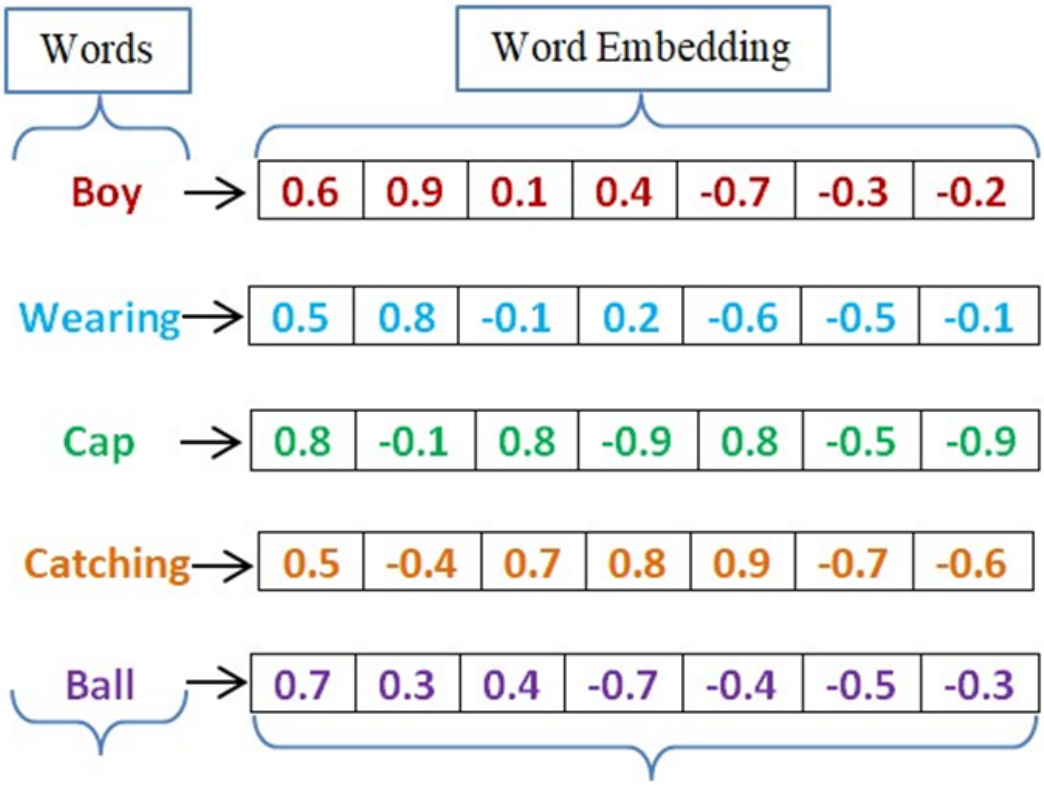

Figure 5: Word Embedding Illustration

- Spatial features: These refer to characteristics associated with the positions and geometrical configurations of objects within a scene. Typically, these features include relative positions, distances between objects, orientations, and spatial context. In our approach, we integrate relative coordinates that represent the spatial relationship context between the subject's bounding box and the object's bounding box, scaled by the size of each box. This normalization aids in generalizing spatial relationships independently of the actual sizes. We also integrate the relative size between the subject and object bounding boxes, based on their respective widths and heights. Given S_{box} and O_{box}, which denote the bounding box coordinates of the subject and the object, respectively:

$$S_{\text{box}} = [s_{x_1}, s_{y_1}, s_{x_2}, s_{y_2}] \tag{19}$$

$$O_{\text{box}} = [o_{x_1}, o_{y_1}, o_{x_2}, o_{y_2}] \tag{20}$$

The relative coordinates can be calculated using formulas (21) and (22):

$$\mathbf{A} = \begin{bmatrix} \frac{s_{x_1} - o_{x_1}}{o_w} \\ \frac{s_{y_1} - o_{y_1}}{o_h} \end{bmatrix} \tag{21}$$

$$\mathbf{B} = \begin{bmatrix} \frac{o_{x_1} - s_{x_1}}{s_w} \\ \frac{o_{y_1} - s_{y_1}}{s_h} \end{bmatrix} \tag{22}$$

$$\begin{cases} s_w = s_{x_2} - s_{x_1} \\ s_h = s_{y_2} - s_{y_1} \end{cases} \tag{23}$$

$$\begin{cases} o_w = o_{x_2} - o_{x_1} \\ o_h = o_{y_2} - o_{y_1} \end{cases} \tag{24}$$

A represents the relative x and y coordinates of the subject's top-left corner with respect to the object's width and height. B represents the relative x and y coordinates of the object's top-left corner with respect to the subject's width and height. Here, s_w and s_h in (23) denote the width and height of the subject box, respectively, while o_w and o_h in (24) denote the width and height of the object box, respectively. The relative ratio can be calculated using the logarithmic function (25):

$$\begin{bmatrix} \log\left(\frac{s_w}{o_w}\right) \\ \log\left(\frac{s_h}{o_h}\right) \\ \log\left(\frac{o_w}{s_w}\right) \\ \log\left(\frac{o_h}{s_h}\right) \end{bmatrix} \tag{25}$$

All these coordinates are stored in a single spatial feature vector $V_{\text{spa}} \in \mathbb{R}^8$, which is passed through a fully connected layer, followed by a ReLU activation function to produce a 256-dimensional spatial feature vector $V_s \in \mathbb{R}^{256}$.

3.3 The scene graph generation step:

Using prior knowledge about subject-object co-occurrences, we compute the relationship indices and edge weights between subjects and objects. These, along with the calculated feature vectors, are passed through the Graph Convolutional Neural Network to generate a weighted digraph (Fig. 6) for the input image. Where:

- The nodes in the graph represent the objects, the subjects, and the predicates of the triplet relationships,
- The content of the nodes includes the subject, the object and the predicate semantics, spatial, and visual features which are extracted from the Word2Vec network and the CNN backbone (EfficientNet-b7, ResNet101, or ResNet152).
- The edges in the graph represent the connections between pairs (object-predicate) and contain the occurrence probability for each pairwise connection. This probability is calculated based on prior knowledge of subject-object co-occurrence, adjusted for specific relationship probabilities between object types and predicates (e.g., "boy, catching, ball")
- The edge direction $(S \rightarrow P \rightarrow O)$ signifies that the subject S node interacts with the object node O through the predicate P (e.g., $boy \rightarrow catching \rightarrow ball$).

