Few-Shot Visual Relationship Co-Localization

Vaibhav Mishra*, Mayank Maheshwari*, Revant Teotia*, Anand Mishra
Indian Institute of Technology Jodhpur
{mishra.4,maheshwari.2,trevant,mishra}@iitj.ac.in
$*$: equally contributed

Abstract

In this paper, we study the problem of co-localizing visual subjects and objects connected via a common predicate across a bag of images. For example, given a small collection of images, each containing a common but latent predicate such as "biting", we are interested in localizing who *is biting* what *i.e., visual subject (*who*) and visual object (*what*) pairs in each of the images. Visual relationship co-localization (VRC as an abbreviation) is a challenging task, even more so than the well-studied object co-localization task. This becomes further challenging when the model has to learn to co-localize unseen predicates based on just the similarity between a few images. To solve VRC, we propose an optimization framework to select a common visual relationship in each image of the bag. Obviously, the optimization framework should, (a) learn visual relationship similarity in a few-shot setting, (b) find the optimal solution despite the combinatorial complexity of the problem. To obtain robust visual relationship representation, we utilize a simple yet effective technique that learns relationship embedding as a translation vector from visual subject to visual object in a shared space. Further, to learn visual relationship similarity, we utilize a proven meta-learning technique commonly used for few-shot classification tasks. Finally, to tackle the combinatorial complexity challenge arising from an exponential number of possible solutions, we use a greedy approximation inference algorithm that selects approximately the best solution. We have extensively evaluated our proposed framework on variations of bag sizes obtained from two challenging public datasets, namely VrR-VG and VG-150, and obtain impressive performance gains over baselines.*

1. Introduction

Localizing visual relationship (<subject, predicate, object>) in images is a core task towards holistic scene interpretation [16, 37]. Often the success of such localization tasks heavily relies on the availability of large-scale annotated datasets. *Can we localize visual relationships in images by looking into just a few examples?* In this paper, towards addressing this problem, we introduce an important and unexplored task of **V**isual **R**elationship **C**o-localization (or VRC in short) in few-shot setup. VRC has the following problem setting: given a bag of b images, each containing a common visual predicate P_{so} connecting a visual subject B_s to a visual object B_o, our goal is to automatically lo-

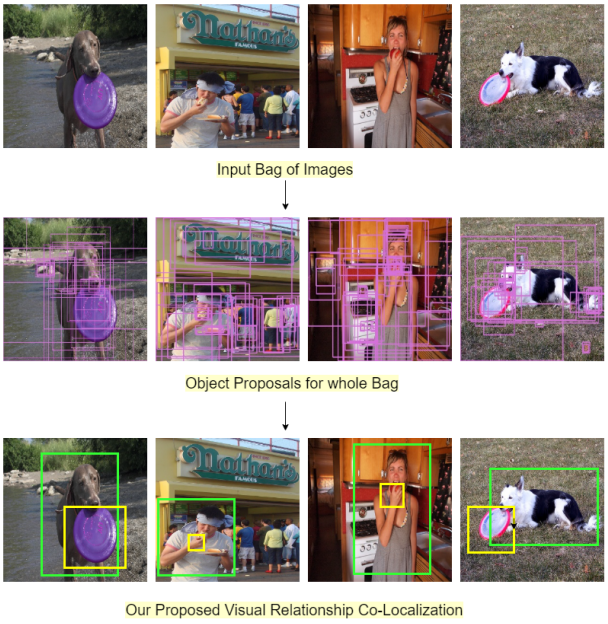

Figure 1. *Given a bag of four images as shown in the first row, can you find the visual subjects and objects that are connected via a common predicate?* Our proposed model in this paper automatically does that. In this illustration, the "biting" predicate is present in all the four images in the first row, and our proposed model localizes those visual subjects and objects in each image that are connected via "biting" as shown in the third row. Note that the category name "biting" is not provided to our approach. Green and yellow bounding boxes indicate the localized visual subject and objects, respectively, using our visual relationship co-localization approach. **[Best viewed in color].**

calize subject and object pairs connected via the common predicate in each of the b images. Note that, during both the training and testing phases, the only assumption is that each image in a bag contains a common predicate P_{so}. However, its category (e.g., eating, riding) is latent.

Consider Figure 1 to understand our goal in a better way. Given a bag of four images, each containing a latent common predicate, e.g. "biting" in this illustration, we aim to localize visual subject and object pairs, such as (dog, frisbee), (man, hot dog), and so on, with respect to the common predicate in each of the images. VRC is significantly more challenging than well-explored object co-localization [30, 26, 12] due to the following: (i) Common objects often share similar visual appearance, but common relationships can visually be very different, for example "dog biting frisbee" and "man biting hot dog" are very different in visual space. (ii) Relationship co-localization requires both visual as well as semantic interpretation of the scene. Further, VRC is also distinctly different from visual relationship detection (VRD) that aims to estimate the maximum likelihood for <subject, predicate, object> tuples from a predefined fixed set. On the other hand, predicates are not provided even during the training phase of VRC, and it has to understand the semantics of visual relationships even for unseen ones.

