Relational Prior Knowledge Graphs for Detection and Instance Segmentation

Osman Ülger[1] Yu Wang[2] Ysbrand Galama[2] Sezer Karaoglu[1] Theo Gevers[1] Martin R. Oswald[1]
[1]University of Amsterdam [2] TomTom
{o.ulger,s.karaoglu,th.gevers,m.r.oswald}@uva.nl {yu.wang,ysbrand.galama}@tomtom.com

Abstract

Humans have a remarkable ability to perceive and reason about the world around them by understanding the relationships between objects. In this paper, we investigate the effectiveness of using such relationships for object detection and instance segmentation. To this end, we propose a Relational Prior-based Feature Enhancement Model (RP-FEM), a graph transformer that enhances object proposal features using relational priors. The proposed architecture operates on top of scene graphs obtained from initial proposals and aims to concurrently learn relational context modeling for object detection and instance segmentation.

Experimental evaluations on COCO show that the utilization of scene graphs, augmented with relational priors, offer benefits for object detection and instance segmentation. RP-FEM demonstrates its capacity to suppress improbable class predictions within the image while also preventing the model from generating duplicate predictions, leading to improvements over the baseline model on which it is built.

1. Introduction

In cognitive psychology, it is well established that humans have a remarkable ability to perceive and reason about the world around them by understanding the relationships between objects [2, 10, 17]. By recognizing how objects relate to each other, humans can build a mental representation of their environment, reason about possible actions and make predictions about outcomes of such actions. This ability is essential for a wide range of tasks, from simple everyday activities like crossing the road to more complex tasks like understanding natural language, planning and decision-making. Similarly, in the field of computer vision, relationships between objects have become an increasingly important topic of research [30, 16, 26]. By leveraging relationships between objects, computer vision systems can enhance their ability to detect and segment objects in images, as well as to reason about their relationships, making it possible to build more sophisticated applications that require a deeper understanding of the visual world.

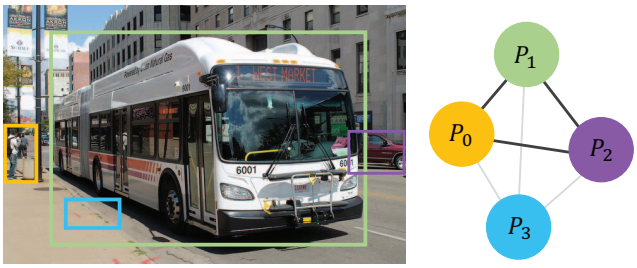

Figure 1. **Problem Setting.** From proposals, a graph is constructed and its edge weights are determined and assigned using prior knowledge. Redundant proposals (*e.g.* in blue) receive low edge weights due to their limited contextual relevance to other objects in the scene. As a result, less contextual feature information from neighboring nodes is aggregated to and from these proposal nodes, which eventually removes them from the final prediction.

In this paper, we explore the role of different types of object relationships in instance segmentation, and propose a method for enhancing object proposals by modeling relationships between them. To this end, we introduce a Relational Prior-based Feature Enhancement Model (RP-FEM), a novel framework which combines a multi-headed attention mechanism to select relevant priors and a graph transformer model to aggregate them. Images are represented as scene graphs, where visual feature representations of proposals are modeled as graph nodes and multi-dimensional edges are obtained from prior knowledge about object-object relations. We propose to represent such relations in a Relational Prior Knowledge Graph (RPKG), which is sourced from the scene graph dataset Visual Genome (VG) [19]. Different from previous works, which enhance proposal features with relational priors for scene graph classification [26] or object detection [16, 30], RP-FEM does not rely on ground truth object regions or an initial classification of proposals. Instead, our method is able to compute relevant relational prior values by attending object neighborhoods in the scene graph with object neighborhoods in the RPKG.

Experimental evaluations on COCO show that using

scene graphs, enhanced with relational priors, is beneficial for object detection and instance segmentation. Our method demonstrates its capacity to filter out predicted areas which are improbable when considered in context to other candidate objects in the scene. Furthermore, RP-FEM shows a remarkable capability of reducing the amount of duplicate predictions of the same object instances. All code is publicly available to facilitate additional research on this topic[1].

In summary, our contributions are as follows:

1. We propose RP-FEM, a novel graph transformer-based model to enhance object proposal features for object detection and instance segmentation using relational priors;

2. We propose and assess multiple types of relational prior knowledge graphs as relational priors to our model. We demonstrate through qualitative and quantitative evaluations that our model is able to improve in cases where object context is relevant when making predictions.

2. Related Work

Instance Segmentation. Instance segmentation is a challenging task in computer vision that involves identifying and segmenting objects within an image. The problem is approached by using convolutional neural networks (CNNs) [12, 24], graph-based methods [4, 31] and transformer-based techniques [9, 6]. Notable works in this field include Mask R-CNN [12], GCNet [4] or Mask2Former [6]. Building on top of such architectures, various works are proposed to utilize prior knowledge in order to enable targeted object search [5], weak supervision [13, 11] or mask refinement [3, 29] with priors originating from bounding box tightness [13, 11], object shapes [5, 3], object contours [11] or object connectivity [29]. Despite these efforts to provide instance segmentation algorithms with more object-specific information with priors, leveraging prior knowledge about object-to-object information still remains an open problem. Although the majority of instance segmentation models benefit from spatial context information in feature space, explicit relationships between objects are ignored.