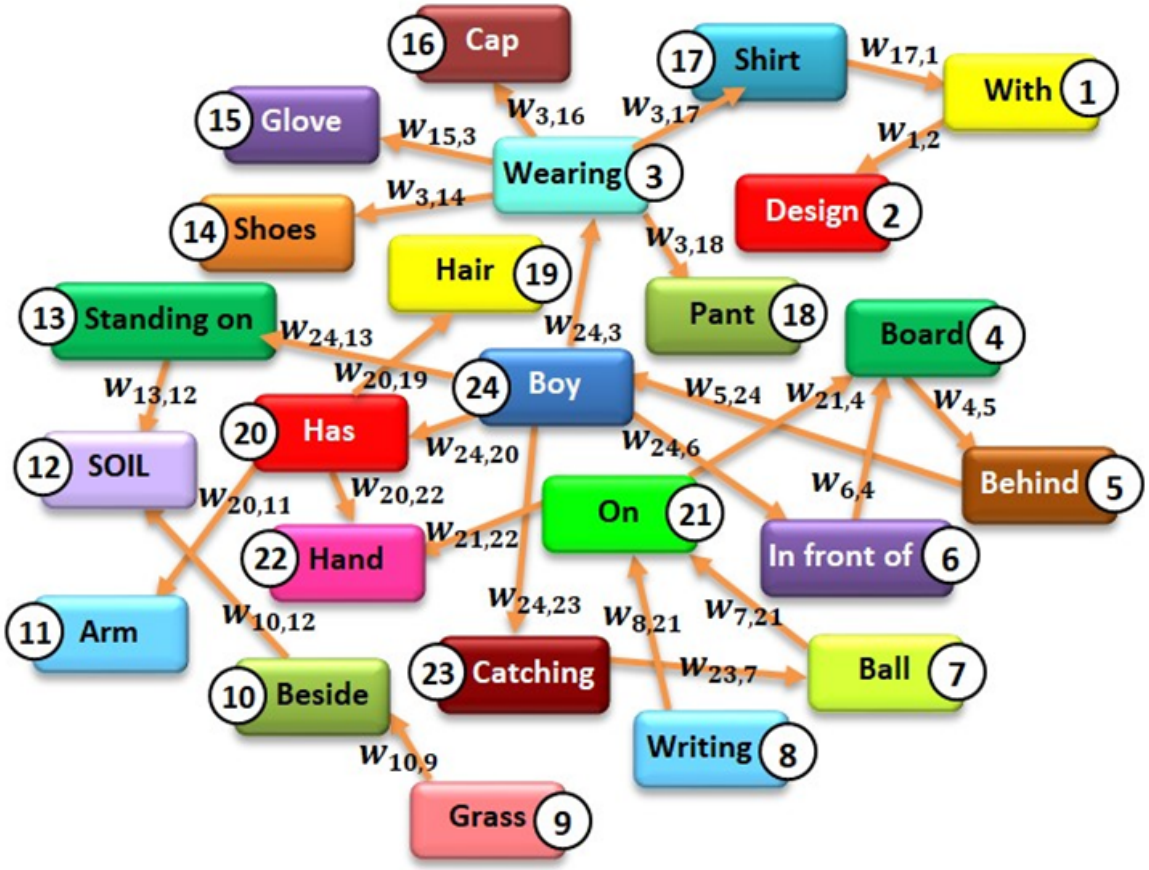

Figure 6: Scene Graph Visualisation

3.4 The Predicate Classification step:

To classify the identified predicates, we use the ranking loss function [25] which is more flexible and better suited for multi-label problems. Unlike traditional loss functions that typically distinguish between positive and

negative relationships (i.e. binary classification), ranking loss allows multiple valid predicates to be associated with a pair of objects rather than enforcing a single correct answer. This function focuses on correctly ordering predicates in a list based on criteria, such as relevance, similarity, or importance, thereby prioritizing the most relevant predicate, rather than solely minimizing classification error.

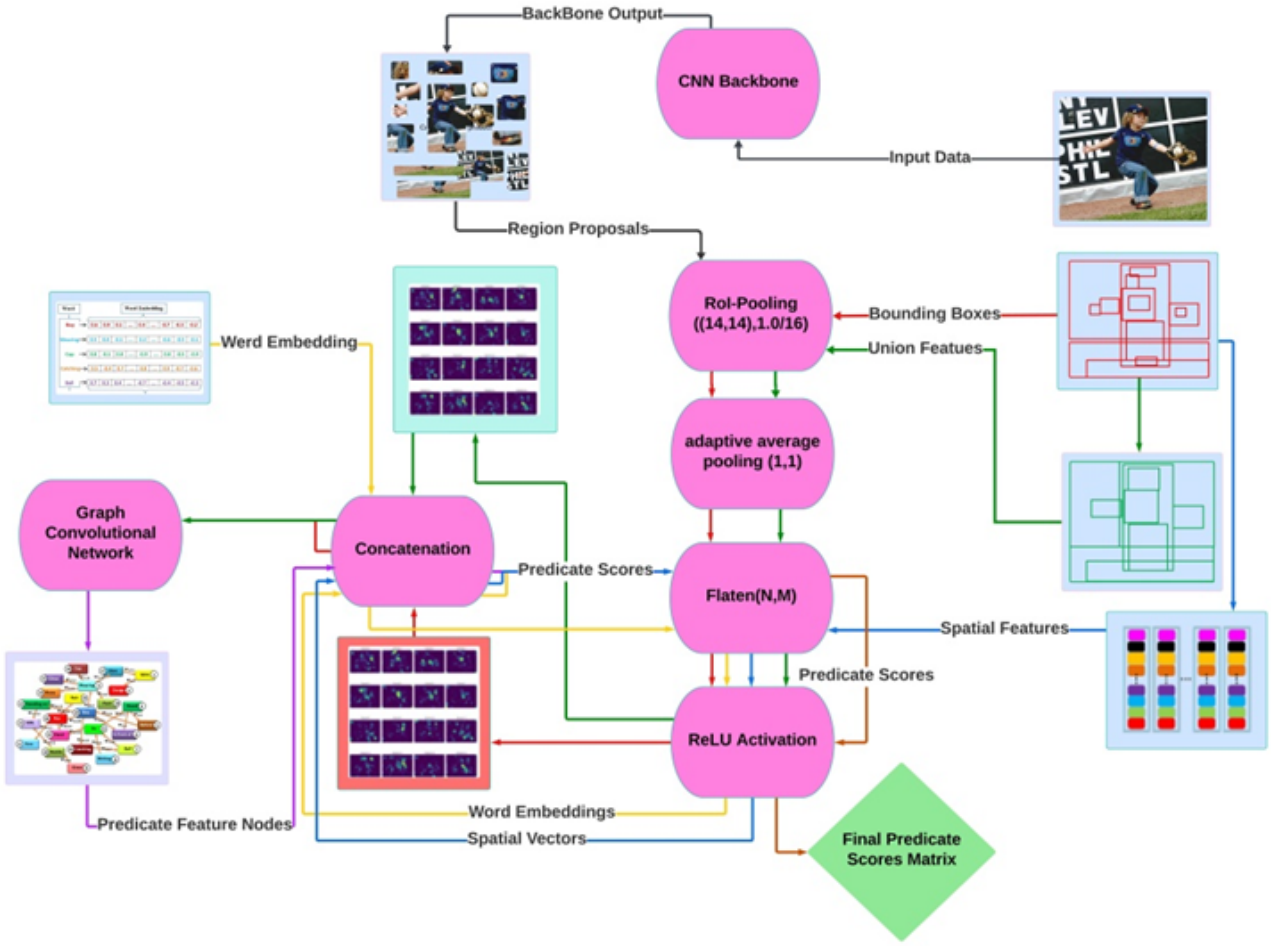

Figure 7: Predicate Detection Approach

Detecting all relationships within an image is challenging for models due to the skewed distribution of relationship instances. This necessitates models capable of zero-shot predicate detection, where relationships between object categories not seen in training can still be recognized. Since zero-shot detection requires generalization to unseen relationships, its performance may be lower than that of conventional relationship detection methods trained on specific relationships. Many factors influencing zero-shot detection performance include relationship complexity, the quality and diversity of training data, feature representation choices, and the effectiveness of the learning algorithm. In this work, we assess the performance of the approach based on three backbone architectures to detect various predicates within visual elements, with a particular emphasis on its performance in identifying unseen predicates.