Visual relationship co-localization (VRC) has many potential applications, examples include automatic image annotation, bringing interpretability in visual search engines, visual relationship discovery. Moreover, co-localizing visual subjects and objects simultaneously on a bag of images also potentially improves the scene's holistic interpretation. To perform visual relationship co-localization, we pose VRC as a labeling problem. In this framework, every possible visual subject and object pair is a potential label for each image. To get the optimal labeling, we define an objective function parametrized by model parameters $\boldsymbol{\theta}$ whose minima corresponds to visual subject-object pairs that are connected via a common latent predicate in all the images. To generalize well on unseen predicates, we follow the meta-learning paradigm to train the model. Just as a good meta-learning model learns on various learning tasks, we train our model on a variety of bags having different common latent predicate in each of them so that the model generalizes to new bags. We use a greedy approximation algorithm during inference, which breaks down the problem into small subparts and combines the subpart solutions greedily.

To evaluate the proposed model's performance, we use two public datasets, namely VrR-VG [20] and VG-150 [33]. The principled formulation of the problem by defining a suitable objective function and our meta-learning-based approach to optimize this objective function helps our method to achieve impressive performance for this challenging task. Further, we present several ablation studies to validate the effectiveness of different components of our proposed framework. We achieve 76.12% co-localization accuracy on unseen predicates of VrR-VG [20] dataset while performing inference on bag size $= 4$.

Contributions of this paper: Our main contributions are two folds:

- We introduce a novel task – VRC (Visual Relationship Co-Localization). VRC has several potential applications and is an important step towards holistic scene interpretation.
- Inspired by the meta-learning paradigm's recent success in solving few-shot learning tasks, we propose a novel framework for performing few-shot Visual Relationship Co-Localization. Our framework learns robust representation for latent visual predicates and is efficacious in performing Visual Relationship Co-Localization with only a few examples.

2. Related Work

2.1. Object Co-localization

Object localization [41, 27, 13, 5] is an important and open problem in computer vision. To address the problem of localizing object overlap between two or more images, object co-localization has been introduced. Object co-localization is related to weakly supervised object localization [40, 18, 6, 34]. To tackle object co-localization, Tang et al. [30] have opted way of combining the box model and image model in such a way that both complements each other as the image model helps to select clean images, and the box model helps to select the boxes that contain the instance of that common object. The latter proposed approaches such as Shaban et al. [26], and Hu et al. [12] opted the lane of Few-Shot learning due to the problem of limited annotated data, the former forms bags of images, and then it finds common object across all the images in a bag while the later one has a support branch and a query branch and there is a common object across both branches, however, the object itself is unknown. While object co-localization is an interesting task, visual relationship co-localization (VRC) requires visual as well as semantic understanding of the scene in the images. To the best of our knowledge, few-shot visual relationship co-localization has not been studied in the literature.

2.2. Visual Relation Detection

Detecting the relationship in an image is an instrumental task in computer vision as it helps in comprehensive scene understanding. Zhang et al. [37] uses the spatial, visual, and semantic features to get the predicted relationship label in the image. The bottleneck to this approach is that it is

limited to those available during training and fails miserably on unseen relationships. Other methods include approaches [38] which takes the object and relations into two different higher dimensional spaces and ensures their semantic similarity and distinctive affinity by using multiple losses. Zhang et al. [39] inculcates new graphical loss to better detect the relationship. Zellers et al. [36] uses a network of stacked Bidirectional LSTMs and Convolutional layers to parse a scene graph and, in between, detect various relationships in the image. With the advancements in graph neural networks, we can see many approaches building upon them. One such approach inspired by GNNs is proposed by Li et al. [19] in which they build upon a sub-graph-based connected graph to detect the visual relationship in a better and efficient way. As compared to visual relation detection, we are distinctively different, as discussed in the introduction of this paper.

2.3. Meta Learning for Few-Shot Learning

Few-shot learning methods [1, 4, 32, 9] are being studied and explored significantly for both computer vision [21, 8] and natural language processing [3, 35, 10, 32, 25, 22]. Few-shot learning is a form of weakly supervised learning approach, and there are two significant methodologies towards solving the few-shot learning problem. Metric-based learning approaches like Siamese Networks [14] which uses a shared CNN architecture for learning the embedding function and uses weighted L1 distance for few-shot image classification, although the problem with this approach is the difference in evaluation metric during training and test due to task shift. Matching Networks [31] uses CNN followed by LSTM architecture for learning the embedding function. Prototypical Networks [28] uses CNN architecture with a squared L2 distance function. Relation Networks [29] proposed to replace the hand-crafted distance metrics with a deep distance metric to compare a small number of images within episodes, each of which is designed to simulate the few-shot setting. Gradient-based meta-learners have two models, namely base-learner and meta-learner, the meta-learner learns across episodes and the base-learner, which is learned inside the episode, the aim is to learn the optimization of the model weights. We adopt a metric-based meta-learning-based approach in this work to perform a few-shot visual relationship co-localization.

3. Approach

Given a bag of b images, $\{I_u\}_{u=1}^{b}$ such that each image of the bag I_u contains a latent common predicate which is present across all the images in the bag. A visual relationship can be represented by identifying the related subject and object from each image which are connected by the same predicate; thus, the goal is to find the ordered set O such that $O = \{(B_i^u, B_j^u)\}_{u=1}^{b}$ where each tuple $< B_i^u, B_j^u >$ corresponds to object proposal pairs in uth image in the bag, and B_i^u and B_j^u are the bounding boxes over visual subject and object respectively. Table 1 represents major notations used in this section.

