IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 August 2024, accepted 22 August 2024, date of publication 28 August 2024, date of current version 10 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3450908

RESEARCH ARTICLE

Relationship-Aware Unknown Object Detection for Open-Set Scene Graph Generation

MOTOHARU SONOGASHIRA[1], MASAAKI IIYAMA[2], (Member, IEEE), AND YASUTOMO KAWANISHI[1], (Member, IEEE)

[1]Information R&D and Strategy Headquarters, RIKEN, Soraku, Kyoto 619-0288, Japan
[2]Faculty of Data Science, Shiga University, Hikone, Shiga 522-8522, Japan

Corresponding author: Motoharu Sonogashira (motoharu.sonogashira@riken.jp)

This work was supported by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (KAKENHI) under Grant JP21H03519, Grant JP22K17920, and Grant JP24H00733.

ABSTRACT Scene graph generation (SGG) aims to detect the relationships of objects in an image. Recently, it has been extended to open-set SGG, which also considers unknown objects unseen in a training phase and thereby enables various applications in complex real-world scenes. However, previous research on open-set SGG addressed unknown object detection simply by thresholding confidence scores from object classification trained only for known objects. In reality, these scores become low for both unknown objects and failure detections of the background since they look different from known objects. Therefore, the current state of the art of open-set SGG cannot distinguish unknown objects from backgrounds, thereby overlooking their relationships. In this paper, we propose a novel relationship-aware unknown detection technique. Our main idea is to exploit the fact that only foreground regions containing objects can have relationships with other regions. To this end, we define a Bayesian model on objects and relationships and derive an algorithm of variational inference, which propagates foregroundness between regions and region pairs to assign foreground regions that have more related objects and relationships. As the results of extensive experiments using a public benchmark for open-set SGG, the proposed technique outperformed previous methods, including the state-of-the-art thresholding technique, in the standard OSGDet metrics regardless of the SGG models with which the proposed technique was combined (e.g., +0.61 improvement in OSGDet@100 with the VCTree model).

INDEX TERMS Open-set, object detection, scene graph generation.

I. INTRODUCTION

Scene graph generation (SGG) is a problem of detecting the relationships of objects in an image. It enables a detailed understanding of complex scenes and has a wide range of applications, such as image retrieval, visual question answering, and robot control [1]. Recently, it has been extended to a new problem called open-set SGG [2], which also aims to detect relationships involving unknown objects whose instances are absent in training images. Owing to the consideration of unknown objects, open-set SGG is more robust than conventional closed-set SGG in complex real-world scenes and thus important for practical SGG applications.

The previous research on open-set SGG only examined baseline methods that detect unknown objects by simple thresholding of confidence scores from object classification, which is trained only for known objects [2]. In reality, these scores become low for image regions that contain not only unknown objects but also failure detections from the background without any objects since both of them tend to have different image features from known objects. Therefore, the current state of the art of open-set SGG cannot distinguish unknown objects and backgrounds, thereby failing to detect the relationships involving the unknown objects.

In this paper, we propose a novel relationship-aware unknown detection technique, which fully makes use of

The associate editor coordinating the review of this manuscript and approving it for publication was Sudhakar Radhakrishnan.

SGG-specific relationship information. Our key observation is that backgrounds cannot have relationships; thus, our idea is to accept regions that are more probably related to other regions as foregrounds and reject others as failure detections from the background. Specifically, we build a Bayesian model that describes the mutual dependencies of objects and relationships in each image and derive a majorization-minimization (MM) algorithm of variational inference. This algorithm iteratively refines the probabilities of known classes produced by an arbitrary closed-set SGG model to separate known and unknown objects in foregrounds while assigning regions with more related objects and relationships to foregrounds. This principle enables the distinction of unknown objects from failure detections of the background, and thereby, the correct detection of their relationships. The results of our extensive experiments using various SGG models confirm the effectiveness of the proposed unknown detection technique, which can detect relationships involving unknown objects that previous techniques cannot. The proposed technique is also model-agnostic, i.e., in principle it can be combined with an arbitrary closed-set SGG model, thereby converting the model into an effective open-set method, as demonstrated by our experimental results.

The major contributions of this work are as follows:

- We present the first relationships-aware methodology for unknown object detection, which we derive by making full use of the mathematically-principled methodology of variational Bayesian inference.
- On the basis of this methodology, we develop the novel technique for open-set SGG, which is model-agnostic (combinable with any SGG models in principle) and more effective than previous techniques in detecting relationships of unknown objects, being able to distinguish unknown objects from backgrounds, as verified by our extensive experiments.
- From a practical perspective, the proposed technique can detect the relationships of unknown objects (in particular, objects dissimilar to known objects) that could not be detected previously, as will be exemplified in the experiments. This benefits various applications based on open-set SGG that require detailed understanding of complex scenes in the real world, where the presence of unknown objects are inevitable.

The rest of this paper is organized as follows. First, in Section II, we review related studies on closed- and open-set SGG (Section II-A), as well as other open-set problems (Section II-B). Then, in Section III, we describe the problem formulation (Section III-A) and the proposed technique for open-set SGG, which consists of a Bayesian model (Section III-B) and an iterative algorithm of variational inference on the model (Section III-C). In Section IV, we present the setting (Section IV-A) and results (Section IV-B) of our extensive experiments on open-set SGG, where we evaluated the proposed technique in comparison with previous techniques combined with various SGG models, along with an ablation study to analyze the details of the proposed technique (Section IV-C). Finally, we summarize this paper with perspectives on future work in Section V.

II. RELATED WORK

In the following, we first review existing studies on SGG in Section II-A. Specifically, we begin with studies on closed-set SGG, e.g., various models and learning techniques. Then, we continue to discuss its extension to open-set SGG, focusing on its standard evaluation protocol and previous methodology, upon which we build this study. Furthermore, we review studies on similar open-set problems, e.g., open-set object detection in Section II-B, clarifying their differences from open-set SGG addressed in this paper.

A. SCENE GRAPH GENERATION

SGG is an extension of image recognition (image-wise classification) and object detection (localization of object regions followed by region-wise classification). Various SGG models based on deep neural networks have been proposed [3], [4], [5], [6], [7], [8], [9], [10] to detect relationships along with objects. SGG-specific learning techniques such as special losses and unbiasing strategies [7], [11], [12], [13], [14], [15], along with techniques that enhances the labels in training datasets themselves [16], have also been developed. In this study, we focus on unknown objects in open-set SGG and apply the proposed unknown detection technique to various representative SGG models; thereby, we show the generality of our model-agnostic unknown detection technique, while in principle it is applicable to any models, including those not considered in this paper. Meanwhile, the SGG-specific learning techniques mainly focus on relationships only, thus being orthogonal to this study. Since the proposed technique only modifies the outputs from an already-trained SGG model, it would be straightforward to combine many such learning techniques, although it is out of the scope of this study.