4 Experimental results

4.1 Dataset

Our approach was examined on the Annotated Visual Relationship Detection (VRD)[35] dataset, one of the most widely used datasets in this area. The VRD dataset consists of 5000 images, with 4000 images dedicated

for training, and 1000 images for testing. It encompasses 70 predicate categories, 100 object categories, and over 37993 relationships. This dataset has been enhanced with detailed annotations, including:

- **Object annotations:** Labeled objects with bounding boxes and descriptions.
- **Attribute annotations:** Additional attributes for objects (e.g., color, size).
- **Relationship annotations:** Relationships between objects (e.g., "person holding racket").
- **Region descriptions:** Descriptions of specific regions in images.

These annotations make the dataset a valuable resource for tasks such as object detection, scene understanding, relationship extraction, and semantic analysis.

4.2 Workstation characteristics:

The approach is developed using Anaconda Navigator, the Spyder editor, Python version 3.10.10, and various libraries including PyTorch v2.0.0, torch_geometric v2.3.0, albumentations v1.3.0, and tabulate v0.9.0. The working environment is based on a Dell Precision 7920 computer outfitted with an Intel(R) Xeon(R) Platinum 8170M CPU processor (104 cores) operating at a base clock speed of 2.10 GHz, and 786432 MB of memory. Additionally, the system includes three NVIDIA Quadro P6000 graphics cards.

4.3 The module's parameters:

Here is a detailed description of the module's parameters:

- **The VRD's mean:** [0.4565, 0.4349, 0.4106]
- **The VRD's standard deviation (std):** [0.2494, 0.2448, 0.2497]
- **Use Object Prior:** Set to True. This indicates that prior knowledge about objects is used, allowing the model to better anticipate relationships based on the objects present in an image.
- **Learning Rate (LR):** The learning rate is set to 0.0001.
- **Learning Gamma (lr_gamma):** The learning gamma is set to 0.1
- **Learning Steps:** The learning rate adjustment occurs at epoch 3
- **Graph Model:** the model utilizes Gated Graph Convolution with an output channel size of 256, and comprises two layers
- **Training session:** The current training session is set to 5
- **Number of Epochs:** The total number of training epochs is set to 12
- **Bach Size:** Set to 1

4.4 Comparative Study: Impact of our Approach (ResNet with Nadam) on Predicate Detection tasks

This section presents a detailed comparative analysis of our approach with various models and optimizers for visual relationship detection, focusing on predicate detection performance (Table 1). The models evaluated include **ResNet152**, **ResNet101**, and **EfficientNet-B7**, paired with two optimizers: **Adam** and **Nadam**. The primary performance metrics considered are **Recall at 50 (R@50)** and **Recall at 100 (R@100)**, which measure the accuracy of correct visual relationships detected within the top 50 and top 100 predictions, respectively. **Zero-shot Recall at 50 (R-ZR@50)** and **Zero-shot Recall at 100 (R-ZR@100)** assess the model's ability to detect relationships that were not present in the training set, thus testing generalization on un-seen data. Additionally, we examine the impact of each approach on training time.

Table 1: Predicate Detection task performances on the VRD dataset

Models	R@50	R@100	R-ZR@50	R-ZR@100	Training Time
ResNet101–Adam	89.4785	94.8244	70.9760	86.6438	04h 11min 22s
ResNet101–Nadam	89.4916	94.9292	71.3185	86.6438	04h 11min 45s
EfficientNet-B7–Adam	89.5702	94.8768	71.4897	86.4726	15h 42min 22s
EfficientNet-B7–Nadam	89.6619	94.9030	71.5753	86.6438	15h 41min 01s
ResNet152–Adam	89.3606	94.7458	70.1199	86.5582	05h 16min 17s
Our (ResNet152–Nadam)	**89.6619**	**94.9555**	**71.8322**	**86.9007**	**05h 15min 18s**

Overall Performance on Recall Metrics (R@50 and R@100)(Fig. 8)

- As shown in (Table 1) and (Fig. 8), **our method (ResNet152–Nadam)** achieves the highest **R@50** score of **89.6619**, outpacing ResNet152–Adam (**89.3606**), ResNet101–Adam (**89.4785**), and EfficientNet-B7–Adam (**89.5702**).
- **Our combination (ResNet152–Nadam)** also leads with an **R@100** score of **94.9555**, surpassing ResNet152–Adam (**94.7458**), ResNet101–Nadam (**94.9292**), and EfficientNet-B7–Nadam (**94.9030**).

This indicates that **our contribution** (**ResNet152-Nadam**) significantly enhances recall at ranks 50 and 100 by capturing more predicates, showcasing the impact of this model combination on detecting visual relationships, which is essential for comprehensive scene understanding.

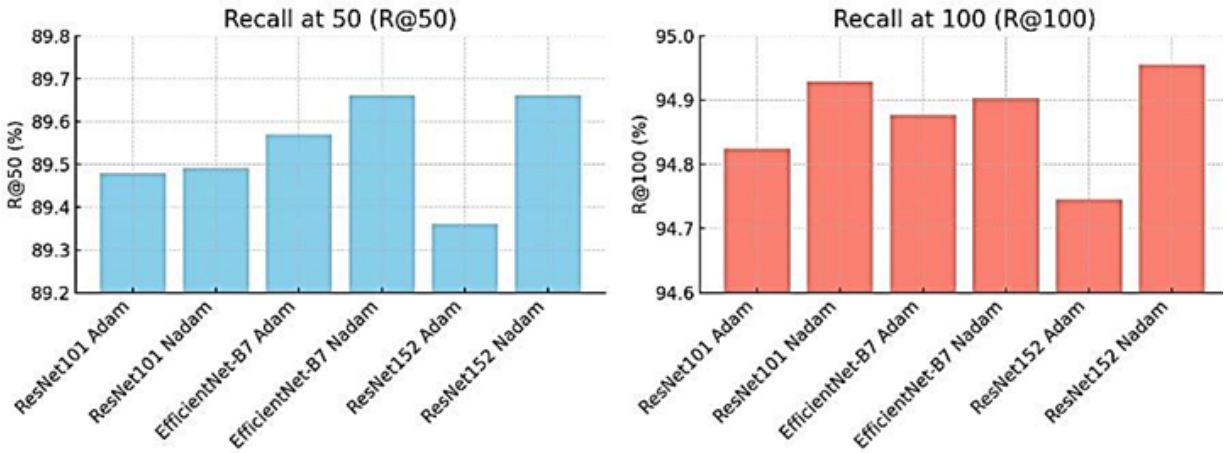

Figure 8: Comparative Model Performance on Recall@50/Recall@100 Metrics (VRD Dataset)

Performance on Zero-shot Recall Metrics (R-ZR@50 and R-ZR@100) ((Fig. 9))

- Our approach (ResNet152-Nadam) again demonstrates superior performance on R-ZR@50 with (71.8332), higher than ResNet152-Adam (70.1199), ResNet101-Nadam (71.3185), and EfficientNet-B7-Nadam (71.5753).

- Our method also leads with a R-ZR@100 of (86.9007), slightly outperforming EfficientNet-B7-Adam (86.4726) and both EfficientNet-B7-Nadam (86.6438) and ResNet101-Nadam (86.6438).