3.1. VRC as a Labeling Problem

We pose VRC as a labeling problem. To this end, given a bag containing b images, we construct a fully connected graph $G = \{V, E\}$ where $V = \{I_u\}_{u=1}^{b}$ is a set of vertices such that each vertex corresponds to an image. The potential label set for each vertex is a set of all possible pairs of object proposals[1] obtained from the corresponding image. Given this graph and label sets, the goal is to assign one label to each vertex of the graph (or equivalently to each image in the bag) such that visual subject-object pair connected via the latent common predicate P_{so}^{*} is assigned to each image.

The labeling problem formulation for the visual relationship co-localization using an illustrative example is shown in Figure 2. Here, we show four images in a bag, i.e., bag size $b = 4$. Each image is represented as a vertex in a fully-connected graph G. To obtain a label set for each of these vertices (or equivalently each image), we first get object proposals using Faster R-CNN [23]. Let $B = \{B_i^u\}_{i=1}^{p_u}$ be a set of object proposals obtained for Image-u, for example, in Figure 2, we get "woman", "sheep", "hat", "bucket", etc. as object proposals for Image-1. Here p_u is the number of object proposals in Image-u. Given these, the label set of this vertex will contain all possible ordered pairs of object proposals. In other words, the cardinality of this label set equal to $p_u \times (p_u - 1)$.

Further, each ordered pair of the object proposals is connected via a latent predicate. Examples of latent predicate in Image-1 (ref. Figure 2) are petting, wearing, etc. These predicates define visual relationships such as "<woman, petting, sheep>", "<woman, wearing, hat>", etc. Suppose $< B_s^u, P_{so}, B_o^u >$ represents that object proposals (bounding boxes) B_s^u and B_o^u of image-u that are connected via a hidden predicate P_{so}. Then, the label set for Image-u or equivalently corresponding vertex-u is given by:

$$\mathcal{L}_u = \{< B_s^u, P_{so}, B_o^u > \mid s \neq o \text{ and } (B_s, B_o) \text{ is an ordered pair of object proposals in image-}u \text{ and } P_{so} \text{ is a latent predicate.}\} \quad (1)$$

A label $l_{u(s,o)} = < B_s^u, P_{so}, B_o^u > \in \mathcal{L}_u$ is an instance (or member) of label set for vertex-u. For simplifying the notation, we write $l_{u(s,o)}$ as l_{ut} from here onwards where t varies from 1 to $|\mathcal{L}_u|$. Further, the optimal label, i.e., the visual subject-object pair that are connected

[1] Object proposals should not be confused with the object in a visual relationship.