In contrast to existing methods, we propose to utilize such contextual information by modeling common knowledge about relations between objects in an image, which can constitute to multiple relation types. We extend the Mask R-CNN architecture and showcase its advantages through the integration of relational priors, which can be readily derived from existing dataset statistics.

[1] https://github.com/ozzyou/RP-FEM

Feature Enhancement with Relational Priors. Relational reasoning for feature enhancement is studied in different areas. Kang et al. [18] propose a graph relation network which embeds more discriminative metric spaces for image classification. Relation distillation networks [8] improve video object detection by modelling appearance and geometric relations via multi-stage reasoning. While these works fall under a collection of works which utilize relationships through combination of instance features [14], a large cohort of other works focus on reasoning about scenes by detecting and classifying relationships from visual input [7, 32], driven by scene graph datasets Visual Relation Detection (VRD) [21] and Visual Genome (VG) [19]. The creation of VG enabled the development of models which use object-to-object relations, often co-occurrence statistics, as external high-level knowledge, which can in turn be used when reasoning about scenes. In previous works, relational prior knowledge about object-object relations is exclusively applied to scene graph classification [26], object classification, [26, 16, 30] or object detection [16, 30].

Sharifzadeh *et al.* [26] show that prior knowledge provides significant improvements in the scene graph and object classification tasks. However, their method relies on available ground truth bounding boxes during inference and therefore a detection or segmentation model cannot be trained end-to-end. For object detection, Jiang *et al.* [16] consider edges in the prior knowledge graph as a supervision signal for edges in the scene graph, predicted from the pairwise L1 difference between the features of each region pair. In our work, the relational prior knowledge is obtained via attention over pairs of readily available object proposal features without loss of feature information. Xu *et al.* [30] propose Reasoning-RCNN in which a category-to-category undirected graph is constructed from classifier weights. Region proposals are then mapped to each class node, essentially forming an initial classification. This graph is then evolved to obtain contextualized class embeddings, which are then concatenated to region proposals.

Our work does not rely on initial classifications and operates fully in the region proposal space, thereby limiting the chance of incorporating wrong prior knowledge as a consequence of misclassification. We furthermore apply our method to instance segmentation, while existing related works exclusively focus on scene graph classification, object classification and object detection. Obtained relational priors represent the importance of each object-object relation in a scene before proposal features are propagated across a fully connected and directed graph.

3. Method

Our aim is to enhance proposal features of an underlying base detection and instance segmentation model - the Mask R-CNN framework in this paper - with relational

prior knowledge using a Relational Prior Knowledge Graph (RPKG) and *context updates*. In this section, we first provide details on how a RPKG is constructed and which relation types are considered in our proposed architecture. Then, we introduce our RP-FEM model to predict edge weights, representing object-object relations, in a fully-connected scene graph using the RPKG. Next, we detail how enhanced proposal features are obtained from the scene graph through context updates. Lastly, we provide details on how the enhanced proposal features are adapted within the Mask R-CNN framework.

3.1. Relational Prior Knowledge Graph

Before training, we build three different Relational Prior Knowledge Graphs (RPKG) from the Visual Genome (VG) [19] dataset once. As the nodes of the RPKG, we use the feature representation $d \in \mathbf{D}^{C \times F}$ of each class $c \in \mathbf{C}$ in the penultimate layer of a Faster R-CNN model [25] pre-trained on C classes with F feature dimensions. We use Faster R-CNN due to its architectural similarity with Mask R-CNN regarding object detection. Using the scene graph annotations in VG, we collect object-object relationships which represent the edges of the RPKG. The different relationship types consist of:

(1) Co-occurrence. Measures how often on average two object classes appear together across the dataset. For each class, the amount of co-occurence with other classes is collected and divided by the individual appearances in a scene.

(2) Relative Orientation. When two objects $\{A, B\}$ appear in the same scene, the relative orientation measures how often object A is *at the center of, left of, right of, above or below* object B. Multiple options, such as *left of* and *above*, can occur at the same time, *i.e.* "A is above and left of B". For each object pair, the 5-dimensional outcome is averaged over all samples in which the objects co-occur.

(3) Relative Distance. Measures the mean distance and mean standard deviation between the locations of two objects which co-occur in an image, relative to the size of its ground truth bounding box and size of the image.

This results in the relational prior knowledge graph $\mathbf{R} = \langle \mathbf{D}, \mathbf{K} \rangle$, with $\mathbf{K} \in \mathbb{R}^{C \times C \times R}$ where R depends on which relations are used. We consider relationships between object categories in VG which overlap with the COCO classes. However, some object category names in VG do not have a one-to-one correspondence with those in COCO, *e.g.* "hair blower" in VG versus "hair dryer" in COCO. In order to mitigate this, we manually link semantically identical object categories. The RPKGs, code to construct them and the dataset-to-dataset mappings of object categories will be made publicly available upon publication.