The improvement in zero-shot performance indicates the ability of our approach (ResNet152-Nadam) to detect relationships that were not explicitly seen during train-ing. This reinforces the idea that ResNet152, when paired with Nadam enhances both traditional and zero-shot predicate detection capabilities, making it more adaptable to diverse real-world settings.

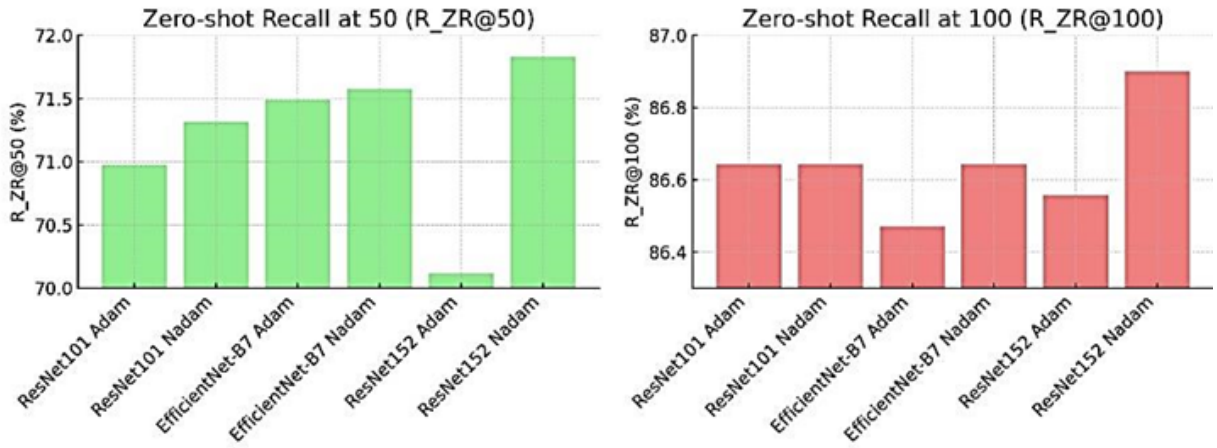

Figure 9: Comparative Zero-Shot Metric Performance of Models on VRD Dataset

Training time analysis (Fig. 10): Our method (ResNet152-Nadam) demonstrates an efficient training time of 5 hours and 15 minutes (ResNet152-Adam takes 5 hours and 16 minutes), which is significantly faster than EfficientNet-B7-Nadam at 15 hours and 41 minutes. The reduced training time for ResNet152-Nadam without compro-mising performance indicates that our combination strikes a good balance between accuracy and computational efficiency, making it more practical for large-scale vis-ual relationship detection tasks.

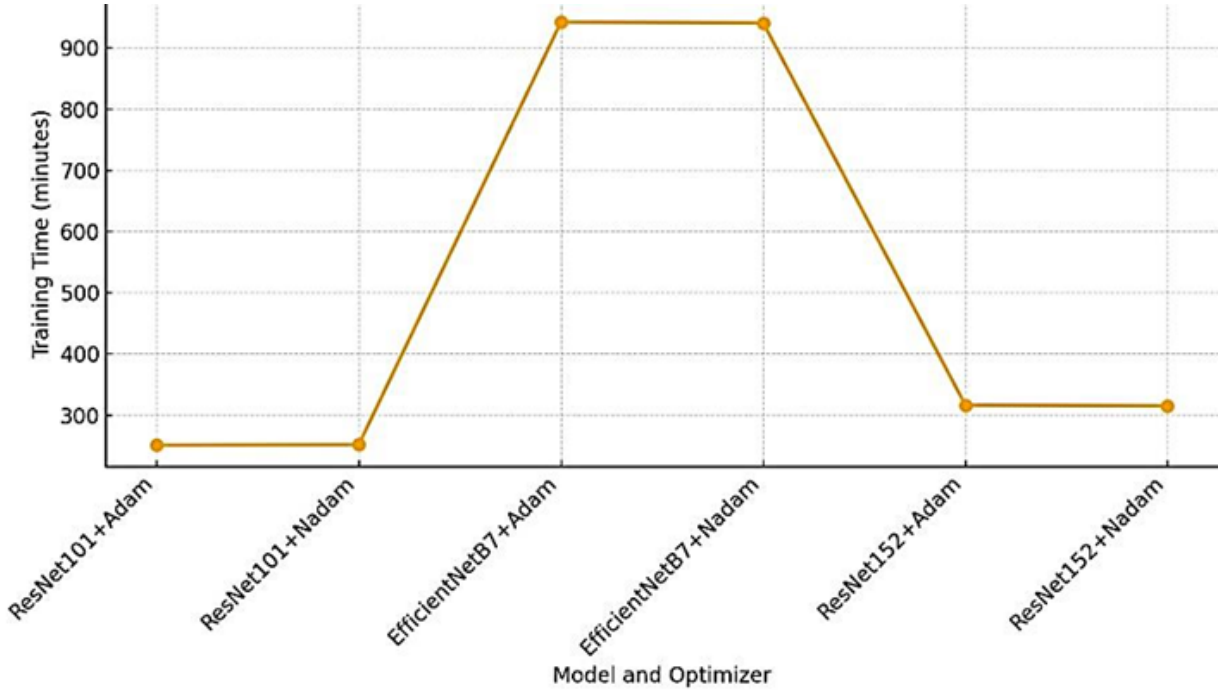

Figure 10: Training Time Comparison for Different Models and Optimizers

Model Comparison: ResNet152 vs. ResNet101 vs. EfficientNetB7

- Our approach (ResNet152-Nadam) outperforms ResNet101 in all key metrics (R@50, R@100, R-ZR@50, R-ZR@100) (Fig. 8 and Fig. 9), demonstrating the advantages of deeper models for visual relationship detection. The increased depth of ResNet152 contributes to better feature extraction and representation, which positively impacts predicate detection performance.

- EfficientNetB7 offers slightly better recall at lower ranks (R@50, R@100)(Fig. 8) when paired with Adam or Nadam. However, its performance comes at a much higher computational cost. The significantly longer training time required for EfficientNetB7 (15 hours 42 minutes) (Fig. 10) suggests that while it may offer marginal performance benefits, it is not as efficient in practice as our method (ResNet152-Nadam).

- The ResNet101 model, though widely used, shows lower performance across all metrics (Fig. 8, Fig. 9, and Fig. 11) compared to our model (ResNet152-Nadam) and EfficientNet-B7-Nadam, indicating that the deeper architecture of ResNet152 is more beneficial for detecting complex visual relationships.

Optimizer comparison: Across all models, Nadam (Fig. 11 and Fig. 13) consistently outperformed Adam (Fig. 11 and Fig. 12) in both general recall (R@50, R@100) and zero-shot recall (R-ZR@50, R-ZR@100). This outcome highlights Nadam's efficacy in fine-tuning model parameters for better stability and convergence, which is especially valuable in handling multi-label relationship classifications in VRD tasks.