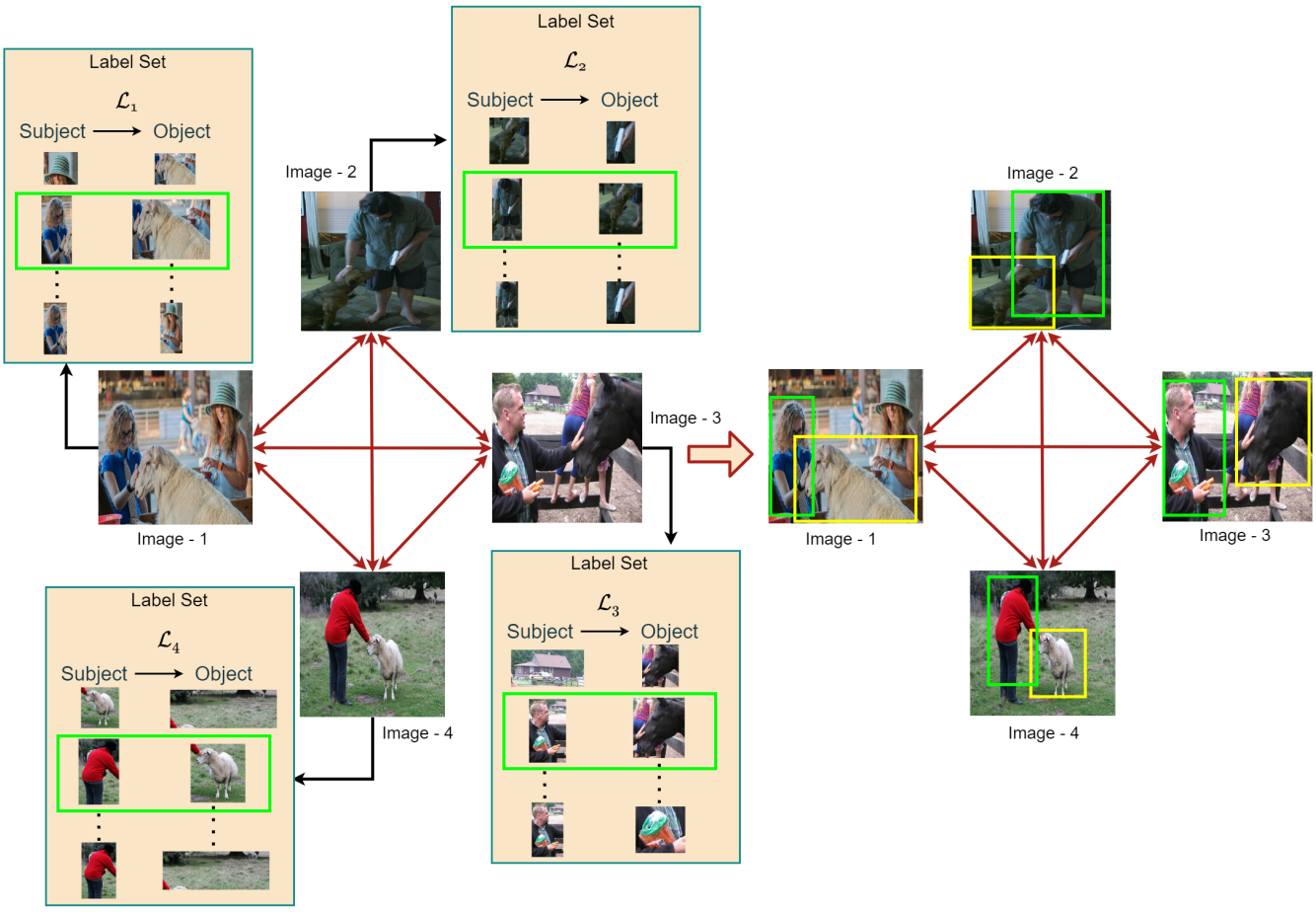

Figure 2. **VRC as a labeling problem.** Given a bag of b images ($b = 4$ in this illustration), we construct a fully connected graph by denoting each image as a node. All the pairs of object proposals in each image constructs the label set for each node. The goal is to find a labeling such that the labels representing common latent predicate are selected for each image, e.g., "petting" in this illustration. We solve this problem by minimizing a corresponding objective function. Refer to Section 3 for more details. **[Best viewed in color]**

via a "common" latent predicate ${P_{so}}^*$ in image-u is represented by: l^*_{ut}. In Figure 2, P^*_{so} = "petting" with visual relationship tuples <woman, petting, sheep>, <man, petting, dog>, <man, petting, horse>, <man, petting, sheep> in Image-1 to 4 respectively. Recall that the goal of the labeling problem is to assign the optimal labels to all of the bag images or, in other words finding an optimal pair of subject and object bounding boxes $< B_s^{u*}, B_o^{u*} >$ for each bag image.

Formulation for the optimal labeling: To solve the labeling problem, we define the following objective function whose minima corresponds to optimal labeling for VRC, i.e., localizing the visual subject-object pairs in each image of a bag that are connected via the common latent predicate:

$$\Psi = \sum_{u=1}^{b} \Big(\min_{t} \Psi_u(l_{ut}) + \sum_{v=1}^{b,u\neq v} \min_{t_1,t_2} \Psi_{uv}(l_{ut_1}, l_{vt_2}, \boldsymbol{\theta}) \Big). \tag{2}$$

In this objective function there are two terms: (i) Unary term $\Psi_u(l_{ut})$ which represents cost of assigning a label $l_{ut} =< B_s^u, P_{so}, B_o^u >$ to image u. Since given an image, any subject-object pair is considered to be equally likely. Therefore, this term of the objective function does not contribute to the optimization. (ii) Pairwise term $\Psi_{uv}(l_{ut_1}, l_{vt_2}, \boldsymbol{\theta})$ represents the cost of image-u taking a label $l_{ut_1} = < B_s^u, P_{so}, B_o^u >$ and image-v taking a label $l_{vt_2} = < B_s^v, P_{so}, B_o^v >$. Here $\boldsymbol{\theta}$ is a learnable model parameter which needs be learnt from few examples. We use a neural model to learn these parameters as described in Section 3.2. Further, the pairwise term of optimization should be defined in such a way that it is lower when hid-

Notation	Meaning
$\mathcal{L}_u$	Label set for Image-u
$l_{so} \in \mathcal{L}_u$	Visual relationship
b	Bag size
p_u	Number of object proposals in Image-u
B_i^u	ith object proposal in Image-u
P_{so}	Latent predicate connecting proposals B_s^u and B_o^u
P_{so}^*	Common latent predicate
$f_\phi(\cdot)$	Relationship embedding network
$R_{\boldsymbol{\theta}}(\cdot,\cdot)$	Visual relationship similarity
$\boldsymbol{\theta}$	Model parameters

Table 1. Notations used in this paper.

den predicates P_{so} of l_{ut_1} and l_{vt_2} are semantically similar, higher otherwise. We compute this pairwise term in Equation 6. Further, to compute this pairwise term, we need to first learn a robust semantic encoding of l_{so} given pair of object proposals $< B_s, B_o >$. In other words, we wish to learn visual relation embedding as follows:

$$f_{l_{so}} = f_\Phi(B_s, B_o), \tag{3}$$

where f_Φ denotes our visual relationship embedding network parameterized by Φ and $f_{l_{so}}$ is the encoding of visual relationship l_{so}. We use a popular relationship encoding network viz. VTransE [37] for computing relationship embedding. However, any other relationship encoder can also be used here.

3.2. Learning to Label with Few Examples

In our problem setting, to be able to generalize well on new bags, the model should be able to learn similarity between visual relationships even when looking into small-size bags at a time. This is usually referred to as a few-shot setting, an active area of research in machine learning. Many learning paradigms exist for addressing the problem in this setting; one of the most successful is Meta-Learning [11, 24]. Therefore, we choose meta-learning to address our optimization problem. To this end, we use one of the metric-based meta-learning approaches [14, 28, 31] i.e., Relation Net [29] to learn the similarity between visual relationships as follows.

Given a pair of visual relationships l_i and l_j, we first obtain their representations f_{l_i} and f_{l_j} respectively using the Equation 3. Then we calculate similarity score between these representations as follows:

$$R_{\boldsymbol{\theta}}(f_{l_i}, f_{l_j}) = w^T K(f_{l_i}, f_{l_j}) + b, \tag{4}$$

where w is a learnable weights matrix and b is the bias vector. Further, K is computed as follows:

$$K(f_{l_i}, f_{l_j}) = \tanh(W_1([f_{l_i}; f_{l_j}]) + b_1) \\ \sigma(W_2[f_{l_i}; f_{l_j}] + b_2) + ((f_{l_i} + f_{l_j})/2), \tag{5}$$

where W_1, W_2 are two learnable weight matrices, b_1, b_2 represent the bias vectors. Further, $\tanh$ and σ represent the hyperbolic tanh and sigmoid activation function respectively.

We train the Relation Net parameters using episodic binary logistic regression loss. To this end, for each bag, we create all possible pairs of l_i and l_j such that they belong to different images in the bag. A pair of l_i and l_j is positive if the predicates of both are the same as the common latent predicate of the bag; otherwise, it is negative.

We finally compute the pairwise cost as negative of the learned similarity metric, i.e.,

$$\Psi_{uv}(l_{ut_1}, l_{vt_2}, \boldsymbol{\theta}) = -R_{\boldsymbol{\theta}}(f_{l_{ut_1}}, f_{l_{vt_2}}). \tag{6}$$

3.3. Inference

The problem of finding the global optimal solution for the optimization function in Equation 2 is an NP-hard problem. The cardinality of the label set of an image is $p_u \times (p_u - 1)$ where p_u is number of object proposals in image-u. Therefore, a brute force technique to find the optimal solution to this labeling problem will take $O\big((\prod_{u=1}^{b} p_u^2)\big)$ time. There are a few approximation algorithms available for solving these kinds of problems [15, 2, 26]. However, we adopt a greedy inference algorithm proposed by Shaban et al. [26] due to its proven superiority over other approximation algorithms.

4. Experiments and Results

4.1. Datasets and Experimental Setup

To quantitatively compare and show the robustness of our proposed approach, we have used two public datasets for all our experiments. Details of both of them can be found in successive paragraphs.

VrR-VG [20]: Visually relevant relationships dataset (VrR-VG in short) is derived from the Visual Genome [17] by removing all the statistically and positionally-biased visual relationships. It contains 58,983 images, 23,375 visual relationship tuples, and 117 unique predicates. Out of these 117, we use 100 for training and 17 for testing. It has become a de facto choice for performing any study on visual relationships.

VG-150 [33]: To test the robustness, we further show results on VG-150. This dataset contains 150 object categories and 50 predicate classes. Out of the 50 predicates, we use 40 and 10 for training and testing, respectively.

To get object proposals for an image, we use Faster R-CNN [23] trained on Visual Genome [17]. We select the

Method	Bag size	VrR-VG		VG-150	
		Bag-CorLoc (%)	VR-CorLoc (%)	Bag-CorLoc (%)	VR-CorLoc (%)
Concat + Cosine Similarity	2	55.90	72.16	50.00	71.42
	4	31.57	70.86	24.40	65.58
	8	30.65	76.85	18.75	67.33
VTransE + Cosine Similarity	2	59.84	73.34	55.67	74.90
	4	36.23	74.20	33.45	71.78
	8	34.64	82.56	26.67	70.85
Concat + Relation Net	2	61.72	75.61	54.55	71.85
	4	35.28	74.02	38.62	72.19
	8	31.24	76.38	29.15	75.55
Our Approach	2	**63.40**	**78.99**	**61.10**	**75.82**
	4	**48.06**	**76.12**	**42.30**	**79.15**
	8	**45.48**	**84.07**	**37.61**	**79.96**

Table 2. **Visual Relationship Co-localization results on unseen predicates.** We observe that our proposed approach outperforms relevant baselines by a significant margin. It shows effectiveness of visual relationship embedding and metric-based meta learning approach to compute visual relationship similarity as components in our approach, and our overall optimization framework.

top-100 most confident object proposals from the Faster R-CNN and remove the "background" class proposals. We further do Non-max suppression with a 0.5 threshold to refine the set of object proposals. Finally, we have 30-50 object proposals for each image. To create the label set for an image, we consider all possible ordered pairs of object proposals for that image as candidates for the typical visual relationship. Since we have almost 30-50 object proposals, we get around 1600 candidates for visual relationship co-localization in each image.

To obtain visual relationship embedding, we train VTransE [37] on visual relationships in VrR-VG containing only the 100 training predicates, and the remaining 17 predicates are only used for testing. To create an image bag of size b, we first select a predicate and then pick b images from the dataset such that each of the b images has at least one visual relationship with the selected predicate. In this way, we get a bag in which all the images share a common predicate. Note that while creating set of training bags (S_{train} such that $|S_{train}| = 10000$) and testing bags (S_{test} such that $|S_{test}| = 500$) we make sure that they do not share any common predicate i.e. $S_{train} \cap S_{test} = \phi$.

Performance Metrics: To evaluate performance of our approach, we use two different metrics, namely Visual Relation(VR)-CorLoc and Bag-CorLoc, both of which are inspired by widely-used localization metric CorLoc [7]. These performance measures are described below:

(i) Visual Relation-CorLoc: In an image, a visual relation candidate prediction is considered to be correct if both its visual subject and visual object localization are correct.[2] *VR-CorLoc is defined as the fraction of test images for which visual subject-object pairs are correctly localized.*

[2]An object proposal is considered to be correct if it's IoU with the target ground-truth bounding box is greater than 0.50.

(ii) Bag-CorLoc: If the common visual relations are correctly predicted for all of the bag images, then we consider that bag to be correctly predicted. *Bag-CorLoc is defined as the fraction of the total number of bags for which the visual subject-object pairs are correctly localized for all of its images.*

4.2. Ablations and Different Problem Settings

VRC being a novel task does not have any direct competitive method to compare with our proposed approach. However, to justify the utility of different modules of our approach and show robustness on unseen visual relationship localization, we perform several ablation studies. Further, we create different variations of the problem setting of VRC. These ablations and variations in problem setting are listed below: **(i) VtransE + Cosine Similarity**: To verify the utility of Relation Net in the proposed approach, as the first ablation, we replace the Relation Net that we use to compute the similarity between two of the relationship embeddings f_{l_i} and f_{l_j} by its like-for-like counterpart, i.e., cosine similarity.

(ii) Concat Embedding + Relation Net: To verify the utility of our relationship embedding encoder network in capturing a holistic understanding from the images, as a second ablation, we replace it with just a trivial concatenation of subject and object embeddings, i.e., $f_{l_i} = [s; o]$ where s and o represent the subject and object feature vector respectively. These feature vectors are the concatenation of FasterR-CNN features, bounding box coordinates, and object class probability scores to replace our relationship embedding. The rest of the method is identical to ours as we are utilizing Relation Net for computing similarity.

(iii) Concat Embedding + Cosine Similarity: In this experiment, we replace both the vital component of our approach, i.e., VtransE and Relation Net, by their like-for-

Method →	Concat + Cosine			VtransE+ Cosine			Concat+ Rel. Net			Our Approach		
Supervision ↓	Bag Size			Bag Size			Bag Size			Bag Size		
	2	4	8	2	4	8	2	4	8	2	4	8
No supervision	72.16	70.86	76.85	73.34	74.20	82.56	75.61	74.02	76.38	78.99	76.12	84.07
Subject Fixed	76.82	78.66	**81.27**	80.37	**83.12**	83.58	**81.07**	**82.88**	**84.60**	83.90	**88.25**	86.67
Subject-Object in one image	**77.03**	**80.20**	79.42	**83.33**	82.40	**84.07**	79.29	81.69	81.45	**87.44**	84.46	**86.95**

Table 3. **Effects of weak supervision on co-localization of relationships.** Here, we observe that just by giving a weak form of supervision, the visual relationship co-localization performance increases significantly for each ablation. The results correspond to VR-CorLoc %.

like counterparts and show how it performs compared to our proposed approach.

Further, In the original problem setting of VRC, only a bag of images is provided (no supervision). While we perform the experiment in this challenging setting, we also relax the problem setting a bit as follows in conducting additional experiments:

(i) Visual subjects in all the images are given: In this setting, along with the bag of images, we assume that a bounding box for the visual subject is also provided in each image. Our goal is to only co-localize those visual objects that connect the given subject via a common predicate in all the images of the bag.

(ii) Both visual subject-object in one image is given: In this setting, both visual subject and object bounding boxes corresponding to the common latent predicate are provided but only for one image of the bag. Given this, our goal is to co-localize visual subjects and objects in the remaining images of the bag.

We show results of these baselines and problem setting variations on datasets presented in Section 4.1, and compare them against our proposed approach in the next section.

4.3. Results and Discussion

We first do a quantitative analysis of our proposed approach in Table 2. We report Bag-CorLoc and VR-CorLoc (refer Section 4.1) in % for bag size varying from 2 to 8. We observe that the baseline approach, which uses cosine similarity as compared to Relation Network as a component to compute the similarity between visual relationship embedding performs remarkably inferior as compared to our approach. For example, baselines that use cosine similarity achieve maximum Bag-CorLoc as 34.64% vs. 45.48% in our approach for bag size = 8 on VrR-VG. This result establishes the utility of metric-based meta-learning in the form of Relation Net used in our framework. Further, to verify the effectiveness of a robust visual relationship embedding technique in our optimization framework, we use a very naive method of representing visual relationships, in other words, by just concatenating visual subject and object representation. Simple concatenation for representing visual relationships is less powerful. For example, it achieves 31.24% Bag-CorLoc for bag size = 8 on VrR-VG, which is significantly inferior to our approach, which by the virtue of the right choice of visual relationship embedding technique and metric-based meta-learning approach achieves 45.48% Bag-CorLoc and 84.07% VR-CorLoc on VrR-VG on bag size = 8. We notice similar performance gain with our approach in VG-150 as well.