3.2. From Prior Knowledge to Useful Knowledge

Our method is designed to enhance proposal features with prior knowledge information. Central to our proposed architecture is the process of retrieving relevant relational information from $\mathbf{R}$ based on the appearance of potential objects in the proposals and to represent the relational information in the proposal feature space. Naturally, proposal features do not consider the relations among them. In order to do so, we first construct a scene graph $\mathbf{S} = \langle \mathbf{P}, \mathbf{E} \rangle$ given a set of proposal features $\{p_i, ..., p_\mathcal{N}\} = \mathbf{P} \in \mathbb{R}^{\mathcal{N} \times \mathcal{F}_\text{p}}$ representing the nodes and a set of edges $\{e_{ii}, e_{ij}, ..., e_{\mathcal{N}\mathcal{N}}\} = \mathbf{E} \in \mathbb{R}^{\mathcal{N} \times \mathcal{N} \times \mathcal{F}_\text{e}}$. In the next step, we predict $\mathbf{E}$ using $\mathbf{P}$ and $\mathbf{R}$, thereby retrieving relevant relational information.

To retrieve this relational information, the aim is to compute the similarity between pairs of node features - or neighborhood features in $\mathbf{S}$ - and pairs of node features in $\mathbf{R}$ using an attention mechanism $\text{att}(\cdot)$. Weighted by the extent of similarity between two node neighborhoods, the edge value of each neighborhood in $\mathbf{R}$ is aggregated. More formally, the attention coefficient $\alpha_{(ij),(uv)}$ is computed between each pair of nodes $[p_i, p_j] \in \mathbf{P}$, representing the queries, and all pairs of nodes $[d_u, d_v] \in \mathbf{R}$, representing the keys. Features of node neighborhoods are stacked and linearly transformed with shared weight matrices to create local, latent neighborhood representations $\hat{p}_{ij} \in \mathbb{R}^{\mathcal{F}_\text{p}+\mathcal{F}_\text{p}}$ and $\hat{d}_{uv} \in \mathbb{R}^{\mathcal{F}_\text{r}+\mathcal{F}_\text{r}}$ for $\mathbf{S}$ and $\mathbf{R}$ respectively. In order to compute the final edge values $\mathbf{E}$ in $\mathbf{S}$, linearly transformed edge values in $\mathbf{R}$, representing the values, are multiplied by the corresponding attention weights:

$$\begin{aligned}\alpha_{(ij),(uv)} &= \frac{\exp\left(\text{att}(\mathbf{W}_q[p_i, p_j], \mathbf{W}_k[d_u, d_v])\right)}{\sum_{u=0}^{C}\sum_{v=0}^{C}\exp\left(\text{att}(\mathbf{W}_q[p_i, p_j], \mathbf{W}_k[d_u, d_v])\right)} \\ &= \frac{\exp\left(\text{att}(\hat{p}_{ij}, \hat{d}_{uv})\right)}{\sum_{u=0}^{C}\sum_{v=0}^{C}\exp\left(\text{att}(\hat{p}_{ij}, \hat{d}_{uv})\right)} \end{aligned} \tag{1}$$

$$e_{(ij),(kl)} = \alpha_{(ij),(kl)})\mathbf{W}_v\mathbf{R}_{kl} \tag{2}$$

$$\mathbf{E}_{ij} = \mathbf{W}_\text{E}\sum_{k=0}^{C}\sum_{l=0}^{C} e_{(ij),(kl)} \tag{3}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_\mathbf{E}$ represent shared weight matrices to obtain latent representations of the queries, keys, values and predicted edge values respectively. The resulting edge matrix $\mathbf{E}$ now weighs the importance of a proposed object to other proposed objects in the scene graph based on relational prior knowledge obtained from occurrences (or lack of occurrences) of such object combinations in the prior knowledge graph. By using proposals to obtain the prior-knowledge based edge matrix $\mathbf{E}$ with attention, our method does not rely on the classification of proposals and

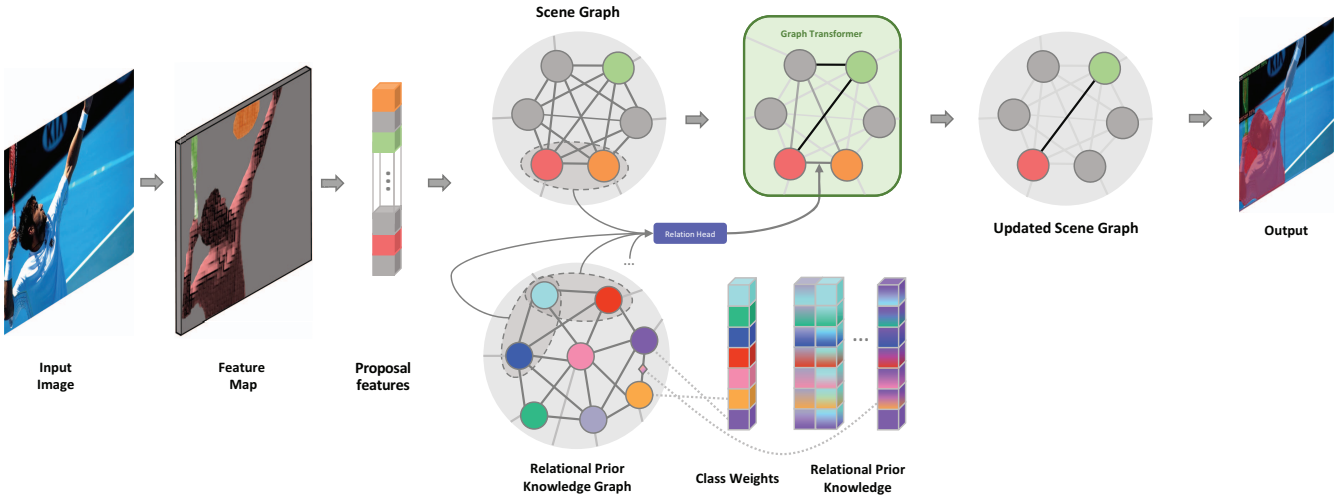

Figure 2. **Method overview.** Our Relation Head updates each edge in the Scene Graph with relational prior knowledge by attending node neighborhoods in the Scene Graph (representing proposals) with node neighborhoods in the Relational Prior Knowledge Graph (representing class embeddings). Original proposal features and predicted edges are fed to a Graph Transformer to obtain an updated Scene Graph. From the updated Scene Graph, bounding boxes and masks are predicted.

therefore avoids possible challenges posed when proposals are incorrectly classified.

3.3. Context Update

After predicting all edge values in $\mathbf{E}$ of the scene graph $\mathbf{S}$, we execute an operation referred to as the *context update*, employing a multi-layered Graph Transformer. The context update ensures that node features in $\mathbf{S}$ are aggregated across the graph to provide each node with more context about the entire scene, as well as the usual, prior-based relations it has with other nodes in such a context. In the process, each node gets informed about its neighboring nodes through messages $\mathbf{m}$, weighted by the edge matrix $\mathbf{E}$ or $\mathbf{A}$, to result in context-aware nodes $\{\mathbf{z}_i, ..., \mathbf{z}_\mathcal{N}\} = \mathbf{Z} \in \mathbb{R}^{\mathcal{N} \times \mathcal{F}_\mathbf{z}}$:

$$\mathbf{f}_{ij}^{(l)} = \mathcal{E}(\delta_{ij})) \qquad \text{with} \quad \delta_{ij} = \begin{cases} \mathbf{E}_{ij} & \text{if } l = 0 \\ \mathbf{A}_{ij}^{(l-1)} & \text{if } l > 0 \end{cases} \tag{4}$$

$$\alpha_{ij}^{(l)} = \sigma\big(\mathrm{LReLU}([\mathbf{f}_{ij}^{(l)} \oplus p_i^{(l)}])\big) \tag{5}$$

$$\mathbf{m}_i^{(l)} = \sum_{j \in \mathcal{I}} \alpha_{ij}^{(l)} \mathbf{f}_{ij}^{(l)} \tag{6}$$

$$\mathbf{z}_i'^{(l)} = \mathbf{LN}\left(\mathbf{z}_i^{(l)} + \mathbf{m}_i^{(l),\text{head}} + \mathbf{m}_i^{(l),\text{tail}}\right) \tag{7}$$

$$\mathbf{z}_i^{(l+1)} = \mathbf{LN}\left(\mathbf{z}_i'^{(l)} + f\big(\mathbf{z}_i'^{(l)}\big)\right) \tag{8}$$

where $\mathcal{F}_\mathrm{z}$ represents the dimension of the output features, l represents the l-th Graph Transformer layer, $\mathcal{E}$ is the transformer function applied to the edge features, $\sigma(\cdot)$ the $\mathrm{Softmax}$ function, LReLU the Leaky ReLU activation function [22], $\oplus$ is concatenation, $\mathbf{A}$ the updated adjacency matrix (explained in Eqs. (9)-(12)), $\mathbf{LN}$ the LayerNorm operation [1] and $f(\cdot)$ is a two layered feed-forward neural network with Leaky ReLU non-linearities between each layer. The first Graph Transformer layer considers the original edge matrix $\mathbf{E}$, whereas in following layers the edge matrix is updated to give $\mathbf{A}$:

$$\mathbf{h}_i^{(l),\text{head}} = \mathcal{H}(p_i^{(l)}) \qquad \mathbf{h}_i^{(l),\text{tail}} = \mathcal{T}(p_i^{(l)}) \tag{9}$$

$$\begin{aligned} \alpha_i^{(l),\text{head}} &= \mathrm{LReLU}\big(\mathcal{A}([\delta_i^{\text{head}} \oplus \mathbf{h}_i^{(l),\text{head}}])\big) \\ \alpha_i^{(l),\text{tail}} &= \mathrm{LReLU}\big(\mathcal{A}([\delta_i^{\text{tail}} \oplus \mathbf{h}_i^{(l),\text{tail}}])\big) \\ \text{with} \quad \delta_i &= \begin{cases} \mathbf{E}_i & \text{if } l = 0 \\ \mathbf{A}_i^{(l-1)} & \text{if } l > 0 \end{cases} \end{aligned} \tag{10}$$

$$\alpha_i^{(l),\text{head+tail}} = \sigma([\alpha_i^{(l),\text{head}} \oplus \alpha_i^{(l),\text{tail}}]) \tag{11}$$

$$\mathbf{A}_i^{(l)} = \alpha_i^{(l),\text{head+tail}} \odot [\mathbf{h}_i^{(l),\text{head}} \oplus \mathbf{h}_i^{(l),\text{tail}}] \tag{12}$$

where additionally $\mathcal{H}$ and $\mathcal{T}$ are the transformer functions applied to proposal features with head or tail indices respectively, $\mathcal{A}$ is the transformer function applied to the concatenated node and edge features and $\odot$ represents element-wise multiplication. A Leaky ReLU activation function is applied to the concatenated features to enable non-linearity. Next, a $\mathrm{Softmax}$ layer is applied on the stacked attention coefficients for the head and tail indices. Finally, the adjacency matrix $\mathbf{A}_i^{(l)}$ is obtained by multiplying the attention coefficients with the transformed proposal features from the head and tail indices, which can then be utilized to properly weigh node features during the feature aggregation step described in Eqs. (4)-(8).

3.4. Mask Prediction

After L iterations of context updates, the edge matrix is discarded and the final, L-th node features $\mathbf{Z}^{(L)}$ are concatenated with the original proposal features $\mathbf{P}^{box}$ to yield output features $\mathbf{O}^{box}$ used for bounding box prediction. The output features $\mathbf{O}^{mask}$ for mask prediction are obtained after concatenation with proposal features $\mathbf{P}^{mask}$ which contain foreground objects found by the box head $\mathcal{B}$. More specifically:

$$\mathbf{O}^{box} = [\mathbf{P}^{box} \oplus \mathbf{Z}^{(L)}] \tag{13}$$

$$\mathbf{O}^{mask} = [\mathbf{P}^{mask}_{\mathcal{B}(\mathbf{O}_{bb})} \oplus \mathbf{Z}^{(L)}] = [\mathbf{P}^{mask}_{fg} \oplus \mathbf{Z}^{(L)}] \tag{14}$$

We ensure that the Mask R-CNN framework is adapted to the shape size increase as a result of the concatenations by adjusting the accepted input size of the box and mask head. Since the prior knowledge graph is fixed, it does not have a gradient and is thus not trained. This allows for training on the image data (COCO) alone. The final output is supervised end-to-end with Mask R-CNN's original loss function, consisting of a term for each of the predictions:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{box} + \mathcal{L}_{mask}. \tag{15}$$

4. Experiments

4.1. Experimental Setup

Datasets and Classes. We obtain the relational priors from ground-truth annotations of the scene graph edges in Visual Genome (VG) [19], which contains over 3000 object categories. However, since VG is not a segmentation dataset, we instead train and evaluate our model on the Common Objects in Context (COCO) [20] dataset which contains 80 classes. Both VG and COCO are licensed under a Creative Commons Attribution 4.0 License. Naturally, VG contains many redundant classes when training and evaluating on COCO, posing the need to select only overlapping classes. Therefore, we automatically assign VG labels to COCO labels which are semantically identical based on WordNet synsets [23] and extract the relational prior knowledge for a class in COCO from all VG classes that have been assigned to it. This results in relational prior knowledge for all possible classes in COCO.

Training. To compare in a fair manner, we set the training parameters similar to [12]. Our models are trained on 8 Tesla V100 GPUs with 32GB memory for 90k iterations with a batch size of 16. The Mask R-CNN model we extend using RP-FEM uses a ResNet-50-FPN backbone and ROIAlign. The feature dimensions $\mathcal{F}_{\text{p}}$, $\mathcal{F}_{\text{r}}$ and $\mathcal{F}_{\text{z}}$ for the node, edge and updated node features are set to 1024 latent dimensions respectively. To sustain a memory-efficient scene graph, we experiment with 128 or 448 proposals originating from the Region Proposal Network (RPN), different from the default 512 in Mask R-CNN. During all experiments, we train the box- and mask head concurrently.

	$AP\uparrow$	$AP_{50}\uparrow$	$AP_{75}\uparrow$	$AP_s\uparrow$	$AP_m\uparrow$	$AP_l\uparrow$
Relation Attention Heads						
1	33.11	52.33	35.65	15.09	34.97	50.10
2	**33.75**	**53.17**	**36.43**	14.85	35.56	**51.52**
4	33.62	53.06	36.14	**15.90**	**35.78**	50.45
Context Updates						
1	**33.75**	**53.17**	**36.43**	14.85	**35.56**	**51.52**
2	33.24	52.87	35.75	**15.70**	35.24	49.37
3	32.73	52.35	35.02	15.09	34.57	49.11
Relationship Types						
Co-occurence	33.88	**53.70**	36.40	15.54	35.75	51.14
Relative Orientation	33.71	53.33	36.04	15.43	35.59	**51.58**
Relative Distance	**33.90**	53.56	**36.65**	**15.56**	**36.16**	51.29
All	33.75	53.17	36.43	14.85	35.56	51.52

Table 1. **Overview of all ablation studies on COCO Instance Segmentation** [20]. Our findings provide the following insights: i) two relation attention heads aid learning best through added complexity; ii) one graph transformer layer is all you need; and iii) overall, Relative Distance proves as the strongest relational prior for instance segmentation.

4.2. Ablation Studies

In this section, we present three ablation studies to investigate the preferred settings for the amount of relation heads, context updates and effect of each relation type. In these initial experiments, 128 proposals are used and results are reported for instance segmentation.

Relation Heads. Having multiple attention heads, *i.e.* attention mechanisms, has been shown to stabilize the learning process in attention-based approaches [27, 28]. Therefore, we first investigate the effect of the number of attention heads on the performance of RP-FEM. In this initial experiment, we set the amount of relation heads to 1, 2 or 4. Results are presented in Table 1 and demonstrate that having multiple relation heads positively impacts the performance when compared to a single head. Given our setting, one potential explanation for the effectiveness of employing multiple heads is that it allows for learning diverse dynamics within a single layer. Interestingly, we observe that 4 relation heads allow the model to more accurately segment instances of small and medium size, while overall, 2 relation heads are preferred. This could indicate that small and medium-sized instances benefit from more dynamics captured by more attention heads. Throughout following experiments, we employ 2 relation heads.

Context Updates. The RP-FEM architecture allows for iterative aggregation of node features, referred to as context updates, using relational priors as edges. This architecture enables the model to enhance scene graph nodes with increased contextual information at each progressive layer. The node and edge features from previous graph

		Object Detection			Instance Segmentation		
	Proposals	$AP \uparrow$	$AP_{50} \uparrow$	$AP_{75} \uparrow$	$AP \uparrow$	$AP_{50} \uparrow$	$AP_{75} \uparrow$
Previous Work							
Mask R-CNN [12]	512	38.5	59.1	42.0	35.0	56.0	37.5
GCNet [4]	2000	37.2	59.0	40.1	33.8	55.4	35.9
MS R-CNN [15]	2000	**38.6**	**59.2**	**42.5**	**36.0**	**55.8**	**38.8**
Baseline Improvement							
Mask R-CNN [12]	128	36.7	56.3	40.3	33.5	53.7	35.7
RP-FEM (Ours)	128	36.2	55.8	39.8	33.9	53.6	36.7
Mask R-CNN [12]	448	38.4	59.2	42.0	35.1	56.1	37.6
RP-FEM (Ours)	448	**38.7**	**59.4**	**42.2**	**35.3**	**56.4**	**37.8**

Table 2. **Quantitative Results of Object Detection and Instance Segmentation on COCO val2017** [20]. RP-FEM succesfully enhances the performance of the Mask R-CNN baseline with relational priors. With considerably less proposals, we also outperform GCNet [4], which similarly tries to model global context for feature enhancement.

transformer layers are propagated to subsequent layers. To determine the optimal number of context updates, this ablation study examines the results for the COCO dataset, as depicted in Table 1. Interestingly, the best performance is achieved with a single context update across most metrics, except for AP_s, where two updates are preferable and result in a significant performance gap. From these findings, we can conclude that it is more important to richly model a single graph transformer layer with multiple relation heads rather than modeling higher-order neighborhood context. For this reason, 1 graph transformer layer is modeled in following experiments.

Relationship Types. In the final ablation study, we conduct an evaluation of the performance of each individual relationship type. The results, presented in Table 1, reveal that specific relationship types offer advantages for different metrics. Notably, the Relative Distance relationship type exhibits superior performance across all metrics, with an average AP of 33.904. This outcome is surprising considering that the literature often uses co-occurrence metrics [16, 30]. Co-occurrence proves more effective when computing AP with an Intersection over Union (IoU) threshold of 50%. This suggests that when considering a larger number of proposal predictions, co-occurrence enhances more proposals successfully compared to other relationship types. Moreover, Relative Orientation appears to play a crucial role in the case of large objects. Interestingly, an ensemble of edge features incorporating all relationship types does not yield the best results for any of the metrics. This observation indicates the potential for exploring additional individual relationship types in future studies.

4.3. Quantitative Analysis

In Table 2, we provide quantitative results of RP-FEM in comparison to previous works which utilize a similar ResNet-50 backbone and two-stage framework for object detection and instance segmentation. We note that previous works such as GCNet [9] and MS R-CNN [15] require a large number of object proposals in order to achieve the reported performance, and comparing against them would require us to reduce the amount by a factor of over 15. Hence, we mainly compare against Mask R-CNN with the number of proposals set to either 128 or 448. When we compare RP-FEM in the object detection task with 128 proposals, we observe that it achieves competitive performance but struggles to outperform Mask R-CNN. In the instance segmentation task, however, RP-FEM performs better overall with an average precision score of 33.9 in comparison to 33.5 achieved by Mask R-CNN. This is likely attributed to lower recall when a small number of proposals is used, consequently allowing less context to be propagated across the scene graph. When the number of proposals is increased to 448, we observe a consistent performance increase over

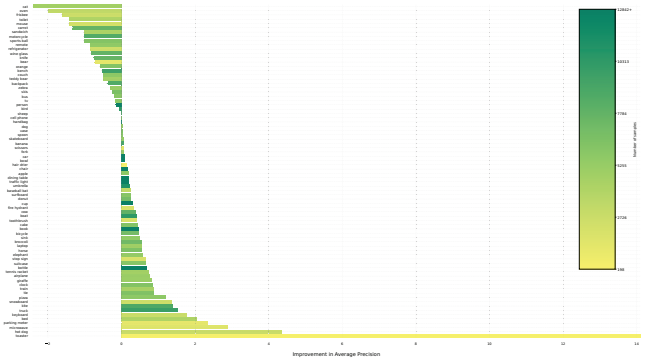

Figure 3. **AP improvement per class.** RP-FEM improves on two thirds of the classes in COCO over Mask R-CNN [12]. Classes with a low number of samples in the dataset particularly benefit from relational prior knowledge.

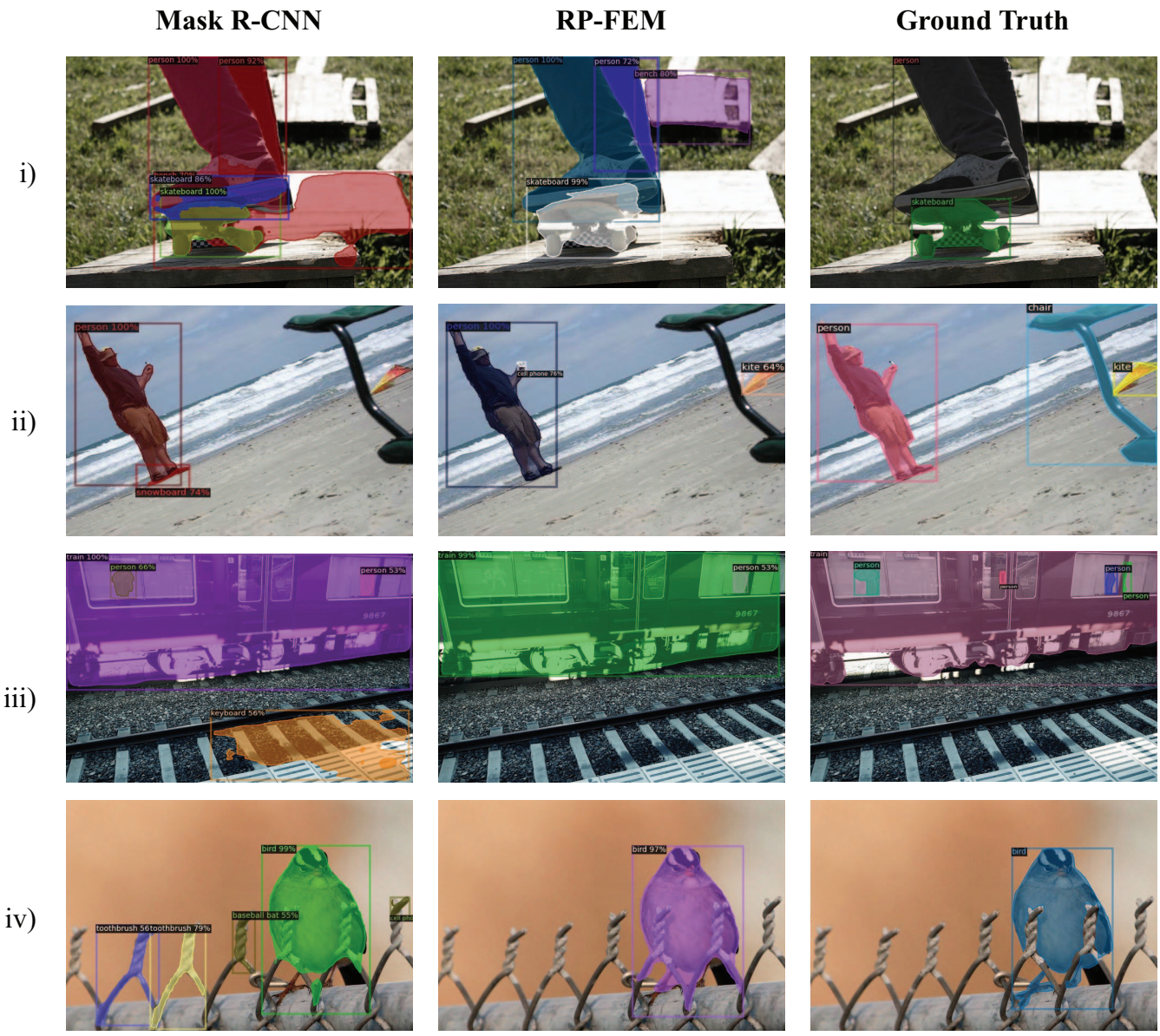

Figure 4. **Qualitative results on COCO [20]** with the following observations: i) our model is able to more accurately predict the correct number of instances, while Mask R-CNN creates duplicate predictions; ii) thanks to contextual information, our model is better able to suppress objects (snowboard) which likely occur with one object (the person), but unlikely in context of multiple objects (the kite). One adverse side effect of relational prior knowledge is the hallucination of objects, such as the cell phone, likely due to many co-occurrences of person and cell phone; iii) RP-FEM filters regions that have visual similarities with other objects (railway as keyboard) if the context does not make sense; iv) multiple instances can be suppressed at once.

Mask R-CNN - even the original version with 512 proposals - across all metrics for both object detection and instance segmentation. GCNet similarly tries to model global context for feature enhancement. This is achieved by applying self-attention on query positions within the image. Our prior-based attention mechanism achieves a superior performance of 35.2 AP over GCNet's 33.8 AP with less than a quarter of the number of proposals (448 versus 2000), while performing on par with MS R-CNN.

In Figure 3, we also report the AP improvement per class of RP-FEM over Mask R-CNN. Our model improves the AP of two thirds of the classes in COCO. Classes which have a low number of samples, such as "toaster", particularly benefit from the incorporation of relational prior knowledge. This indicates that RP-FEM can serve as a promising approach for applications with a long-tailed class distribution. Our analysis indicates the effectiveness of proposal feature enhancement with relational priors and mod-

eling global context.

4.4. Qualitative Analysis

A qualitative analysis of our model's performance on the COCO dataset yields insightful observations, highlighted in Fig. 4, which can be summarized as follows:

Proposal suppression using context. Our model leverages contextual information to effectively suppress objects that are likely to appear in the context of a single object but are improbable when multiple objects are present. This contextual awareness empowers our model to make more informed and contextually consistent predictions. Moreover, our model exhibits a remarkable ability to identify and filter out incorrectly predicted regions that bear visual similarities to certain objects. For example, in Figure 4, the third row depicts a region containing a railway that Mask R-CNN mistakenly identifies as a keyboard due to their visual resemblance. In contrast, RP-FEM successfully discards this region, ensuring that predictions align with the logical context of objects in the scene. Additionally, our model showcases the capability to suppress multiple instances simultaneously, enhancing the instance segmentation process. This efficiency leads to more accurate outputs.

Accurate instance count prediction. Our model demonstrates improved accuracy in predicting the correct number of instances compared to Mask R-CNN, likely guided by the co-occurrence relational prior knowledge. Unlike Mask R-CNN, our model avoids generating duplicate predictions better, resulting in a more precise instance segmentation output. For example, in Figure 4, Mask R-CNN predicts multiple instances of the skateboard, while RP-FEM correctly identifies one skateboard only.

Through this qualitative analysis, we highlight the strengths of our model in accurately predicting instance counts, leveraging contextual information, filtering regions with conflicting visual similarities, and efficiently suppressing multiple instances. These findings showcase the advancements and superior performance achieved by our approach on the COCO dataset.

4.5. Limitations

Our model has the following limitations. When the number of proposals and/or classes in the relational prior knowledge graph grows, predicting each edge in the scene graph becomes costly in terms of memory consumption. This can be addressed by computing the edges sparsely or iteratively, but both workarounds have their price in accuracy or computation time. Furthermore, the incorporation of relational prior knowledge can cause the detection or segmentation model to hallucinate objects. In Figure 4, for example, RP-FEM hallucinates a cell phone in the hand of the man, likely due to many co-occurrences of both objects.

5. Conclusion

The understanding of relationships between objects play an important role in human perception and reasoning. In this work, we explored whether utilizing relationships can play a beneficial role in the tasks of object detection and instance segmentation. To this end, we proposed a Relational Prior-based Feature Enhancement Model which employs the unique capability of suppressing multiple region proposals when they present themselves in a context that is unlikely to be consistent. Furthermore, our model provides a more accurate notion of instance counts, reducing the amount of duplicate object detections and instance segmentations around the same object. Our quantitative results further confirm that our model can outperform its base model, as well as comparative models which model context, all the while using less object proposals. We find that, in particular, classes with a low number of samples benefit strongly from the incorporation of relational prior knowledge. We encourage future work to explore linguistic relationships and to experiment with stronger backbones.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

[2] Michael F. Bonner and Russell A. Epstein. Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature Communications*, 2021.

[3] Caio de Moraes Braz, Luiz Felipe D. Santos, and Paulo A. V. Miranda. Graph-based image segmentation with shape priors and band constraints. In *DGMM*, 2022.

[4] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *ICCV*, 2019.

[5] Long Chen, Weiwen Zhang, Yuli Wu, Martin Strauch, and Dorit Merhof. Semi-supervised instance segmentation with a learned shape prior. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, 2020.

[6] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022.

[7] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. *CoRR*, 2017.

[8] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. Relation distillation networks for video object detection. *CoRR*, 2019.

[9] Bin Dong, Fangao Zeng, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Solq: Segmenting objects by learning queries. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *NeurIPS*, volume 34, pages 21898–21909. Curran Associates, Inc., 2021.

[10] Michelle Greene. Statistics of high-level scene context. *Frontiers in Psychology*, 4, 2013.

[11] Shengyu Hao, Gaoang Wang, and Renshu Gu. Weakly supervised instance segmentation using multi-prior fusion. *CVIU*, 2021.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *CVPR*, 2017.

[13] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. In *NeurIPS*, 2019.

[14] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. *CoRR*, 2017.

[15] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring R-CNN. *CVPR*, 2019.

[16] Chenhan Jiang, Hang Xu, Xiaodan Liang, and Liang Lin. Hybrid knowledge routed modules for large-scale object detection. *CoRR*, 2018.

[17] Daniel Kaiser, Genevieve L. Quek, Radoslaw M. Cichy, and Marius V. Peelen. Object vision in a structured world. *TiCS*, 23(8):672–685, 2019.

[18] J. Kang, R. Fernandez-Beltran, D. Hong, J. Chanussot, and A. Plaza. Graph relation network: Modeling relations between scenes for multilabel remote-sensing image classification and retrieval. *GRSS*, 2020.

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

[20] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, 2014.

[21] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. *CoRR*, 2016.

[22] A.L. Maas, A.Y. Hannun, and A.Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013.

[23] George A. Miller. WordNet: A lexical database for English. In *Human Language Technology*, 1994.

[24] Pedro O. Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *NIPS'15*, 2015.

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[26] Sahand Sharifzadeh, Sina Moayed Baharlou, and Volker Tresp. Classification by attention: Scene graph classification with prior knowledge. *CoRR*, 2020.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[28] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *ICLR*, 2018.

[29] Sara Vicente, Vladimir Kolmogorov, and Carsten Rother. Graph cut based image segmentation with connectivity priors. In *CVPR*, 2008.

[30] Hang Xu, Chenhan Jiang, Xiaodan Liang, Liang Lin, and Zhenguo Li. Reasoning-rcnn: Unifying adaptive global reasoning into large-scale object detection. In *CVPR*, 2019.

[31] Rongtao Xu, Ye Li, Changwei Wang, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. Instance segmentation of biological images using graph convolutional network. *Engineering Applications of Artificial Intelligence*, 110:104739, 2022.

[32] Ji Zhang, Kevin J. Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph generation. *CoRR*, 2019.