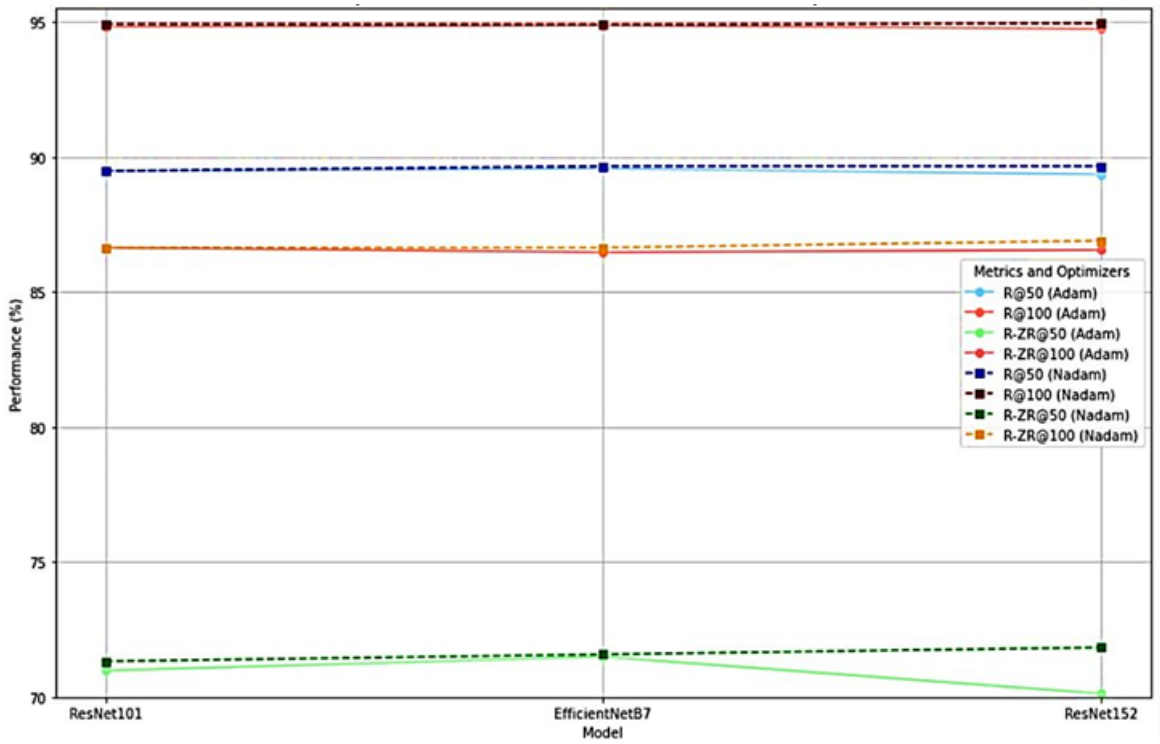

Figure 11: Comparative Performance of Models with Adam vs. Nadam Optimizers

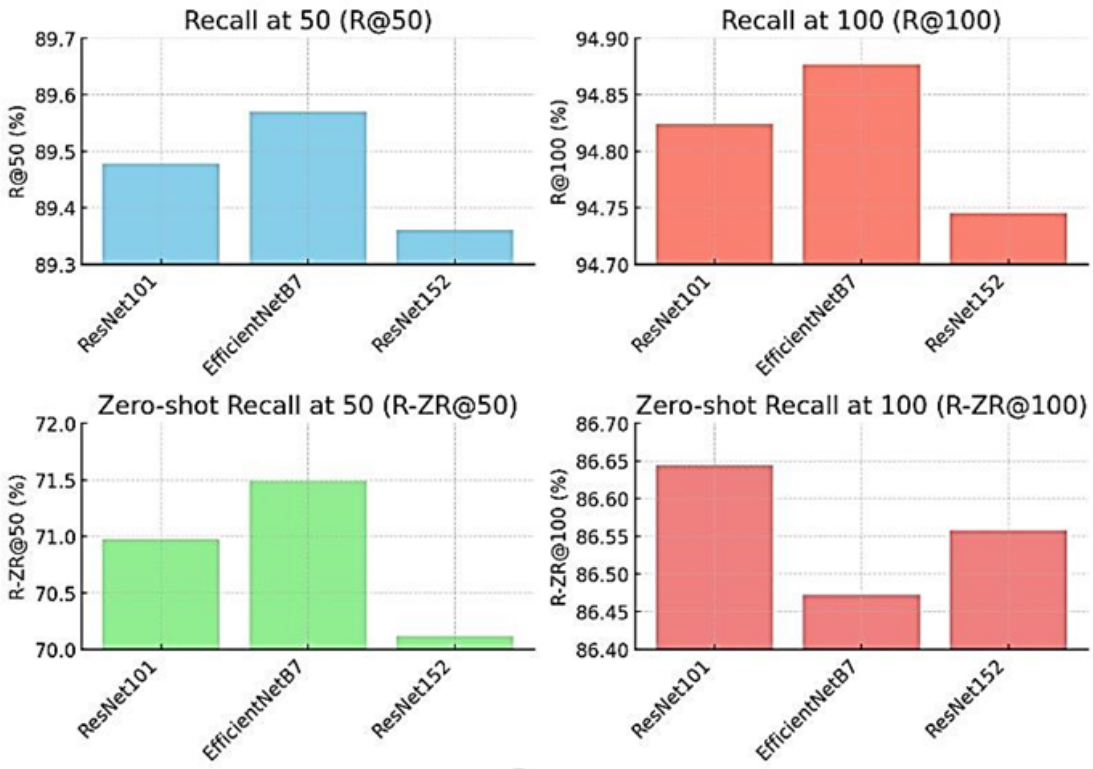

Figure 12: Performance Comparison of Models using Adam Optimizer

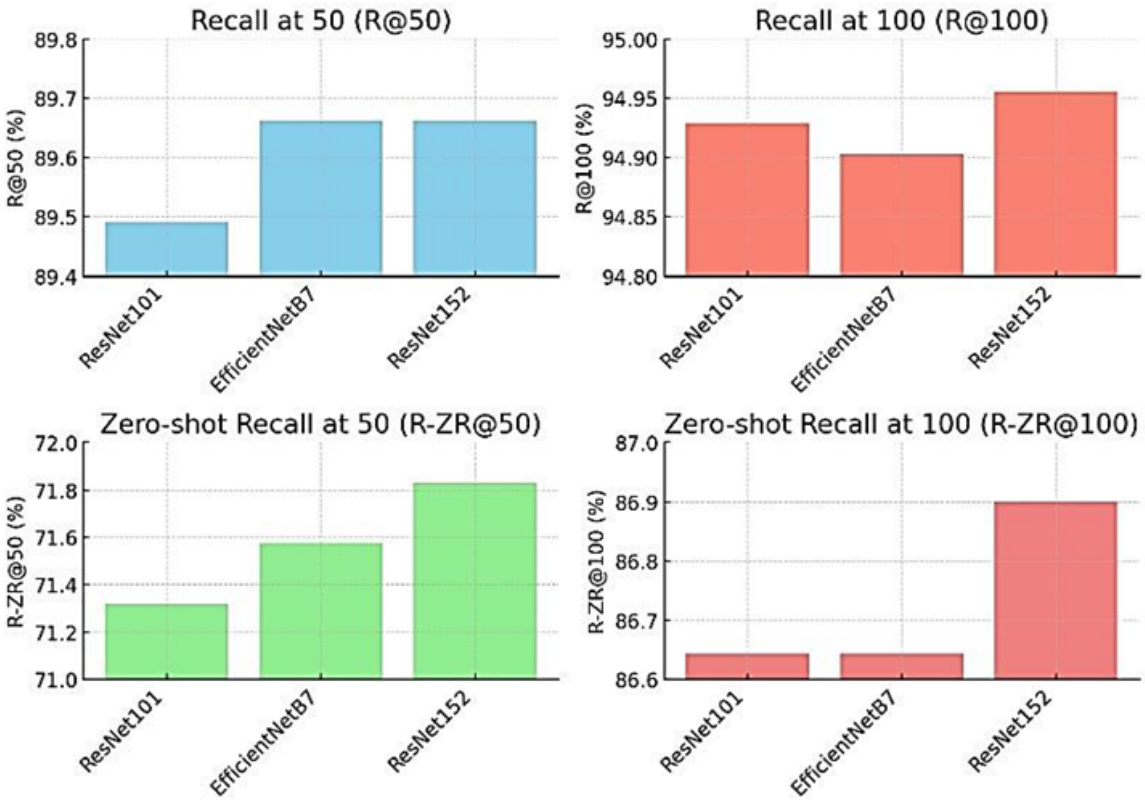

Figure 13: Performance Comparison of Models with Nadam Optimizer

This analytical study demonstrates that our approach (ResNet152-Nadam) excels in the predicate detection task, achieving the highest recall across all metrics and superior zero-shot generalization. The depth of ResNet152 enables it to capture complex patterns more effectively than shallower architectures, whereas EfficientNet-B7, although competitive, requires longer training times. This depth advantage, along with the efficiency of Nadam, helps ResNet152 manage "object-pair combination explosion" by accurately predicting novel relationships, making it suitable for complex, real-world VRD applications. Although EfficientNet-B7 shows slightly higher recall in certain cases, our approach (ResNet152-Nadam) offers a balanced solution with robust performance and practical efficiency.

4.5 State-of-the-Art Comparison

Here's (Table 2): a comparative analysis of our results alongside state-of-the-art ap-proaches on the VRD dataset. This table provides a structured comparison based on the backbone model, recall metrics (R@50, R@100, R-ZR@50, and R-ZR@100), and optimizer strategies used in each approach, which are relevant to our own setup.

Table 2: Predicate Detection Comparison on VRD Dataset

Model Name	Backbone	Optimizer	R@50	R@100	R-ZR@50	R-ZR@100
VRD-CI (2020)	VGG16	SGD	76.66	82.43	N/A	N/A