We also did extensive experiments with minor tweaks in the original setting of VRC by relaxing it a bit. We have shown VR-CorLoc for all those experiments in Table 3 on the VrR-VG dataset. We can see that once we relax the strictness in problem setting a little bit, in other words, by providing subject boxes, the VR-CorLoc increases significantly for each of the ablation and prominently if we see our approach for bag size two and four, it increases to 83.90% and 88.25% from 78.99% and 76.12% respectively. In the other scenario where we relax the condition by only giving subject and object bounding box for only one image in the bag, the VR-CorLoc score increases to 87.44% and 84.46% from 78.99% and 76.12% for bag size two and four, respectively. This also shows that by providing slightly more supervision (either annotating bounding boxes for subject corresponding to a common predicate in all the images or annotating subject-object pair corresponding to a common predicate in one image), the visual relationship co-localization of our approach significantly improves.

We perform an extensive qualitative analysis of the proposed approach on VrR-VG. A selection of visual relationship co-localization results by our approach is shown in Figure 3.[3] Here we show a bag of images in each column. The subject and object co-localization on these bags is shown using green and yellow colors, respectively. We observe that our approach successfully co-localizes the visual subject and objects connected via a latent predicate by just looking into four images in the bag. Specifically, we observe that in the fourth column where the latent predicate is 'Following,' if we see for humans also, it would be a challenging task to localize this relationship in each image if we had never known of relationship beforehand and different combinations of subject and object following to each other, for example "a cow *following* to another cow" in row-1, "a sheep *following* to a man" in row-2 and so on. However,

[3]More visual results are presented in Supplementary Material.

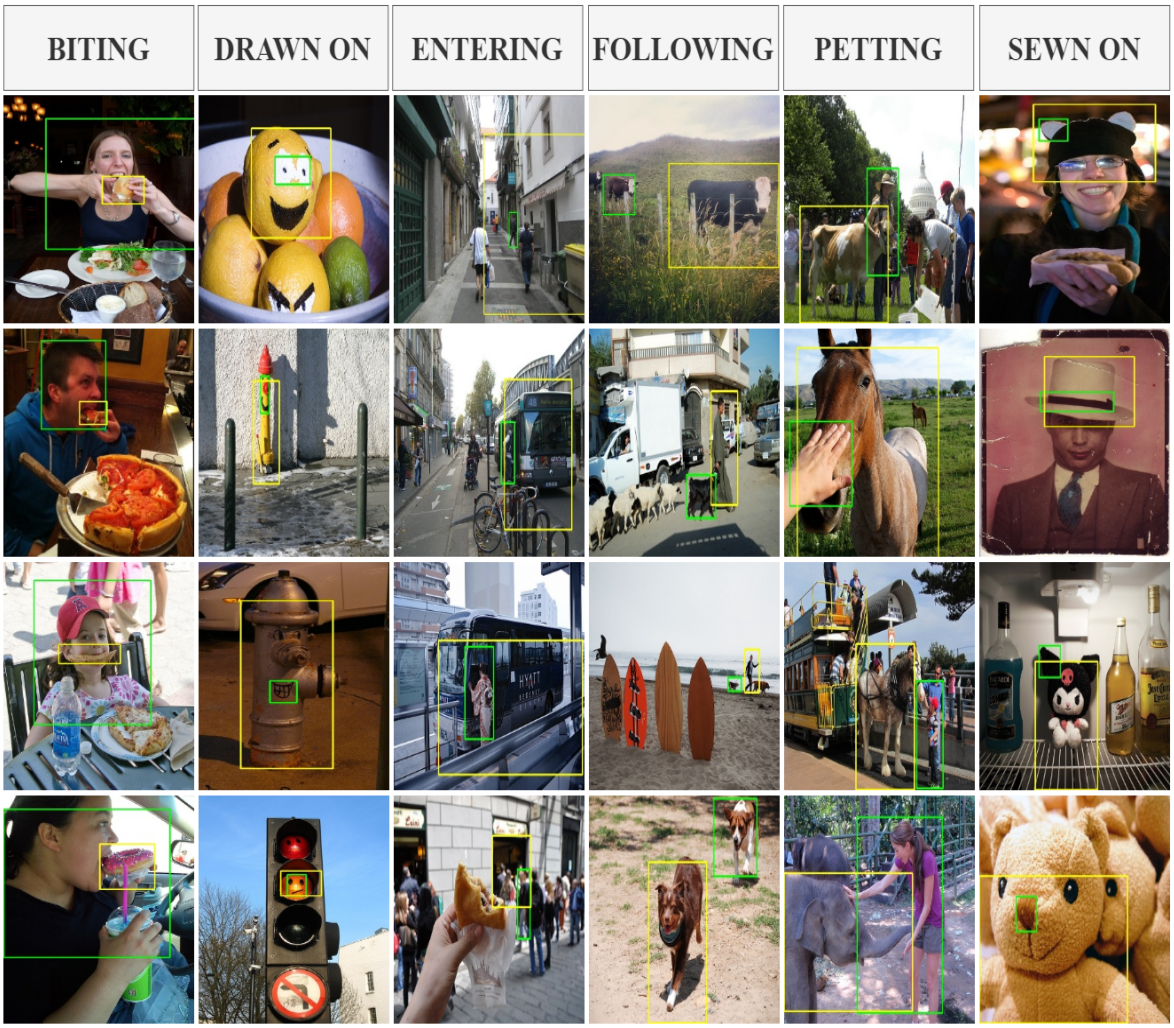

Figure 3. We show some of the qualitative results on the VrR-VG dataset. Each column is a bag of images (bag size = 4), having a common latent predicate in all its images. The common latent predicate is written on top of each column. Our approach localizes the visual subject-object pairs in each image of the column, which are connected through that common latent predicate by drawing bounding boxes around them. The green and yellow boxes correspond to the localized visual subject and object, respectively. It is to be noted that all of **these predicates are never seen during the training phase**. **[Best viewed in color and 200% zoom in]**.

our proposed approach does an excellent job in localizing these relationships. Note that all the relationships shown in Figure 3 are 'unseen' during the training phase.

To sum up, VRC is a challenging problem even for humans. We tackle VRC by posing it as a labeling problem and solving an optimization problem in a meta-learning framework. To verify our approach's robustness, we have provided an extensive comparison to all the relevant baselines. The results on challenging public datasets are affirmative and confirm our proposed approach's superiority over all the comparing baselines in most cases.

5. Conclusion

We presented a novel task, namely a few-shot visual relationship co-localization (VRC), and proposed a principled optimization framework to solve this by posing an equivalent labeling problem. Our proposed model successfully co-localizes many different visual relationships with reasonably high accuracy by just looking into few images. We also show visual relationship co-localization in two more exciting settings, firstly when the subject is known in all the images, and we have to co-localize objects. Secondly, when the subject and object pair is annotated for one image in the bag, and we need to transfer this annotation to the remaining images in the bag. In both these settings, our proposed

method has been found effective indicating utility of VRC in visual relationship discovery and automatic annotation. We firmly believe the novel task presented in this paper and benchmarks shall enable future research avenues in visual relationship interpretation and, thereby, scene understanding.

References

[1] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2002.

[2] Martin Bergtholdt, Jörg Kappes, Stefan Schmidt, and Christoph Schnörr. A study of parts-based object class detection using complete graphs. *International journal of computer vision*, 2010.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *neurIPS*, 2020.

[4] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 2018.

[5] Kai Chen, Hang Song, Chen Change Loy, and Dahua Lin. Discover and learn new objects from documentaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[6] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[7] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Localizing objects while learning their appearance. In *European conference on computer vision*, 2010.

[8] Gary Doran and Soumya Ray. A theoretical and empirical analysis of support vector machine methods for multiple-instance classification. *Machine learning*, 2014.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks, 2017.

[10] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. *arXiv preprint arXiv:1810.10147*, 2018.

[11] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.

[12] Tao Hu, Pascal Mettes, Jia-Hong Huang, and Cees GM Snoek. Silco: Show a few images, localize the common object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[13] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[14] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, 2015.

[15] Vladimir Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE transactions on pattern analysis and machine intelligence*, 2006.

[16] Ranjay Krishna, Ines Chami, Michael S. Bernstein, and Li Fei-Fei. Referring relationships. In *CVPR*, 2018.

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR*, 2016.

[18] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. Weakly supervised object localization with progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[19] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: an efficient subgraph-based framework for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[20] Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. Vrr-vg: Refocusing visually-relevant relationships. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[21] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multi-

ple instance learning. *arXiv preprint arXiv:1412.7144*, 2014.

[22] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. *arXiv preprint arXiv:1412.7144*, 2016.

[23] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, 2015.

[24] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, 2016.

[25] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*, 2018.

[26] Amirreza Shaban, Amir Rahimi, Shray Bansal, Stephen Gould, Byron Boots, and Richard Hartley. Learning to find common objects across few image collections. In *ICCV*, 2019.

[27] Yunhan Shen, Rongrong Ji, Shengchuan Zhang, Wangmeng Zuo, and Yan Wang. Generative adversarial learning towards fast weakly supervised detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[28] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning, 2017.

[29] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, 2018.

[30] Kevin Tang, Armand Joulin, Li-Jia Li, and Li Fei-Fei. Co-localization in real-world images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[31] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, 2016.

[32] Yaqing Wang, Quanming Yao, James Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning, 2020.

[33] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[34] Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Danet: Divergent activation for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[35] Leiming Yan, Yuhui Zheng, and Jie Cao. Few-shot learning for short text classification. *Multimedia Tools and Applications*, 2018.

[36] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[37] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5532–5540, 2017.

[38] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. Large-scale visual relationship understanding. In *Proceedings of the AAAI conference on artificial intelligence*, 2019.

[39] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[40] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[41] Bohan Zhuang, Lingqiao Liu, Yao Li, Chunhua Shen, and Ian Reid. Attend in groups: a weakly-supervised deep learning framework for learning from web data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.