The only previous study on open-set SGG [2] focused on formulating the new open-set problem, proposing an evaluation protocol, and presenting initial experimental results. Consequently, it compared baseline methods built by combining closed-set SGG models with a simple unknown detection technique based on thresholding of output probabilities from object classification. While such a technique has extensively been used in open-set classification problems as a strong baseline [17], we must accept foregrounds only and reject backgrounds in SGG. Consequently, the effectiveness of the previous technique is limited since it tends to confuse unknown objects and backgrounds due to their visual difference from known objects. In this paper, we propose a novel SGG-specific relationship-aware technique for unknown object detection, which also uses output probabilities from relationship classification. To the best of our knowledge, this is the first technique that can exploit rich information from the relationship classification for unknown object detection. By making full use of relationships along with objects, the proposed inference technique is able to

distinguish unknown objects from backgrounds and thereby outperform the previous thresholding technique that can make use of objects only. In addition, we perform extensive experiments using various SGG models, including but not limited to all those considered in the previous study. Note that open-set SGG is not to be confused with open-vocabulary SGG [18], which involves classifying objects into individual unseen classes and thus requires additional information to associate them with seen classes, as in zero-shot learning.

B. OPEN-SET DETECTION

Open-set classification has recently been extended to open-set object detection [19], [20], which assumes multiple objects in each image. Technically, unknown detection in open-set object detection only considers objects but not their relationships, thus being simpler than the proposed relationship-aware technique.

Overall, the state-of-the-art methods for open-set tasks (including open-set SGG [2]) lack the ability to detect unknown objects by exploiting the rich context information in their relationships. To fill this gap, we propose a relationship-aware technique of unknown object detection for the first time, thereby realizing effective detection of relationships involving unknown objects.

III. RELATIONSHIP-AWARE UNKNOWN OBJECT DETECTION

A. OPEN-SET SCENE GRAPH GENERATION

We first review the problem setting of open-set SGG [2]. For each image, we localize each image region $i \in \mathcal{J}$ that may contain an object and classify it into an object class in $\mathcal{K} \cup \{u, b\}$, where $\mathcal{J} = \{1, \ldots, n\}$ is the set of the indices for n regions, $\mathcal{K}$ is the set of known object classes, u is an unknown object class, and b is a background class, which indicates the absence of any object in the region. We also classify each region pair $ij \in \mathcal{P}$ into a relationship class in $\mathcal{F} \cup \{b\}$, where $\mathcal{P} = \{ij\}_{i,j \in \mathcal{J} \text{ s.t. } i \neq j}$ is the set of the indices for $n(n-1)$ region pairs, $\mathcal{F}$ is the set of foreground relationship classes, and the background class b indicates the absence of any relationship in the pair. Since relationships are represented by subject-predicate-object word triplets [3], we distinguish the bidirectional combinations of two regions as different pairs (ij and ji) but ignore the self-loop pair of the same region (ii or jj), hence the definition of $\mathcal{P}$.

To solve this problem, we utilize a closed-set SGG model that takes the image and returns the probabilities of the object classes for each region i and those of the relationship classes for each region pair ij. In particular, we use the known object and foreground relationship probabilities, denoted by P^o_i and P^r_{ij}, respectively, where $o \in \mathcal{K}$ and $r \in \mathcal{F}$. In addition, the model selects one known object class $c_i \in \mathcal{K}$ and one foreground relationship class $c_{ij} \in \mathcal{F}$, which are basically the maximum-score classes after resolving region overlapping [6], thereby enabling classification. However, the unknown object class u is not included in these object classes.

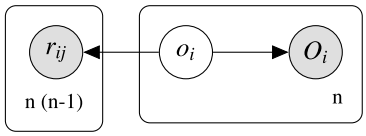


FIGURE 1. Graphical-model representation of the Bayesian model of the proposed inference. Gray and white circles denote the observed and hidden variables. $O_i \in \{k, b\}$ and $o_i \in \{k, u, b\}$ are object classes input from a model and output of the inference, respectively, and $r_{ij} \in \{f, b\}$ is the relationship class from the model, where k, u, f, and b are the known, unknown, foreground, and background classes, respectively; n and $n(n-1)$ are the numbers of regions and their pairs, respectively.

This is because the model cannot see unknown objects in training by the definition of the open-set problem [17] and thus cannot tell if each region belongs to the unknown class. To realize the open-set SGG, we invoke an unknown detection technique, whose input from the SGG model is both of the known object and foreground relationship class probabilities for each region and region pair, along with the selected known object class for each region, while the output is a known or unknown class and its foregroundness]core for each region. As with typical open-set techniques [17], this unknown detection is performed in testing only, thereby being training-free except for hyperparameter tuning by validation. Note that the training-free nature of the proposed technique also eliminate the risk of overfitting to training data. In addition, the hyperparameter tuning does not overfit to the types of relationships in validation data since the proposed technique does not distinguish between individual foreground relationship classes in $\mathcal{F}$ but only considers whether each pair is foreground or background, owing to our model design in Section III-B.

In the rest of this section, we present the proposed inference-based unknown detection technique, which can distinguish whether each region belongs to the unknown object or background class, unlike the previous thresholding technique [2].

B. MODEL

We build a Bayesian model to describe the mutual dependencies of the objects and relationships in the image while constraining background regions without objects not to have relationships with other regions. We embed this relationship constraint as a distribution in our Bayesian model, which makes known and unknown object probabilities higher for regions with more related objects and relationships, thereby accepting foregrounds and rejecting backgrounds. Note that this Bayesian model is different from those used in the previous Bayesian approaches to closed-set SGG, which did not consider unknown objects but had other purposes such as model design [5] and unbiasing [12]. We depict this model graphically in Fig. 1.

1) OBSERVED DISTRIBUTIONS

a: OBSERVED OBJECT DISTRIBUTION

To reduce the amount of information to be estimated, we do not distinguish individual known object classes in $\mathcal{K}$ and

merge them into single known-objects class k. Note that we use this class only in the proposed inference and will restore individual known classes in the finalization described in Section III-C2. Let $\mathcal{O}^{\setminus u} = \{k, b\}$ be the set of the object classes other than the unknown u, whose probability is not provided by the closed-set SGG model since it can classify objects into known and background classes only. After summing the input probabilities over $\mathcal{K}$ into the single probability of k, we have the observed object distribution $p(O_i)$ for each region i, where $O_i \in \mathcal{O}^{\setminus u}$, $p(O_i = k) = \sum_{o \in \mathcal{K}} P_i^o$, and $p(O_i = b) = 1 - p(O_i = k)$. We assume the independence between regions, i.e., $p(O_{ij}) = p(O_i)p(O_j)$, where we have introduced shorthand notation for paired classes $O_{ij} = \{O_i, O_j\}$.

b: OBSERVED RELATIONSHIP DISTRIBUTION

Similarly to the known object classes, we do not distinguish individual foreground relationship classes in $\mathcal{F}$ and merge them into single foreground class f. Let $\mathcal{R} = \{f, b\}$ be the set of the relationship classes. After summing the input probabilities over $\mathcal{F}$ into the single probability of f, we have the observed relationship distribution $p(r_{ij})$ for each region pair ij, where $r_{ij} \in \mathcal{R}$, $p(r_{ij} = f) = \sum_{r \in \mathcal{F}} P_{ij}^r$, and $p(r_{ij} = b) = 1 - p(r_{ij} = f)$.

2) LATENT DISTRIBUTIONS

a: LATENT OBJECT DISTRIBUTION

Let $\mathcal{O} = \{k, u, b\}$ be the set of all object classes, including the unknown u. To obtain the separate probabilities of these classes, we estimate latent object distribution $q(o_i)$ for each region i, where $o_i \in \mathcal{O}$. As with the observed distributions, we assume the independence between regions, i.e., $q(o_{ij}) = q(o_i)q(o_j)$, where $o_{ij} = \{o_i, o_j\}$.

b: CONDITIONAL OBJECT DISTRIBUTION

To associate the observed O_i and the latent object class o_i, we also define conditional object distribution $q(O_i|o_i)$ for each region i. Again, we assume the independence between regions, i.e., $q(O_{ij}|o_{ij}) = q(O_i|o_i)q(O_j|o_j)$.

c: LATENT RELATIONSHIP DISTRIBUTION

To associate objects with relationships and impose the relationship constraint, we introduce conditional relationship distribution $q(r_{ij}|o_{ij})$ for each region pair ij, which is conditioned on the latent object classes of regions i and j, defined as follows:

$$q(r_{ij}|o_{ij}) = \begin{cases} q^f(r_{ij}) & \text{if } o_i, o_j \in \mathcal{O}^f, \\ [r_{ij} = b] & \text{otherwise,} \end{cases} \tag{1}$$

where $r_{ij} \in \mathcal{R}$ as in the observed distribution, $q^f(r_{ij})$ is the latent relationship distribution that is activated when both objects are foreground, $\mathcal{O}^f = \{k, u\}$ is the set of the foreground object classes, and brackets $[\cdot]$ denotes the logical function whose value is one if its argument is true and zero otherwise. This definition states that the pair is always background with probability one if one or both of the two regions are background. Thus, this distribution represents our constraint that non-object regions never have relationships in a Bayesian way. Meanwhile, since we have no prior knowledge on the case where both regions are foreground, we estimate $q^f(r_{ij})$.

C. INFERENCE

To estimate the latent distributions, we consider to minimize the Kullback–Leibler (KL) divergence [21] between the observed and latent marginal relationship distributions, i.e., $L_{\text{orig}} = \sum_{ij \in \mathcal{P}} \mathrm{KL}\left(q(r_{ij}) \| p(r_{ij})\right)$, where we consider the divergence for each region pair and take their sum over all region pairs. While we may obtain the latent distribution $q(r_{ij})$ by marginalizing the product of the latent distributions in our model, it is not obvious how to optimize the nonlinear objective function efficiently. Furthermore, this divergence contains no observed distributions in our model other than the relationship distribution $p(r_{ij})$, preventing us from making use of the observed object distribution $p(O_{ij})$.

Instead of directly minimizing the original KL divergence, we derive its upper bound as follows:

$$\begin{aligned} L_{\text{orig}} &= \sum_{ij \in \mathcal{P}} \mathrm{E}_{q(r_{ij})}\left[\ln \frac{q(r_{ij})}{p(r_{ij})}\right] \\ &= \sum_{ij \in \mathcal{P}} \mathrm{E}_{q(r_{ij}, O_{ij}, o_{ij})}\left[\ln \frac{q(r_{ij})}{p(r_{ij})}\right] \qquad (2) \\ &= \sum_{ij \in \mathcal{P}} \mathrm{E}_{q(r_{ij}, O_{ij}, o_{ij})}\left[\ln \frac{q(r_{ij}, O_{ij}, o_{ij})}{p(r_{ij})p(O_{ij}, o_{ij}|r_{ij})}\right. \\ &\quad \left. - \ln \frac{q(r_{ij}, O_{ij}, o_{ij})}{q(r_{ij})p(O_{ij}, o_{ij}|r_{ij})}\right] \qquad (3) \\ &\le \sum_{ij \in \mathcal{P}} \mathrm{KL}\left(q(r_{ij}, O_{ij}, o_{ij}) \| p(r_{ij}, O_{ij}, o_{ij})\right) = L, \qquad (4) \end{aligned}$$

where $\mathrm{E}_{\cdot}[\cdots]$ denotes the expectation with respect to the distribution in its subscript. Here, we have applied Gibbs' inequality (i.e., the nonnegativity of the KL divergence) [21] to the second term in Eq. (3) and also defined the observed joint distribution in Eq. (4) as the product of the observed distributions in our model and auxiliary object distribution $p(o_{ij}|O_{ij}, r_{ij})$:

$$p(r_{ij}, O_{ij}, o_{ij}) = p(r_{ij})p(O_{ij})p(o_{ij}|O_{ij}, r_{ij}). \tag{5}$$

Then, we ensure that the bound is tight, i.e., close to the original KL divergence, by optimizing the bound with respect to the auxiliary distribution. Meanwhile, the latent joint distribution is simply the product of all latent distributions in our model:

$$q(r_{ij}, O_{ij}, o_{ij}) = q(r_{ij}|o_{ij})q(O_{ij}|o_{ij})q(o_{ij}), \tag{6}$$

where we assume that the relationship class r_{ij} is conditionally independent from the observed object classes O_{ij} given the corresponding latent object classes o_{ij}. This can be seen as a MM algorithm [22], which optimizes a parameterized

surrogate function, although the general methodology of MM algorithms does not assume a specific problem. Specifically, we newly designed the Bayesian model (Section III-B), the original objective function (Eq. (2)) and the auxiliary distributions (Eq. (3)) for our problem of open-set SGG; hence, the resulting objective (Eq. (4)) and their solutions are also novel. In the following, we present the optimal solution for each distribution.

a: AUXILIARY OBJECT DISTRIBUTION

To tighten the bound L, we minimize it with respect to the auxiliary distribution $p(o_{ij}|r_{ij}, O_{ij})$ for each ij. This is achieved when the corresponding KL term in Eq. (4) is zero, which holds if its arguments are equal [21]. Equating the right-hand sides of Eqs. (5) and (6), we obtain the following proportionality:

$$p(o_{ij}|r_{ij}, O_{ij}) \propto q(r_{ij}|o_{ij})q(O_{ij}|o_{ij})q(o_{ij}), \tag{7}$$

where we have ignored constant factors with respect to o_{ij}. By normalizing the right-hand side with respect to o_{ij} (either numerically or analytically), we obtain the optimal solution of $p(o_{ij}|O_{ij}, r_{ij})$.

b: LATENT RELATIONSHIP DISTRIBUTION

For each ij, we extract the terms depending on $q^f(r_{ij})$ in the right-hand side of Eq. (4) after substituting Eqs. (5) and (6):

$$L = \sum_{o_i, o_j \in \mathcal{O}^f} q(o_{ij}) \, \mathrm{E}_{q^f(r_{ij})q(O_{ij}|o_{ij})} \times \left[\ln \frac{q^f(r_{ij})}{p(r_{ij})p(o_{ij}|r_{ij}, O_{ij})} \right] + \text{const.} \tag{8}$$

$$= \sum_{o_i, o_j \in \mathcal{O}^f} q(o_{ij}) \, \mathrm{E}_{q^f(r_{ij})} \left[\ln \frac{q^f(r_{ij})}{p(r_{ij})v(r_{ij})} \right] + \text{const.}, \tag{9}$$

where

$$v(r_{ij}) = \exp\left(\mathrm{E}_{q(O_{ij}|o_{ij})} \left[\ln p(o_{ij}|r_{ij}, O_{ij}) \right] \right), \tag{10}$$

and in the sum of Eq. (8) we have used Eq. (1), which states that $q^f(r_{ij})$ appears only when $o_i, o_j \in \mathcal{O}^f$. Noticing that the log-expectation term is a KL divergence except for the unnormalized denominator, we can minimize the right-hand side of Eq. (9) by making the numerator and denominator proportional:

$$q^f(r_{ij}) \propto p(r_{ij})v(r_{ij}). \tag{11}$$

By normalizing the right-hand side with respect to r_{ij}, we obtain the optimal solution of $q^f(r_{ij})$.

c: LATENT OBJECT DISTRIBUTIONS

For each i, we optimize L with respect to $q(o_i)$ and $q(O_i|o_i)$ simultaneously. First, we extract the terms depending on them in the right-hand side of Eq. (4) after substituting Eqs. (5) and (6):

$$L = \mathrm{E}_{q(O_i|o_i)q(o_i)} \left[\sum_{j \in \mathcal{J}_{\setminus i}} \mathrm{E}_{q(O_j|o_j)q(o_j)} \left[\sum_{kl \in \mathcal{P}_{ij}} \mathrm{E}_{q(r_{kl}|o_{kl})} \times \left[\ln \frac{q(r_{kl}|o_{kl})q(O_i|o_i)q(o_i)}{p(r_{kl})p(O_i)p(o_{kl}|O_{kl}, r_{kl})} \right] \right] \right] + \text{const.} \tag{12}$$

$$= 2(n-1) E_{q(O_i|o_i)q(o_i)} \times \left[\ln \frac{q(O_i|o_i)q(o_i)}{p(O_i)w(O_i, o_i)} \right] + \text{const.}, \tag{13}$$

where

$$[b] w(O_i, o_i) = \exp\left(\frac{1}{2(n-1)} \sum_{j \in \mathcal{J}_{\setminus i}} \mathrm{E}_{q(O_j|o_j)q(o_j)} \times \left[\sum_{kl \in \mathcal{P}_{ij}} \mathrm{E}_{q(r_{kl}|o_{kl})} \left[\ln \frac{p(r_{kl})p(o_{kl}|O_{kl}, r_{kl})}{q(r_{kl}|o_{kl})} \right] \right] \right), \tag{14}$$

$\mathcal{J}_{\setminus i} = \mathcal{J} \setminus \{i\}$ is the set of all region indices other than i, $\mathcal{P}_{ij} = \{ij, ji\}$ is the set of the two pair indices involving regions i and j, and in Eq. (12) we have decomposed the distributions of paired object classes O_{ij} and o_{ij}. Similarly to the latent relationship distribution, we can minimize the right-hand side of Eq. (13) by making the numerator and denominator proportional:

$$q(O_i|o_i)q(o_i) \propto p(O_i)w(O_i, o_i). \tag{15}$$

By normalizing the right-hand side with respect to O_i and o_i, we obtain the optimal solution of $q(O_i|o_i)q(o_i)$. Then, we can simply obtain $q(o_i)$ by marginalizing out O_i numerically, i.e., by computing $\sum_{O_i \in \mathcal{O}^{\setminus u}} q(O_i|o_i)q(o_i)$.

While the normalized explicit solution would be complicated, we can better interpret the inference by considering a simplified version. Assume that the auxiliary and conditional object distributions are unconditional, i.e., $p(o_{ij}|O_{ij}, r_{ij}) = p(o_{ij})$ and $q(O_i|o_i) = q(O_i)$. Then, the optimal auxiliary distribution becomes $p(o_{ij}) = q(o_{ij}) = q(o_i)q(o_j)$ since $L = \mathrm{KL}(q(o_{ij}) \| p(o_{ij})) + \text{const.}$ with respect to the distribution $q(o_{ij})$ (with which $q(r_{ij}|o_{ij})$ is also constant) from Eq. (4). Also, the optimal relationship distribution becomes $q^f(r_{ij}) = p(r_{ij})$ since the right-hand side of Eq. (10) is constant with respect to r_{ij}. With these distributions, the log-expectation term inside the second sum in Eq. (14) is as simple as $\ln q(o_i)q(o_j) + (1 - [o_i, o_j \in \mathcal{O}^f]) \ln b_{kl}$ since $\ln \frac{p(r_{kl})}{q(r_{kl}|o_{kl})}$ is equal to zero if $o_k, o_l \in \mathcal{O}^f$ and $\ln b_{kl}$ otherwise due to the relationship constraint in Eq. (1). After taking its expectation with respect to $q(O_j|o_j)q(o_j)$ and ignoring the constant terms $\mathrm{E}_{q(o_j)}[\ln q(o_j)] + \ln b_{kl}$ with respect to o_i, we substitute the resulting $w(O_i, o_j)$ into Eq. (15) and marginalize out O_{ij}; thereby, we obtain the following simplified optimal object distribution:

$$q(o_i) \propto \bar{q}(o_i) \exp\left(\frac{[o_i \in \mathcal{O}^f]}{2(n-1)} \sum_{j \in \mathcal{J}_{\setminus i}} f_j \left(-\ln b_{ij} b_{ji} \right) \right), \tag{16}$$

where $f_j = q(o_j \in \mathcal{O}^f) = \sum_{o_j \in \mathcal{O}^f} q(o_j)$, $b_{ij} = p(r_{ij} = b)$, $b_{ji} = p(r_{ji} = b)$, and overbar $\bar{\cdot}$ on q denotes the current estimate, which have resulted from the substituted auxiliary distribution. Note that the negative log factor is

increasing with respect to $f_{ij} = 1 - b_{ij}$ and $f_{ji} = 1 - b_{ji}$. We observe that the optimal solution is basically its current estimate updated with an exponential factor, which increases the foreground probabilities of region i when each paired region j or their pairs ij and ji have higher foreground probabilities. Therefore, as intended, we can accept the regions that have many relationships as foreground while rejecting the others as background. Meanwhile, since the probabilities of all foreground classes are uniformly updated, the ratio between the known and unknown object probabilities does not change. This limitation is alleviated in the full version in Eq. (15) owing to the use of the observed object class O_i, which does not appear in Eq. (16) due to the simplifying assumptions, as will be seen experimentally in Section IV-C.

Since the optimal solutions depend on each other, we initialize the latent distributions and iteratively update them one by one, i.e., fixing the other distributions than the one being currently updated. Here, to facilitate efficient parallel computation, we update the object distributions for all regions simultaneously, thereby performing fixed-point iteration, where each distribution depends on the current estimates of the others. Similarly, we simultaneously update the relationship distributions for all region pairs, which do not depend on each other. Before updating each set of distributions, we substitute the current auxiliary distribution in Eq. (7) into its optimal solution to tighten the bound and also simplify computation.

Algorithm 1 summarizes the algorithm of the proposed inference-based unknown detection. Intuitively, this algorithm propagates the foregroundness between regions and pairs so that each region becomes foreground if it is involved with more foreground regions in more foreground pairs. This is realized by the model (Section III-B) that describes the dependency of object and relationship classes between regions and pairs under the constraint ($q(r_{ij}|o_{ij})$ in Eq. (1)) that only foreground regions (objects) can constitute foreground pairs (relationships) while adding the unknown class to the known and background object classes predicted by the SGG model (extending object class O_i to o_i). More specifically, at each iteration, the latent object distributions $q(O_i|o_i)q(o_i)$ of each regions are updated using other distributions, i.e., the observed object and relationship distributions $p(O_i)$ and $p(r_{ij})$ from the SGG model and the current estimates of the latent relationship distributions $q^f(r_{ij})$, focusing on how likely the other regions paired with this region and the pairs themselves are foreground. For example, if the region has many relationships (i.e., foreground pairs, which necessarily involves foreground regions due to the constraint), its foreground probability is increased (as seen in the simplified version in Eq. (16)). Similarly, the latent relationship distribution $q^f(r_{ij})$ of each pair is updated using the observed distributions and the current estimate of the object distributions, focusing on how likely the regions involved in the pair are foreground, e.g., if the two regions involved in the pair are likely to be foreground, the pair itself is more likely to be foreground (i.e., has a relationship). This will further reinforces the confidence that the involved regions are foreground in the next iteration. Substitution of the currently-optimal estimate of the auxiliary distribution $p(o_{ij}|r_{ij}, O_{ij})$ (Eq. (7)) in each update ensures that the upper bound of the KL objective function (Eq. (4)) is as tight as possible, i.e., the bound approximates the original KL divergence well. We will discuss the number of iterations in Section IV-A.

Algorithm 1 Algorithm of variational-Bayesian inference for relationship-aware unknown object detection.

In the rest of this section, we describe the initialization and finalization parts of this algorithm.

Input: Object classes and probabilities $\{c_i, \{P_i^o\}_{o\in\mathcal{K}}\}_{i\in\mathcal{J}}$ and relationship probabilities $\{\{P_{ij}^r\}_{r\in\mathcal{F}}\}_{ij\in\mathcal{P}}$ from a SGG model

1: Sum up the input probabilities to obtain observed distributions $\{p(O_i)\}_{i\in\mathcal{J}}$ and $\{p(r_{ij})\}_{ij\in\mathcal{P}}$ according to Section III-B1.
2: Initialize latent distributions $\{q(O_j|o_i), q(o_i)\}_{i\in\mathcal{J}}$ and $\{q^f(r_{ij})\}_{ij\in\mathcal{P}}$ according to Section III-C1.
3: **repeat**
4: Update the latent object distributions $\{q(O_j|o_i), q(o_i)\}_{i\in\mathcal{J}}$ by Eq. (15) while substituting Eq. (7).
5: Update the latent relationship distributions $\{q^f(r_{ij})\}_{ij\in\mathcal{P}}$ by Eq. (11) while substituting Eq. (7).
6: **until** the number of iterations is reached.
7: Finalize the classes and scores according to Section III-C2.

Output: Object classes and their scores $\{c_i', s_i'\}_{i\in\mathcal{J}}$

1) INITIALIZATION

Before the inference, we initialize each latent object distribution using two scalar hyperparameters, i.e., foreground and unknown scales $a, t \in [0, 1] \subset \mathbb{R}$, whose values are constant among images and regions, as follows:

$$\begin{aligned} q(o_i = k) &= \frac{aK_i}{K_i + t}, \\ q(o_i = u) &= \frac{at}{K_i + t}, \\ q(o_i = b) &= 1 - a, \end{aligned} \tag{17}$$

where $K_i = p(O_i = k)$. Here, a determines the initial foreground probability, and t supplements the unknown probability missing in the observed distribution $p(O_i)$.

Meanwhile, to initialize each conditional object distribution $q(O_i|o_i)$, we use the uniform distribution except when $o_i = k$, in which case we use the observed distribution $p(O_i)$ to ensure that the first update takes account of the difference with respect to o_i. To initialize each relationship distribution $q^f(r_{ij})$, we simply use the uniform distribution.

2) FINALIZATION

After the inference, we set the class and score of each region by using the updated latent probabilities of the known and unknown object classes, denoted by $k_i = q(o_i = k)$ and $u_i = q(o_i = u)$, respectively. Since k_i is the sum over individual known classes, we split it into individual known probabilities using the original known probabilities from the SGG model before summed into K_i. Combining them with u_i, we obtain updated class-wise probability Q_i^o, which is equal to $\frac{k_i}{K_i} P_i^o$ if $o \in \mathcal{K}$ and u_i if $o = u$.

Then, we set the updated class c_i' of the region to the unknown class u if its updated probability Q_i^u is larger than that of the original class $Q_i^{c_i}$; otherwise, we keep the original known class c_i selected by the SGG model. Since $Q_i^u : Q_i^{c_i} = t : P_i^{c_i}$ at the initial state in Eq. (17), the proposed technique with no iterative updates assigns the same regions to the unknown class as the previous thresholding technique [2] with threshold t, which facilitates their comparison in Section IV-C.

Finally, we set the score s_i' of the region to the scaled probability of its updated class $(K_i + t)Q_i^{c_i'}$, where we multiply the normalizing denominators of the initial values in Eq. (17) to restore the score variations among regions when t is small. This score reflects the relationship-based foregroundness and thereby better distinguishes backgrounds than the original score $s_i = P_i^{c_i}$, which closed-set SGG without unknown detection or the thresholding technique [2] would assign to the region. Note that we do not modify the original foreground relationship class c_{ij} of each region pair and its score $s_{ij} = P_{ij}^{c_{ij}}$ since the updated latent relationship distribution $q^f(r_{ij})$ is conditioned by foreground object classes, while SGG metrics expect unconditional classes and scores [2].

IV. EXPERIMENTS

A. SETTING

We evaluated the effectiveness of the proposed inference-based unknown detection technique for open-set SGG in comparison with previous techniques. For fair comparison, we followed the evaluation protocol of the previous study on open-set SGG [2]. Specifically, we employed its open-set SGG dataset, which is a version of Visual Genome [1], the most widely-used SGG dataset, augmented with unknown objects. This dataset contains various natural images, typically featuring multiple objects with indoor or outdoor backgrounds, as exemplified in Fig. 2. The numbers of images are 46707, 11131, and 33135 for the training, validation, and testing splits, respectively, while on average each image contains 11.5 known objects, plus 10.7 unknown objects in the validation and testing splits only. As an evaluation metric, we employed open-set SGDet (OSGDet) [2], an unknown-aware extension of a standard close-set recall metric [3], which counts ground-truth region pairs that are correctly detected. As in the previous study [2], we considered multiple versions of the metric, i.e., OSGDet@20, @50, and @100,

(a) (b) (c)

FIGURE 2. Examples of (a) training, (b) validation, and (c) testing images. Each image depicts a scene with multiple objects of various classes, which may have relationships with each other. Meanwhile, the training images do not contain the objects of the unknown class.

where each number denotes the number of predictions per image. Here, the predicted pairs in each image are ranked by their scores, each of which is the product of one relationship score s_{ij} and two object scores s_i', s_j', and only the specified number of the highest-score pairs are used.

To construct each open-set SGG method, we combined a closed-set SGG model and an unknown detection technique. Here, we used one of three unknown detection technique: the dummy technique with no unknown detection (i.e., conventional closed-set SGG), the previous thresholding technique [2], and the proposed inference technique. Meanwhile, to prove the model-agnostic effectiveness of the proposed technique, we employed multiple SGG models compared in the previous open-set SGG study [2] along with others. To evaluate different SGG models in a consistent manner, we employed two publicly-available implementations for closed-set SGG, denoted by Implementation 1 [11] and 2 [5], respectively, and extended them to the open-set setting by adding unknown detection.

After training each model on the training images, we combined it with each technique and tuned its hyperparameters by optimizing OSGDet@100 over the validation images, following the open-set protocol [2]. While the thresholding technique has a threshold parameter only, the inference technique has two scale parameters and the number of iterations. To enable efficient tuning, we first fixed the foreground scale to uniform 0.5 and the number of iterations to 100 and tuned the unknown scale by coarse-to-fine grid search with stride 0.1 and then 0.01 in the value range $[0, 1]$. Then, fixing the unknown scale at its best value, we tuned the foreground scale in the same manner. Finally, fixing both scales, we fine-tuned the number of iterations by iteratively doubling and halving the current value until no improvement was observed. Note that these three hyperparameters are the only parameters of the proposed technique that affect the behavior of its algorithm; the other parameters, i.e., the probabilities of the discrete latent distributions on the object and relationship classes, are automatically determined thorough iterative optimization for each testing image. Furthermore, while these hyperparameters are determined on validation data before testing, their actual values do not influence the performance of the proposed technique so much; for example, the validation OSGDet@100 values of the VCTree model during the coarse-to-fine tuning were 15.95, 15.98, and 16.00 after determining the unknown, foreground scale, and the number

TABLE 1. Average evaluation metrics in the testing data for the combinations of unknown detection techniques and models in two implementations.

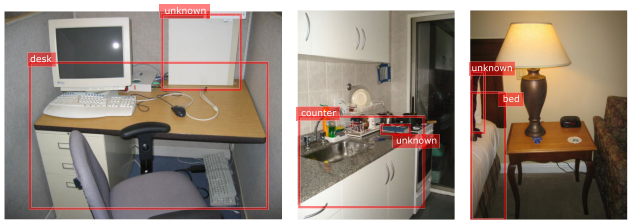
Impl.	Model	Technique	OSGDet @20	OSGDet @50	OSGDet @100
1 [8]	Freq [6]	None	6.73	9.05	10.74
		Threshold [2]	6.84	9.44	11.55
		Inference	**7.15**	**9.78**	**12.11**
	IMP [3]	None	6.33	9.03	10.96
		Threshold [2]	6.34	9.10	11.31
		Inference	**6.99**	**10.05**	**12.57**
	Motif [6]	None	8.89	11.54	13.23
		Threshold [2]	9.11	12.37	14.91
		Inference	**9.52**	**13.08**	**15.75**
	VCTree [8]	None	8.68	11.20	12.81
		Threshold [2]	8.85	12.02	14.45
		Inference	**9.04**	**12.48**	**15.06**
2 [5]	MSDN [4]	None	5.53	7.11	8.21
		Threshold [2]	5.77	7.73	9.21
		Inference	**6.11**	**8.28**	**9.87**
	GRCNN [5]	None	5.92	7.43	8.27
		Threshold [2]	6.18	8.22	9.70
		Inference	**6.44**	**8.63**	**10.03**
	RelDN [7]	None	5.79	7.43	8.54
		Threshold [2]	6.04	8.00	9.50
		Inference	**6.34**	**8.53**	**10.08**

of iterations, respectively, which indicates consistent but moderate improvement at each stage and thereby the stability of the proposed algorithm with respect to hyperparameters.

B. RESULTS

Table 1 summarizes the resulting metric values averaged over the testing images. Note that we do not compare the performances of different models but of the unknown detection techniques for each model. From these results, we can see that the proposed inference technique always outperformed no unknown detection (denoted by None) and the thresholding technique, regardless of the model and the number of predictions. Thus, we can conclude that the proposed technique is more effective than the existing techniques in a model-agnostic manner, being able to turn any closed-set SGG models into an open-set method. We note that newer models did not necessarily achieve higher metric values due to their different design goals. For example, VCTree [8] was more focused on dealing with the effect of biased class distributions, which is not reflected in the standard metrics. Still, we can easily combine such models with the proposed technique for better open-set performance, as exemplified here. We also note that Implementation 1 had the tendency to yield higher metric values than Implementation 2, although we evaluated them under the same open-set protocol [2]. This can be attributed to the differences in their implementation details, and the effectiveness of the proposed technique for each model were consistent.

Regarding computational efficiency, the previous thresholding [2] and proposed inference techniques took 0.013 and 0.392 seconds, respectively, per testing image on average using an Intel Core i9-10980XE CPU and a NVIDIA Tesla V100 GPU. The additional computational time of the proposed technique is due to the iterative nature of its algorithm in Algorithm 1, which can be regarded as the cost of its increased effectiveness shown in Table 1. Note that the number of iteration in this experiment is the optimal value determined by hyperparameter tuning for each model. In practice, if speed is a matter of concern, we can reduce the number of iterations without sacrificing the effectiveness so much, This is because the proposed method already outperforms the previous method at its initial state (i.e., with zero iteration) owing to its initialization scheme, as will be seen from the ablation study in Section IV-C.



(a) (b) (c)

FIGURE 3. Testing images where inference successfully detected unknown objects that thresholding could not. In each image, the proposed inference technique correctly detected an unknown object related to another object, not confusing them with backgrounds.

To better analyze the results of the proposed technique, we found out predicted region pairs that actually contributed to the improvement in OSGDet@100, i.e., pairs that newly became top-100 high-score predictions by the inference technique and also increased the number of correctly-detected ground truths. In summary, we had such pairs in 13 percent of the testing images in the case of VCTree (the newest model in Implementation 1). We visualize some of these images in Fig. 3, marking the predicted region pairs of interest. In Fig. 3(a), the inference successfully detected an unknown object, which was textureless and thus confused with backgrounds by the other techniques, on a desk. In Fig. 3(b), it detected an unknown object, which looked like an object but whose class could not be learned in training, on a kitchen counter. In Fig. 3(c), it could detect an unknown object, which was out of view and whose class was not obvious, on a bed. These results demonstrate the importance of relationships in robust unknown object detection and prove the ability of the proposed technique to distinguish unknown objects from backgrounds, which is our main aim and also one of the main contributions in this study, while being more effective than the previous techniques in terms of the standard metrics as show in Table 1.

We also visualize failure cases of the proposed technique combined with VCTree in Fig. 4. In Fig. 4(a), this method wrongly detected a cup on a shelf as an unknown, seemingly because it looked differently from typical cups seen in training. In Fig. 4(b), the method regarded an unknown object (headset) on a table as a glass, which would be the closest known class in terms of appearance. The proposed inference technique, whose main aim is to separate unknown objects

(a) (b)

FIGURE 4. Testing images where inference failed in detection. In each image, the proposed inference technique wrongly detected an known object as unknown or an unknown object as known.

from backgrounds, does not impose any prior knowledge on object classes and thus cannot distinguish between known and unknown classes well. Possible solutions include improving the class separation with additional training losses [19], or introducing trainable parameters into the Bayesian model to explicitly learn the different relationship statistics of known and unknown classes.

In addition to the above-mentioned models, we evaluated state-of-the-art closed-set SGG methods in the open-set setting. Note that the current state of the art of open-set SGG is the thresholding technique [2], and these closed-set methods were not intended for open-set SGG. Nevertheless, for the sake of completeness, we evaluate these methods under the open-set SGG protocol for the first time. Specifically, we evaluated Relationformer [23] and RelTR [24], which reportedly outperform other state-of-the-art methods [9]. These closed-set methods do not consider unknown objects and thus always treat them as known or background. We used their publicly-available implementations with default hyperparameters, while matching Relationformer's evaluation to the open-set benchmark [2]. As the result, Relationformer and RelTR scored 9.67 and 8.92 in testing OSGDet@100, respectively, underperforming most of the other models without unknown detection techniques ("None") in Table 1. These low open-set performances can be attributed to the one-stage design of the recent models, which tends to miss unknown objects [19], unlike two-stage models (e.g., VCTree) that first detect all objects using class-agnostic region proposal. These results reveals inherent limitation of the closed-set methods, which cannot detect unknown objects nor their relationships. Moreover, such recent transformer-based models with many parameters typically require more computational resource, thereby being less scalable with respect to the training dataset size than traditional SGG models evaluated in Table 1. In contrast, the proposed testing-time-only technique does not incur additional training cost while achieving higher open-set performance when combined with the scalable traditional models.

While the thresholding is a common technique for unknown detection in open-set tasks [17] not limited to SGG, we compared the proposed technique with another open-set technique [19], which uses an energy function as an alternative unknown-class score. We combined this technique with VCTree and tuned its threshold parameter using the validation data. As the result, it scored 13.27 in testing OSGDet@100, underperforming the thresholding [17] and the proposed technique in Table 1 and thereby indicating its unsuitability for SGG.

To summarize, the proposed technique outperformed the previous techniques both quantitatively (Table 1) and qualitatively (Fig. 3) regardless of combined models. These findings indicate that our inference-based approach greatly helps detect the relationships of unknown objects, including those the previous techniques could not detect, using general object and relationship classification results of SGG models to discover unknown objects without relying on their specific architecture and implementation details. We attribute this superiority of the proposed technique to its ability to consider relationships in distinguishing unknown objects from backgrounds, which the previous thresholding technique that consider only object classification scores is unable to do. In particular, the information on the existence of relationships, which we explicitly utilize in our algorithm via the specific Bayesian model with the relationship constraint in Section III-B, is effective in distinguishing unknown objects that are not similar to typical known objects, as exemplified in Fig. 3(a). This leads to the robustness of open-set SGG methods based on the proposed technique in complex scenes, which contrasts sharply with closed-set methods that easily fail in the presence of unknown objects by confusing them with known objects or ignoring as backgrounds. While the proposed technique has such strength in distinguishing unknown objects from backgrounds, as seen from the successful cases in Fig. 3, it also has weakness in distinguishing the unknown objects from known objects, as exemplified by the failure cases in Fig. 4. This is because our algorithm does not incorporate any prior knowledge on the difference between unknown and known objects in the model in Section III-B but infers their statistics only via image-wise iterative optimization, although this prior-free design also enables the proposed inference to deal with a wide range of scenes regardless of object types.

C. ABLATION STUDY

To further investigate the performance improvements achieved by the proposed technique, we compared three versions of the inference-based unknown detection: (1) The *initial* version, which directly uses the initial values of known and unknown probabilities in Eq. (17) without iterative updates. As mentioned in Section III-C2, the resulting classes are the same as the thresholding technique [2] (with a different threshold value after tuning) while the scores of the regions assigned to the unknown class are updated to a constant value proportional to the threshold. (2) The *simple* version, which uses the simplified object distribution in Eq. (16). The classes are still the same as the thresholding since the known-unknown score ratios are invariant between updates, but the scores are updated to reflect relationship-based foreground probabilities. (3) The *full* version, which uses the object distribution in Eq. (15) with

TABLE 2. Ablation study results comparing different versions of the proposed inference-based unknown detection technique.

Version	OSGDet@100
Initial	14.97
Simple	15.02
Full	**15.06**

no simplifications, thereby being able to update both classes and scores. In this experiment, we used the OSGDet@100 metric and the VCTree model.

Table 2 summarizes the results. The initial version was already better than the thresholding technique (shown in Table 1) by uniformly increasing the scores of all low-known-score regions to a constant value, but unable to distinguish unknowns and backgrounds. The simplified version made a further improvement by producing foreground-aware scores, but still no difference in classes. The full version outperformed the other versions, indicating that the proposed inference could make improvements in both classes and scores.

V. CONCLUSION

Open-set SGG, which detects the relationships of both known and unknown objects, is more practical than closed-set SGG. However, unknown detection in open-set SGG has ignored relationships and could not distinguish unknown objects from backgrounds, which has been a problem in detecting relationships of unknown objects. In this paper, on the basis of the idea that foreground regions cannot have relationships, we proposed relationship-aware unknown detection. Methodologically, we realized this by deriving an iterative algorithm of variational-Bayesian inference based on a Bayesian model, which increases the foregroundness of regions with more related objects and relationships. As demonstrated by the results of our extensive experiments, the proposed technique can consistently outperform previous techniques regardless of the model. These contributions of this study, i.e., the first relationships-aware methodology for unknown object discovery and the novel model-agnostic technique for open-set SGG, benefit various applications in complex real-world scenes.

As indicated by the experimental results, there are several pros and cons of the proposed method. The proposed inference approach can deal with various scenes, whether they are indoor or outdoor, since its Bayesian model does not assume specific image properties. One of the resulting advantages is that it can handle as many objects as the combined SGG model can find, since the Bayesian model has no predefined number of known or unknown objects, and the variational inference can naturally normalize the increased number of object distributions. Another advantage is that it can fix typical mistakes by the SGG model (e.g., mis-detected backgrounds with vivid color or complex texture, or undetected human-made objects that look different from natural objects) by adjusting their foregroundness according to relationships rather than appearances, on which the Bayesian model does not impose any prior knowledge. On the other hand, the main limitation of this study is that the proposed technique does not explicitly distinguish between known and unknown objects, as observed in Section IV, due to its prior-free nature. This limitation may be alleviated by introducing trainable components in the inference as a future improvement.

This study is the first attempt to develop a practically-effective method specialized to open-set SGG, taking account of both unknown objects and relationships. Therefore, various promising directions exist for future work. For additional performance boost, the most naive approach is to combine the proposed technique with newer SGG models or various learning techniques, both of which are orthogonal to unknown detection. A more complex but promising approach would be to integrate unknown object detection into a SGG model itself, which realizes unknown-object-aware relationship detection that can fully exploit the different statistics of known and unknown objects. Finally, extension to the open-world setting [19], i.e., iterative learning for distinguishing individual unknown classes as new known ones, will further enhance the real-world applicability of open-set SGG.

REFERENCES

[1] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, May 2017.

[2] M. Sonogashira, M. Iiyama, and Y. Kawanishi, "Towards open-set scene graph generation with unknown objects," *IEEE Access*, vol. 10, pp. 11574–11583, 2022.

[3] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3097–3106.

[4] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1270–1279.

[5] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph R-CNN for scene graph generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 670–685.

[6] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5831–5840.

[7] J. Zhang, K. J. Shih, A. Elgammal, A. Tao, and B. Catanzaro, "Graphical contrastive losses for scene graph parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 11527–11535.

[8] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to compose dynamic tree structures for visual contexts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 6612–6621.

[9] R. Li, S. Zhang, and X. He, "SGTR: End-to-end scene graph generation with transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 19464–19474.

[10] J. Lu, L. Chen, H. Guan, S. Lin, C. Gu, C. Wang, and G. He, "Improving rare relation inferring for scene graph generation using bipartite graph network," *Comput. Vis. Image Understand.*, vol. 239, Feb. 2024, Art. no. 103901.

[11] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 3713–3722.

[12] W. Li, H. Zhang, Q. Bai, G. Zhao, N. Jiang, and X. Yuan, "PPDL: Predicate probability distribution based loss for unbiased scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 19425–19434.
[13] J. Yang, C. Wang, L. Yang, Y. Jiang, and A. Cao, "Adaptive feature learning for unbiased scene graph generation," *IEEE Trans. Image Process.*, vol. 33, pp. 2252–2265, 2024.
[14] R. Zhang, G. An, Y. Hao, and D. O. Wu, "Bridging visual and textual semantics: Towards consistency for unbiased scene graph generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–18, Apr. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10502321
[15] M. Zhao, Y. Kong, L. Zhang, and B. Yin, "Class correlation correction for unbiased scene graph generation," *Pattern Recognit.*, vol. 149, May 2024, Art. no. 110221.
[16] L. Li, J. Xiao, H. Shi, H. Zhang, Y. Yang, W. Liu, and L. Chen, "NICEST: Noisy label correction and training for robust scene graph generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 10, 2024, doi: 10.1109/TPAMI.2024.3387349.
[17] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," 2016, *arXiv:1610.02136*.
[18] T. He, L. Gao, J. Song, and Y.-F. Li, "Towards open-vocabulary scene graph generation with prompt-based finetuning," in *Proc. ECCV*, Tel Aviv-Yafo, Israel, 2022, pp. 56–73.
[19] K. J. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian, "Towards open world object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5826–5836.
[20] A. Gupta, S. Narayan, K. J. Joseph, S. Khan, F. S. Khan, and M. Shah, "OW-DETR: Open-world detection transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 9225–9234.
[21] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed., New York, NY, USA: Springer, 2006.
[22] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Statistician*, vol. 58, no. 1, pp. 30–37, Feb. 2004.
[23] S. Shit, R. Koner, B. Wittmann, J. Paetzold, I. Ezhov, H. Li, J. Pan, S. Sharifzadeh, G. Kaissis, V. Tresp, and B. Menze, "Relationformer: A unified framework for image-to-graph generation," in *Proc. ECCV*, Tel Aviv-Yafo, Israel, 2022, pp. 422–439.
[24] Y. Cong, M. Y. Yang, and B. Rosenhahn, "RelTR: Relation transformer for scene graph generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 11169–11183, Apr. 2023.

MOTOHARU SONOGASHIRA received the B.S. degree in engineering, and the M.S. and Ph.D. degrees in informatics from Kyoto University, Kyoto, Japan, in 2013, 2015, and 2018, respectively. He is currently a Researcher with the Multimodal Data Recognition Research Team, RIKEN Guardian Robot Project. His research interests include image restoration, deconvolution, superresolution, optical flow, and variational Bayes.

MASAAKI IIYAMA (Member, IEEE) received the B.S. degree in engineering informatics and the M.S. and Ph.D. degrees in informatics from Kyoto University, in 1998, 2000, and 2006, respectively. Previously at ACCMS as a Research Associate, from 2003 to 2006. From 2006 to 2009, he was with the Graduate School of Economics, Kyoto University as an Assistant Professor, and as an Associate Professor, from 2009 to 2015, and ACCMS as an Assistant Professor, from 2009 to 2021. He is currently a Professor with the Faculty of Data Science, Shiga University. His research interests include computer vision, 3D modeling, and pattern recognition.

YASUTOMO KAWANISHI (Member, IEEE) received the B.Eng. degree in engineering, and the M.Inf. and Ph.D. degrees in informatics from Kyoto University, Japan, in 2006, 2008, and 2012, respectively. He became a Postdoctoral Fellow with Kyoto University, in 2012. In 2014, he moved to Nagoya University, Japan, as a Designated Assistant Professor, where he became an Assistant Professor, in 2015, and a Lecturer, in 2020. Since 2021, he has been the Team Leader of the Multimodal Data Recognition Research Team, RIKEN Guardian Robot Project. His main research interests include robot vision for environmental understanding and pattern recognition for human understanding, especially pedestrian detection, tracking, retrieval, and recognition. He is a member of IIEEJ and IEICE. He received the Best Paper Award from SPC2009 and the Young Researcher Award from the IEEE ITS Society Nagoya Chapter.