DR-Net (2017)	VGG16 / ResNet101	Adam	80.78	81.90	N/A	N/A
VRD-RANS (2021)	VGG16 / ResNet18	Adam	85.10	92.59	58.68	78.70
BRNN (2019)	Modified VGG-net	Adam	85.21	90.33	N/A	N/A
VRD-DSR (2018)	VGG16	Adam	86.01	93.18	19.03	23.29
VRTS (2021)	VGG16 / ResNet101	Adam	89.23	94.46	N/A	N/A
ADVRD-SA-4 (2022)	VGG16 / ResNet18	Adam	89.61	94.31	20.23	24.56
Our model	**ResNet152**	**Nadam**	**89.66**	**94.96**	**71.83**	**86.90**

Top performers for R@50 and R@100 (Fig. 14): Our model (ResNet152-Nadam) achieved the highest recall scores (89.66% for R@50 and 94.96% for R@100), out-performing all listed models, including VRTS (ResNet101-Adam) and ADVRD-SA-4 (VGG16/ResNet18-Adam). This performance underscores the advantages of deeper architectures and optimizers for capturing complex relationships in VRD tasks.

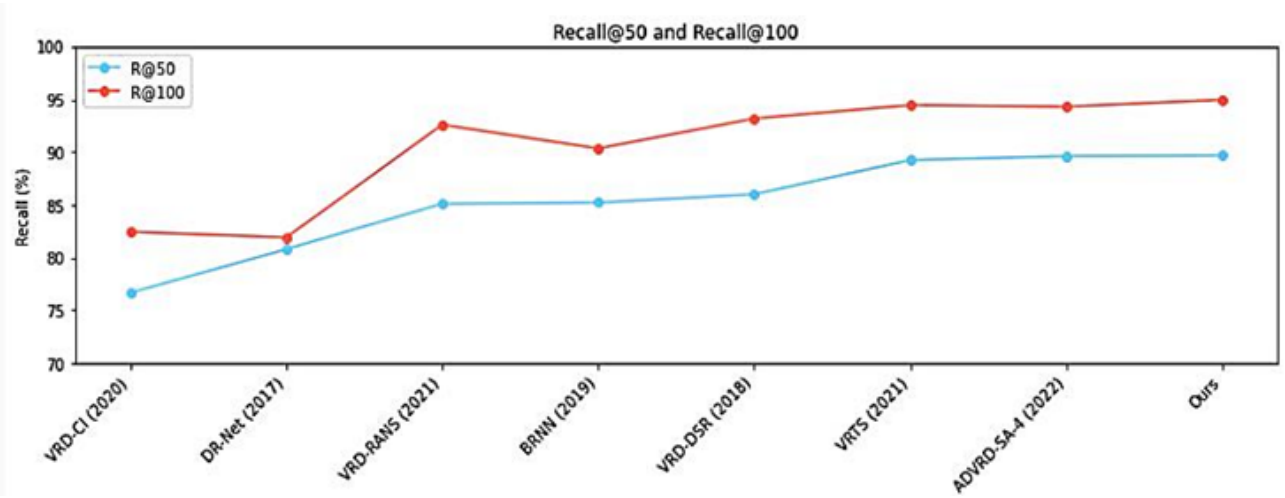

Figure 14: Comparative Analysis of State-of-the-Art Model Performance on R@50 and R@100 for the VRD Dataset

Zero-shot prediction performance (Fig. 15): Among models that report zero-shot performance, ours shows a significant improvement, achieving 71.83% for R-ZR@50 and 86.90% for R-ZR@100 (Fig. 14). These metrics surpass all other models in (Table 2), indicating our model's superior ability to detect relationships not seen during training, which is critical for applications in dynamic, real-world environments, potentially due to the Nadam optimizer, which combines RMSProp with Nesterov momentum for adaptive learning.

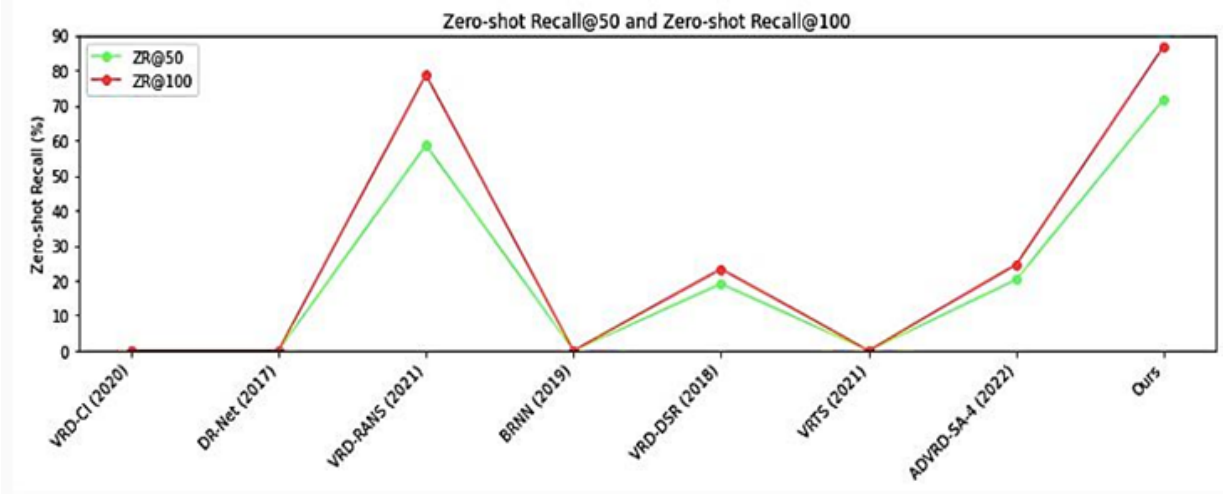

Figure 15: Comparative Analysis of State-of-the-Art Model Zero-Shot Performance on the VRD Dataset

Optimizer strategies: Several state-of-the-art approaches have used the Adam optimizer, which has shown strong effectiveness in visual relationship detection, particularly in handling complex relationships. However, the use of the Nadam optimizer, which combines Adam and Nesterov, in our approach, has demonstrated very high efficiency across all three backbone models, especially when paired with ResNet152. It outperforms the results obtained by Adam in most recall metrics (Fig. 11) and surpasses the state-of-the-art methods (Fig. 15). The advantage is likely due to Nesterov's accelerated Gradient, which enhances convergence. Notably, while SGD performed reasonably well in VRD-CI, its efficiency has consistently lagged behind that of Adam and Nadam.

Backbone models (Fig. 16): ResNet-based backbones, such as ResNet101, ResNet18, and ResNet152, generally achieve higher performance recall metrics than VGG-based models. This is likely due to the greater depth of the residual models, which are able to extract unseen and complex features more efficiently. Although models combining multiple backbones, such as integrating VGG16 with different ResNet configurations, can also be competitive, a single deep ResNet (such as the ResNet152 model used in our method) tends to deliver superior performance.

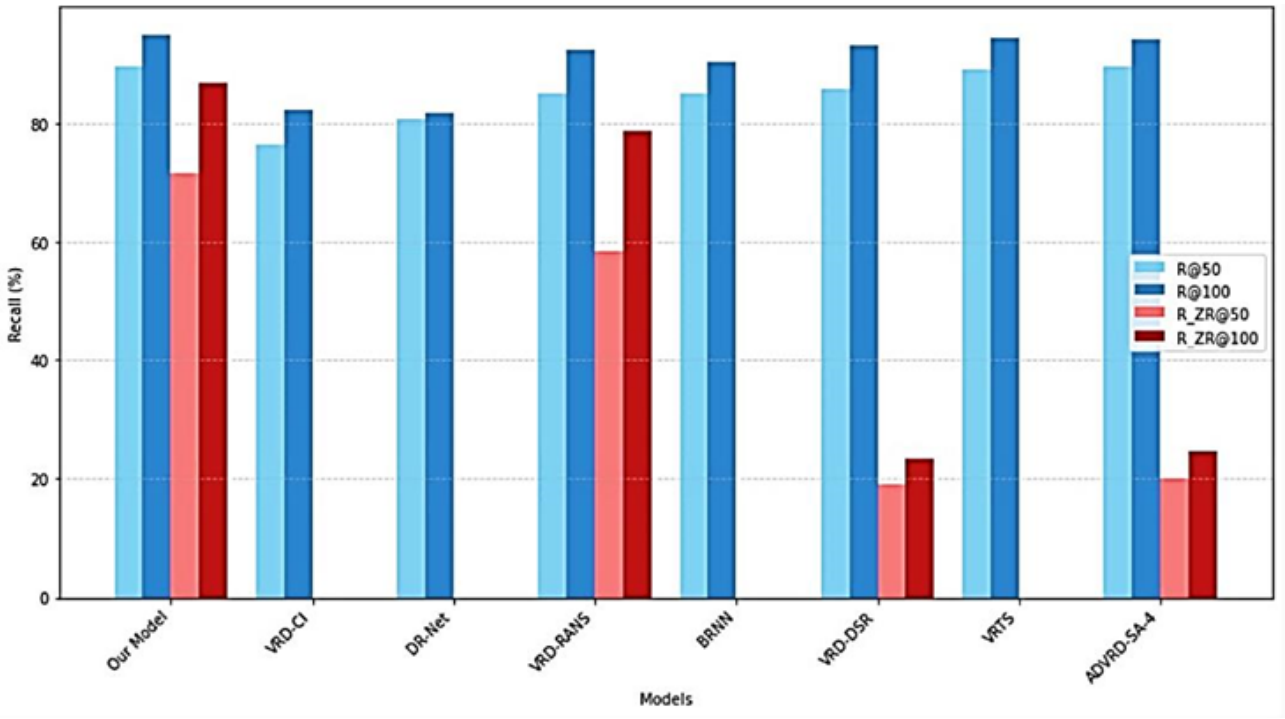

Figure 16: Comparative Analysis of State-of-the-Art Model Performance on the VRD Dataset

Here's the comparative bar chart (Fig. 16) for the models on the VRD dataset. The chart includes:

- R@50 and R@100 in shades of blue to show overall recall performance.
- R-ZR@50 and R-ZR@100 in shades of red for zero-shot recall metrics, where available.

This comparison highlights **the strengths of our model in terms of global recall and zero-shot detection measures, proving the competitive performance of our method in the domain of predicate detection**. The use of **Nadam** and the deeper **ResNet152** backbone were a key factor in achieving these results.

Future research may benefit from this approach by integrating deeper models and effective optimizers with techniques like attention mechanisms or multi-task learning. Such enhancements could further improve accuracy and computational efficiency in VRD tasks, advancing applications that demand precise understanding of unseen relationships in dynamic environments.

5 Conclusion

This article introduces a new method that leverages the ResNet152 model form feature extraction, paired with the Nadam optimizer for model optimization. Additionally, it includes an in-depth comparative analysis featuring EfficientNet-B7 a ResNet101 as alternative backbones for the Faster R-CNN feature extraction framework. We assess the effectiveness of each model, alongside the Adam and Nadam optimizers, in comparison to our approach for predicate detection tasks. Furthermore, we present a comparative study against recent state-of-the-art methods to contextualize our results within the latest advancements in the field. Our results on the VRD dataset show that our combination of ResNet152 with the Nadam optimizer achieved the highest scores across all recall metrics and significantly outperformed current state-of-the-art methods in this task. These results underscore the critical importance of model architecture and optimizer selection in enhancing predicate detection capabilities.

Our approach, along with this comparative analysis, serves as the foundation for ours future research efforts aimed at advancing visual relationship detection applications, particularly in complex real-world scenarios

where relationships are often unseen. The insights gained from this study will guide the development of more robust and efficient models capable of handling the complexities of visual relationships in divers contexts.

Declarations

Funding

This research did not receive any specific grant from agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interests/Competing Interests

The authors declare that they have no conflicts of interest related to this work.

Data Availability Statement

The dataset used in this article is the open-access VRD dataset, which can be accessed via the following link: `https://cs.stanford.edu/people/ranjaykrishna/vrd/`

References

[1] S. Cebollada, L. Payá, M. Flores, A. Peidró, O. Reinoso, "A state-of-the-art review on mobile robotics tasks using artificial intelligence and visual data," *Expert Systems with Applications*, vol. 167, p. 114195, 2021. doi: .

[2] J. Fan, P. Zheng, and S. Li, "Vision-based holistic scene understanding towards proactive human–robot collaboration", *Robotics and Computer-Integrated Manufacturing*, vol. 75, p. 102304, 2022. doi:

[3] Y. Li, W. Ouyang, B. Zhou, K. Wang, X. Wang, "Scene Graph Generation from Objects, Phrases and Region Captions," *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, Venice, Italy, pp. 1261–1270, 2017. doi:

[4] J. Wald, H. Dhamo, N. Navab, F. Tombari, "Learning 3D Semantic Scene Graphs From 3D Indoor Reconstructions," *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*,Seattle, WA, USA, pp. 3960–3969, 2020. doi:

[5] D. Xu, Y. Zhu, C. B. Choy, L. Fei-Fei, "Scene Graph Generation by Iterative Message Passing," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 5410–5419, 2017. doi:

[6] B. Dai, Y. Zhang, D. Lin, "Detecting Visual Relationships with Deep Relational Networks," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 5715–5723, 2017. doi:

[7] Z.S. Hung, A. Mallya, S. Lazebnik, "Contextual Translation Embedding for Visual Relationship Detection and Scene Graph Generation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 3820–3832, 2021. doi:

[8] S. Xiao, W. Fu, "Visual Relationship Detection with Multimodal Fusion and Reasoning," *Sensors*, vol. 22, no. 20, p. 7918, 2022. doi:

[9] L. Zhang, Y. Wang, H. Chen, J. Li, Z. Zhang, "Visual relationship detection with region topology structure," *Information Sciences*, vol. 564, pp. 384–395, 2021. doi:

[10] A. Tato, R. Nkambou, "Improving Adam Optimiser", *Proc. ICLR Workshop Track*, Vancouver, BC, Canada, 2018. [Online]. Available: `https://openreview.net/forum?id=HJfpZq1DM`

[11] S. Xiao, W. Fu, "Optimizing Continuous Prompts for Visual Relationship Detection by Affix-Tuning," *IEEE Access*, vol. 10, pp. 70104-70112, 2022. doi:

[12] M.-G. Gan, Y. He, "Adaptive depth-aware visual relationship detection," *Knowledge-Based Systems*, vol. 247, p. 108786, 2022. doi:

[13] Z. Liu, W.-S. Zheng, "Learning multimodal relationship interaction for visual relationship detection," *Pattern Recognition*, vol. 132, p. 108848, 2022. doi:

[14] Y. Zhu, X. Liang, B. Lin, Q. Ye, J. Jiao, L. Lin, X. Liang, "Configurable Graph Reasoning for Visual Relationship Detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 1, pp. 117–129, 2022. doi:

[15] H. Zhou, C. Zhang, M. Zhao, Y. Luo, C. Hu, "Improving Visual Relationship Detection With Two-Stage Correlation Exploitation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 7, pp. 2751–2763, 2021. doi:

[16] M.-J. Chiou, R. Zimmermann, J. Feng, "Visual Relationship Detection With Visual-Linguistic Knowledge From Multimodal Representations," *IEEE Access*, vol. 9, pp. 50441–50451, 2021. doi:

[17] L. Wang, P. Lin, J. Cheng, F. Liu, X. Ma, J. Yin, "Visual relationship detection with recurrent attention and negative sampling," *Neurocomputing*, vol. 434, pp. 55-66, 2021. doi:

[18] Y. He, G. Duan, G. Luo, X. Liu, "Robust Visual Relationship Detection towards Sparse Images in Internet-of-Things," *Wireless Communications and Mobile Computing*, vol. 2021, pp. 1–10, 2021. doi:

[19] H. Zhang, Z. Kyaw, S.-F. Chang, T.-S. Chua, "Visual Translation Embedding Network for Visual Relation Detection," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 3107-3115, 2017. doi:

[20] Y. Xu, H. Yang, S. Li, X. Wang, M. Cheng, "Contextual Coefficients Excitation Feature: Focal Visual Representation for Relationship Detection," *Applied Sciences*, vol. 10, no. 3, p. 1191, 2020. doi:

[21] Y. Li, Y. Wang, Y. Zhu, "Visual Relationship Detection with Contextual Information," *Computers, Materials & Continua*, vol. 63, no. 3, pp. 1575-1589, 2020. doi:

[22] X. Liu, M.-G. Gan, "RDBN: Visual relationship detection with inaccurate RGB-D images," *Knowledge-Based Systems*, vol. 204, p. 106142, 2020. doi:

[23] Y. Zhan, J. Yu, T. Yu, D. Tao, "Multi-task Compositional Network for Visual Relationship Detection," *International Journal of Computer Vision*, vol. 128, no. 8, pp. 2146–2165, 2020. doi:

[24] Y. Dai, C. Wang, J. Dong, C. Sun, "Visual relationship detection based on bidirectional recurrent neural network," *Multimedia Tools and Applications*, vol. 79, no. 47, pp. 35297–35313, 2020. doi:

[25] K. Liang, Y. Guo, H. Chang, X. Chen, "Visual Relationship Detection With Deep Structural Ranking," *Proc. AAAI Conf. on Artificial Intelligence (AAAI)*, New Orleans, LA, USA, 2018. doi:

[26] F. Plesse, A. Ginsca, B. Delezoide, F. Prêteux, "Modelling relations with prototypes for visual relation detection," *Multimedia Tools and Applications*, vol. 80, no. 15, pp. 22465-22486, 2021. doi:

[27] Y. Shi, P. Long, K. Xu, H. Huang, Y. Xiong, "Data-driven contextual modeling for 3D scene understanding," *Computers & Graphics*, vol. 55, pp. 55-67, 2016. doi:

[28] Y. Xia, W. Xu, L. Zhang, X. Shi, K. Mao, "Integrating 3D structure into traffic scene understanding with RGB-D data," *Neurocomputing*, vol. 151, pp. 700-709, 2015. doi:

[29] B. Satouri, K. Satori, A. El Abderrahmani, "Genetic algorithms and bundle adjustment for the enhancement of 3D reconstruction," *Multimedia Tools and Applications*, vol. 79, no. 39, pp. 29265–29288, 2020. doi:

[30] A. Shaabana, Z. Gharaee, P. Fieguth, "Video Relationship Detection Using Mixture of Experts," *IEEE Access*, vol. 11, pp. 26048–26058, 2023. doi:

[31] M. Qu, G. Deng, D. Di, J. Cui, T. Su, "Dual attentional transformer for video visual relation prediction," *Neurocomputing*, vol. 550, p. 126372, 2023. doi:

[32] S. A. M. Mohamud, A. Jalali, M. Lee, "Hierarchical reasoning based on perception action cycle for visual question answering," *Expert Systems with Applications*, vol. 241, p. 122698, 2024. doi:

[33] L. He, X. Hu, "The application of GCN algorithm in building construction knowledge graph updating under the combination of artificial intelligence and knowledge management," *International Journal of Cognitive Computing in Engineering*, vol. 6, pp. 65–73, 2025. doi:

[34] Y. Xue, S. Ren, W. Cai, Y. Ma, L. Senhadji, H. Shu, J. Wu, "Make your choice for multimodal knowledge graph completion," *Knowledge-Based Systems*, vol. 323, p. 113750, 2025. doi:

[35] C. Lu, R. Krishna, M. Bernstein, L. Fei-Fei, "Visual Relationship Detection with Language Priors," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 3266–3274, 2016. doi: