

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

Image-Collection Summarization using Scene-Graph Generation with External Knowledge

ITTHISAK PHUEAKSRI^{1,2} , MARC A. KASTNER³ , (Member, IEEE),
YASUTOMO KAWANISHI^{2,1} , (Member, IEEE), TAKAHIRO KOMAMIZU^{4,1} , (Member, IEEE),
ICHIRO IDE^{1,4} , (Senior Member, IEEE)

¹Graduate School of Informatics, Nagoya University, Nagoya, Aichi 464-8601 Japan (e-mail: phueaksri@cs.is.i.nagoya-u.ac.jp, ide@i.nagoya-u.ac.jp)

²Guardian Robot Project, Information R&D and Strategy Headquarters, RIKEN, Seika, Kyoto 619-0288 Japan (e-mail: yasutomo.kawanishi@riken.jp)

³Graduate School of Informatics, Kyoto University, Kyoto, Kyoto, 606-8501, Japan (e-mail: mkastner@i.kyoto-u.ac.jp)

⁴Mathematical and Data Science Center, Nagoya University, Nagoya, Aichi 464-8601 Japan (e-mail: taka-coma@acm.org)

Corresponding author: Phueaksri Itthisak (e-mail: phueaksri@cs.is.i.nagoya-u.ac.jp).

This work was supported in part by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, through Grants-in-Aid for Scientific Research JP21H03519 and JP23K16945.

ABSTRACT Summarization tasks aim to summarize multiple pieces of information into a short description or representative information. A text summarization task is a task that summarizes textual information into a short description, whereas in an image collection summarization task, also known as the photo album summarization task, the goal is to find the representative visual information of all images in the collection. In recent years, scene-graph generation has shown the advantage of describing the visual contexts of a single-image, and incorporating external knowledge into the scene-graph generation model has also given effective directions for unseen single-image scene-graph generation. Following this trend, in this paper, we propose a novel scene-graph-based image-collection summarization model. The key idea of the proposed method is to enhance the relation predictor toward relationships between images in an image collection incorporating knowledge graphs as external knowledge for training a model. To evaluate the proposed method, we build an extended annotated MS-COCO dataset for this task and introduce an evaluation process that focuses on estimating the similarity between a summarized scene graph and ground-truth scene graphs. Traditional evaluation focuses on calculating precision and recall scores, which involve true positive predictions without balancing precision and recall. Meanwhile, the proposed evaluation process focuses on calculating the F-score of the similarity between a summarized scene graph and ground-truth scene graphs which aims to balance both false positives and false negatives. Experimental results show that the use of external knowledge in enhancing the relation predictor achieves better results compared with existing methods.

INDEX TERMS Image collection summarization, multiple-image summarization, semantic images summarization, scene-graph generation, scene-graph summarization

I. INTRODUCTION

With the increase of digital content, especially images in the real world, image understanding tasks such as classification and retrieval have become more important than ever to make the contents easy to access. However, most existing research focuses on single-image understanding whereas understanding an image collection is still challenging. In recent years, the task of understanding an image collection is focused in various applications [1], such as semantic image retrieval [2], [3], Web-image concept understanding [4],

[5], and multiple-image summarization [6]–[9]. Generating a summarized scene graph also shows an advantage in visual storytelling [10], [11] and video summarization [12] applications. The typical first stage in understanding an image collection is to understand the overall context and find a representation of it, e.g., in the form of words, sentences, or scene graphs. Compared to other methods, a scene graph has the advantage of its ability to represent the contexts of images by describing objects and their relationships. The task of generating a scene graph is used in various tasks such

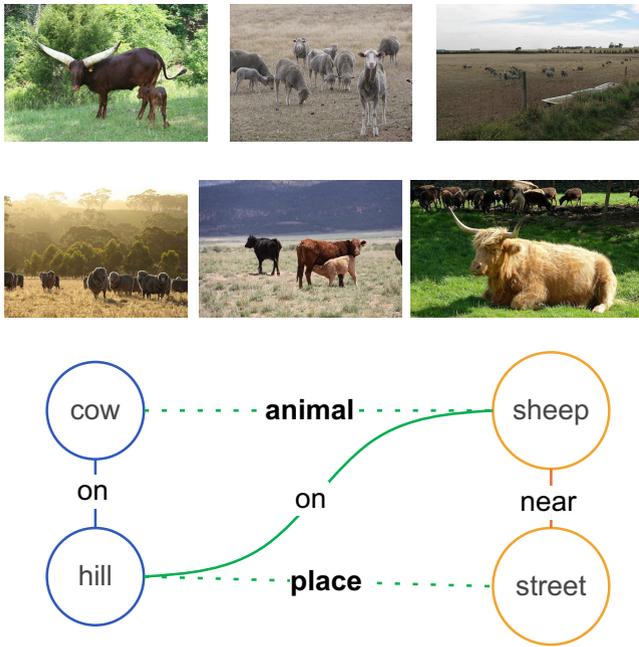


FIGURE 1. Example of generating a scene graph representation of an image collection. The dotted line represents the semantic relationships, and the solid green line shows the inferred relation of the summarized scene graph.

as single-image captioning [13], [14], image retrieval [3], [15], [16], and multiple-image context summarization [7]. However, scene-graph generation is commonly introduced to generate a scene graph of a single image. Whereas, summarizing an image collection into a summarized scene graph shows advantages in understanding the overall contexts and using it in image querying application [6]. However, the common challenge in scene graph summarization is estimating the relationships between different object category pairs detected in different images. In order to improve summarizing information of an image collection, we aim to understand the relationships between objects detected in different images by employing external knowledge graphs. Figure 1 shows an example of summarizing an image collection into a combined scene graph representation, which can describe the overall context by estimating the similar concepts of their visual objects. For example, we human can find the common occurring objects of an image collection which are cow, sheep, hill, and street, and their relationships such as cow-on-hill and sheep-on-street. Based on the external knowledge, where both street and hill are places whereas sheeps and cows are animals. Then, based on the knowledge graph, we can understand that hill is a commonplace for animals. Therefore, we can assume the contexts as cow-on-hill, sheep-on-hill, and sheep-near-street. This example shows the advantage of using external knowledge in finding the relationships across multiple images.

To generate a summarized scene graph of an image collection, a naïve approach would be to do so in the inference phase [18], [19]. However, this is restricted to understanding

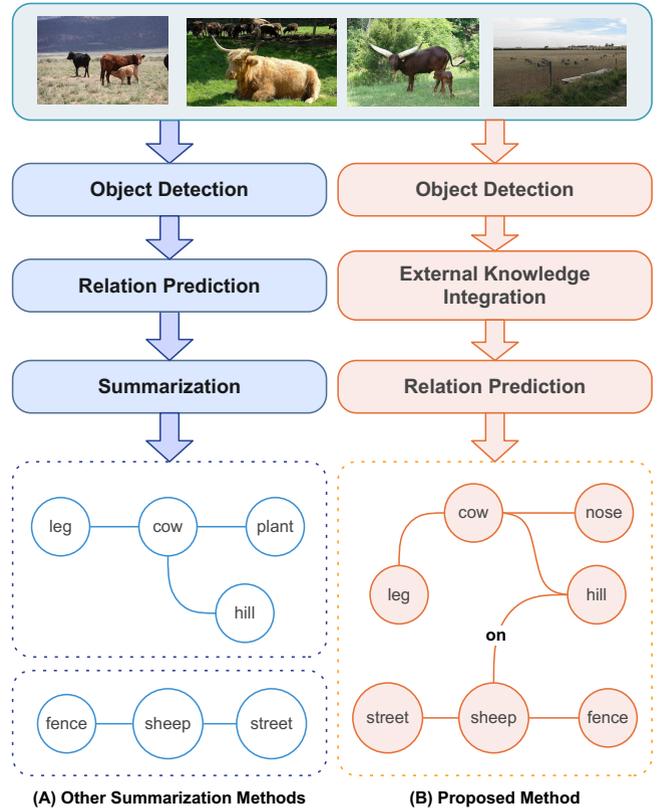


FIGURE 2. Comparison between (A) other summarization methods [2], [8], [9], [17]; Scene-graph generation with summarization process and (B) proposed method; End-to-end scene-graph summarization.

the relationships between object categories. Based on a similar motivation, we [8], [9] have previously proposed a scene-graph summarization method using graph theory for generating a caption of an image collection. Thus, we needed to provide a concept generalization process that aims to find the common concept words from an image collection to refine the final caption, which was performed by generalizing words. However, it would reduce details in the final caption, such as replacing cow or sheep with animal instead of describing both as summarized information such as cow and sheep. Therefore, the common challenge is to find relationships between different objects without reduction of details. For example, in the case of sheep on street and cow on hill, if we can utilize the external knowledge where, both street and hill are places for an animal, we can conclude both are in similar contexts such as places for living. Thus we can infer these indirect relationships such as sheep-on-hill. Based on this idea, the proposed method enhances the relation predictor of the scene-graph generation process so that it can generate generalized relationships of objects which can grasp the relationships of different objects in the same category across images. Inspired by the use of external knowledge to generate a scene graph for an unseen image [20], [21], we have decided to follow this idea.

In order to realize a scene graph summarization approach

of an image collection using external knowledge, there are three hurdles. First, we need to model external knowledge for the training process, for which we incorporate ConceptNet [22], a knowledge graph of commonsense semantic information. Second, we need to integrate the knowledge graph into the relation predictor of the scene-graph generation model. Lastly, we need to construct a summarized scene graph by combining all image information and then generate a final scene graph. Figure 2 shows a comparison between the generation of a summarized scene-graph with the summarization process in the inference phase of conventional methods and the proposed method. It demonstrates the case of finding a relationship between two sub-graphs by joining their location, *hill*.

Furthermore, it is also challenging to estimate the confidence score of each relationship to obtain a final scene graph, whereas a typical scene-graph generation method obtains the final scene graph based on confidence scores. To improve the estimation process, we employ PageRank [23] to re-calculate the node scores for selecting relationships in the process. However, a remaining challenge is the limitation of a dataset specific to the scene-graph summarization task. We hence construct a dataset for evaluating the proposed method based on the MS-COCO dataset [24] which is a popular image captioning dataset and widely used across various tasks including image retrieval and image summarization. In order to evaluate our work on a summarized scene graph, we introduce an evaluation process that evaluates the similarity score of a summarized scene graph based on the F-score whereas previous works focus only on precision. By this, the proposed evaluation process can account for false negatives.

Our contributions can be summarized as follows:

- We propose a scene-graph summarization method to generate a summarized scene graph of an image collection that has indirect relationships by inference using the external knowledge graphs into the relation prediction process.
- We introduce a sub-graph confidence score for estimating a summarized scene graph of an image collection.
- We introduce an evaluation process for evaluating a summarized scene graph by calculating the F-score which evaluates both false positives and false negatives of a generated scene graph.

II. RELATED WORK

In this section, we review related work on three topics; *Image Collection Summarization* which discusses work that aims to generate summarized information of an image collection, *Scene-Graph Generation* which discusses methods to generate image information in graph form, and *Knowledge Graph* which introduces the external knowledge that is used in the proposed method.

A. IMAGE COLLECTION SUMMARIZATION

Image collection summarization is the task of generating a representative summary of an image collection. Traditionally,

it aims to find a representative information in the form of image, textual, or scene-graph representation.

a: Image Representation

Summarizing an image collection is typically introduced in the photo album summarization task, which aims to find an image that represents an image album. Yu et al. [10] proposed a model composed of three hierarchically attentive Recurrent Neural Networks (RNNs) to encode album photos, select representative photos, and generate a story. Wang et al. [25] proposed a model with a hierarchical photo-scene encoder and reconstructor for generating an album story. Moreover, many work also find a representative image of an image collection using a clustering algorithm, such as Self-Organization Map (SOM) [26], [27] or k -Medoids [17] to cluster images and then represent some of them as an image representation of an image collection.

b: Textual Representation

Textual information is a popular summarization form for an image collection summarization task, which is represented as keywords, tags, phrases, or sentences. In summarizing an image collection into keywords or tags, Samani et al. [28] proposed a semantic summarization method for an image collection that utilizes the domain ontology as an input of the system by providing knowledge about the concept domain, e.g., Colosseum and Trevi Fountain. Zhang et al. [29] proposed a model to analyze an image collection and generate appropriate visual summaries and textual topics, e.g., sunset, sky, and sun. For summarizing an image collection into phrases, Trieu et al. [7] proposed a new task named multi-image summarization, that aims to generate a descriptive summary of an image collection such as styles of bags. They also introduced a new dataset for this task by collecting 2.1 million images from Web pages and then building collections of images, each consisting of at least five images. Li et al. [6] introduced a new task called context-aware captioning, which aims to describe an image collection in another context from different image collections. We [8], [9] introduced a method to generate a caption of an image collection based on a summarized scene graph based on graph theory [30].

c: Scene Graph Representation

As scene graphs are widely used for describing visual objects and their relationships for a single image [31], they are also used for describing multiple images. Pasini et al. [2] proposed an image-collection summarization method based on frequent subgraph mining and represents an image collection in a sub-graph form on the MS-COCO dataset [24]. Yang et al. [32] introduced a challenging task, named Panoptic Video Scene Graph Generation (PVSG), which aims to generate a summarized scene graph of real-world data and contributed a new panoptic video dataset for this task.

In the proposed method, we aim to describe an image collection by a scene graph, focusing on integrating external knowledge into the learning process.

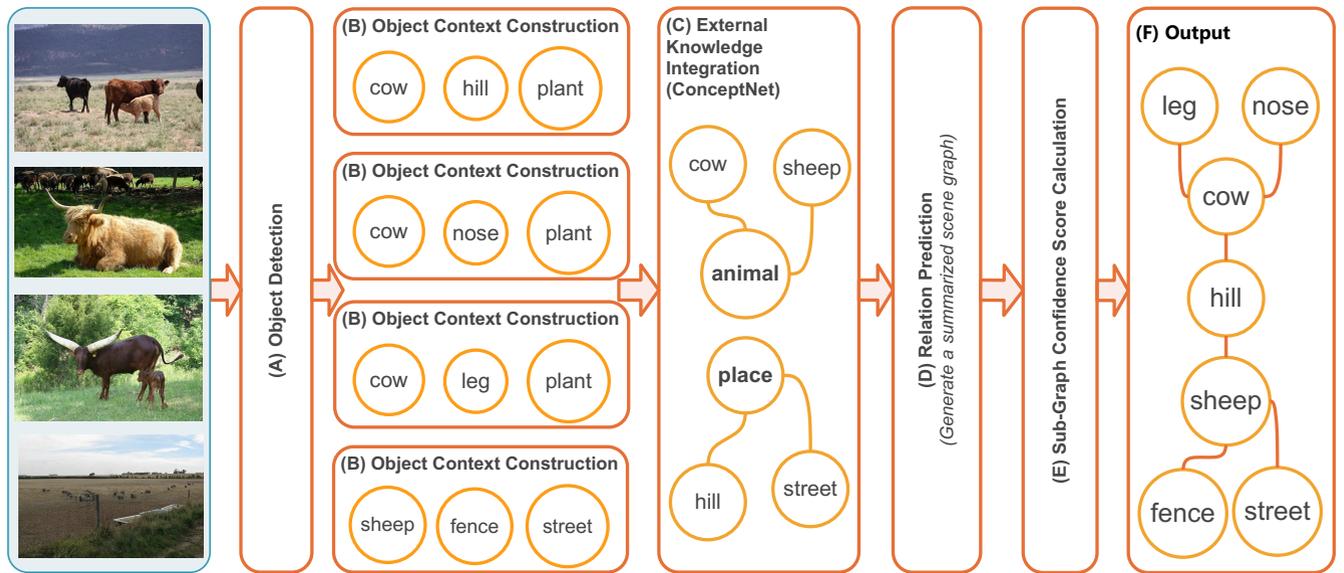


FIGURE 3. Overview of the proposed method consisting of five components: (A) *Object Detection* detects features from each image in an image collection, (B) *Object Context Construction* constructs the contextualized representation of the estimated context of each image, (C) *External Knowledge Integration* finds the knowledge graphs based on the object contexts and integrates the knowledge graphs and contextualized representations, (D) *Relation Prediction* predicts relationships between each combination of object contexts and contextualized representations, and (E) *Sub-Graph Confidence Score Calculation* calculates scores of all objects from the relation prediction and then generates a summarized scene graph as an (F) *Output*.

B. SCENE-GRAPH GENERATION

Scene-graph generation [31] is a popular technique in describing relationships between objects in an image. The relationships of objects are generally represented in triplets which consist of subject, predicate, and object. A common scene-graph generation architecture is divided into two main processes comprising object detection to detect the objects inside the image and relationship prediction to find the edges between the objects. In recent years, it has been widely introduced and implemented on the Visual Genome dataset [33] and the MS-COCO dataset [24]. In addition, scene-graph generation is also adapted to various applications, such as image captioning [34] and image retrieval [3], and has been shown to improve their results. Various techniques are introduced in scene-graph generation; Neural Motif [35] is built with Faster R-CNN [36] with plenty of backbones, such as ResNet-50 [37] and ResNext-101 [38], then computes and propagates through Bidirectional Long Short-Term Memory (BiLSTM) [39] for predicting relations. VCTree [40] is a scene-graph generation technique composed of dynamic tree structures which show the advantage of the use of a binary tree in finding co-occurrence and usual relationships between objects by allowing a dynamic structure. Iterative Message Passing (IMP) [41] is an end-to-end scene-graph model using standard RNNs and improves the prediction via message passing. From the long-tail problem of the scene-graph dataset, the most recent work [42] aims to introduce a technique to solve the bias of the dataset. Relation Transformer for Scene Graph Generation (RelTR) [43] is a one-stage end-to-end scene-graph generation technique that uses an attention mechanism and gives a fixed number of subjects, objects, and relationships to generate a scene graph.

In the proposed method, we use scene graphs as a means to model the relationships between images in an image collection.

C. KNOWLEDGE-GRAPH

A knowledge-base is widely used to enrich models, especially text-generation models [44]. ConceptNet [22] and Wikipedia dataset¹ are popular knowledge-bases that are used in the generation process. ConceptNet is a knowledge-graph that represents general knowledge and commonsense information, while the Wikipedia dataset is structured knowledge data with detailed information on each topic. In recent years, the knowledge-graph has become a popular knowledge-base on various generation processes, mainly focusing on capturing commonsense reasoning during the generation. To tackle the long-tail issues of scene-graph generation mentioned above, integrating knowledge-graphs to improve the generation is a widely introduced strategy, and results show its advantage. Moreover, a knowledge-graph is additionally implemented in an image retrieval task which aims to reason on the semantic context and generalize the concepts inside an image [29].

In the proposed method, we use ConceptNet which is a knowledge-graph to enhance the relation predictor for finding unseen relationships across images.

III. PROPOSED METHOD: SCENE-GRAPH SUMMARIZATION MODEL

From the idea of enhancing the relation predictor with external knowledge for predicting unseen relationships, we

¹<https://www.tensorflow.org/datasets/catalog/wikipedia/> (Accessed Sep 4, 2023)

build the proposed method by adapting an existing scene-graph generation method, Neural Motif [35]. The proposed method starts with extracting visual features from each image and then finds contextualized representations of each image following the Neural Motif approach. Next, we incorporate external knowledge into all contextualized representations. Lastly, we predict the relationship of each object in the contextualized representations and reconstruct them as a summarized scene graph as illustrated in Fig. 3.

The proposed method has five main components. The *Object Detection* component detects the visual features of images and modifies them for detecting objects in an image collection. The *Object Context Construction* component finds the contextualized representations of images. To generate a summarized scene graph from contextualized representations of an image collection, we introduce the *External Knowledge Integration* component to find the indirect relationships between detected objects and an encoder to encode them into the *Relation Prediction* component to generate a relationship between objects. Lastly, we introduce the *Sub-Graph Confidence Score Calculation* component that calculates the confidence scores of objects.

A. OBJECT DETECTION

The first part is the Object Detection component that detects a set of region proposals; Faster R-CNN [45] with ResNet-101 [37] is used as a detector backbone which shows good performance in scene-graph generation [42] compared with other backbones [31]. Following the scene-graph generation, from each image, a set of region proposals $B = \{b_1, \dots, b_n\}$ is predicted. Each region proposal b_i consists of a feature vector \mathbf{f}_i and an object label probability \mathbf{l}_i for the training phase.

In the inference phase, we modify an object detector to parse multiple images into the Relation Prediction component that generates a summarized scene graph of an image collection. Based on a single-image scene-graph generation model, we build an object detector backbone to detect image features \mathbf{f}_n , and proposals b_n . Then, we combine all image features and all proposals as:

$$F = \{\{\mathbf{f}_{n,1}, \dots, \mathbf{f}_{n,M}\}\}_{n=1, \dots, N}, \quad (1)$$

$$B = \{\{b_{n,1}, \dots, b_{n,M}\}\}_{n=1, \dots, N}, \quad (2)$$

where F is a set of feature vectors of all images, N is the number of images, M is the number of region proposals of each image, and B is a set of proposals of all images. All predicted sets of region proposals and each region proposal consists of a feature vector \mathbf{f} and an object label probability \mathbf{l} , which is used in the Object Context Construction component.

B. OBJECT CONTEXT CONSTRUCTION

The second component is the Object Context Construction component that constructs a contextualized representation of a set of region proposals by concatenating them into a linear sequence which is sorted by detected locations, $[(b_1, \mathbf{f}_1, \mathbf{l}_1), \dots, (b_N, \mathbf{f}_N, \mathbf{l}_N)]$. Then, a bidirectional LSTM [39] is used as:

$$C = \text{biLSTM}(\{\mathbf{f}_n; W_1 \mathbf{l}_n\}_{n=1, \dots, N}), \quad (3)$$

where C is a set of object contexts, in which each object context contains the hidden state of each element in the linearization of B , W_1 is a parameter that maps to the distribution prediction represented in the matrix, and \mathbf{l}_n is a probability vector of object labels. Each object context is used to decode a class label with an LSTM as:

$$\mathbf{h}_n = \text{LSTM}_n([\mathbf{c}_n; \hat{\mathbf{o}}_{n-1}]), \quad (4)$$

$$\hat{\mathbf{o}}_n = \text{onehot}(\arg\max(W_o \mathbf{h}_n)) \in \mathbb{R}^{|\mathcal{C}|}, \quad (5)$$

where \mathbf{c}_n is an object context vector in a set of object contexts, C , \mathbf{h}_n is a hidden state that is used in the relation predictor, and $\text{onehot}(\cdot)$ embeds a scalar value into a one-hot vector. W_o is a parameter that maps to the hidden state.

C. EXTERNAL KNOWLEDGE INTEGRATION

Based on the idea of integrating external knowledge to enhance the relation predictor, there are two stages. First, we need to build the external knowledge from ConceptNet [22] for the proposed method. Then, we build the encoding layer to encode the external knowledge for incorporating it into the relation predictor, whose knowledge-graphs are built based on the class labels of a set of object contexts, C .

1) Knowledge-Graph Construction

The objective in constructing the knowledge-graph for this task is to build a word-embedding knowledge-graph from ConceptNet. Since ConceptNet provides various aspects of relation information, we build a knowledge-graph focusing on semantic relations consisting of “*relatedTo*”, “*similarTo*”, and “*synonym*” to improve the relation prediction of similar objects. In the building process, we first initialize the word collection to retrieve the semantic relations from VG200 [46] which consists of 150 labels by giving a class pair, (x, y) , and then employ the connection between (x, y) as (V_x, V_y) . Next, we gather all possible semantic paths from V_x to V_y as $P_{(x,y)}$. Lastly, we employ GloVe word embedding [47] to encode all words.

2) Knowledge-Graph Integration

For the External Knowledge Graph Integration component, we first build a Graph Convolutional Network (GCN) using the GlobalSortPool operator [48], which enables learning from nodes on graph topology instead of summing them up, as an encoder for a knowledge-graph. Then, all knowledge-graphs of class pair (x, y) in vector form are encoded into knowledge feature vectors as:

$$\mathbf{e}_{\text{kb}}^{(x,y)} = \text{GlobalSortPool}(\mathbf{N}^{(x,y)}), \quad (6)$$

where $\mathbf{N}^{(x,y)}$ represents all embedding nodes from $P_{(x,y)}$.

In the training and evaluating processes, we first retrieve all predicted class pairs from the object context as (x, y) . Next,

we retrieve all possible connection paths from the knowledge-graph, $P_{(x,y)}$. Lastly, all of them are encoded into $e_{kb}^{(x,y)}$ and then concatenated into each contextualized representation to estimate relationships as discussed in the *Relationship Prediction* component.

D. RELATION PREDICTION

The obtained object contexts by the previous process are used in the Relation Prediction component, in which a set of regions, B , and objects are encoded by a bidirectional LSTM as:

$$E = \text{biLSTM}([c_n; W_2 \hat{c}_n]_{n=1, \dots, N}), \quad (7)$$

where E is a set of edge contexts, in which each edge context contains the states of the bounding-box regions and W_2 is a mapping parameter of \hat{c}_n . Each edge context is combined with the knowledge embedding and predicts the relation of each pair as:

$$g_{i,j} = (W_h e_i)(W_t e_j) f_{i,j}, \quad (8)$$

$$r_{i,j} = \text{argmax}([g_{i,j}; e_{kb}^{(i,j)}] W_r), \quad (9)$$

where e_i and e_j are edge context vectors of head and tail, W_h and W_t are parameters of heads and tails, $f_{i,j}$ is a feature vector for the union of two bounding boxes, W_r is a parameter that maps to the relation predictor, e_{kb} is a knowledge embedding vector, and $r_{i,j}$ is a relation vector which is transformed into the relation and probability score by using softmax as an activation function.

E. SUB-GRAPH CONFIDENCE SCORE CALCULATION

From the implementation of multiple images to generate a summarized scene graph which aims to generate all possible relationships across images, we also need to re-estimate the relationship scores in a generated scene graph instead of using only confidence scores. The estimation aims to estimate triplet scores which are calculated from subject, predicate, and object confidence by analogy of PageRank [23].

To calculate a score, we first find summarized scores of each object using its confidence score as object scores as:

$$\text{obj_score}_i = \sum_{j=0}^N \text{obj_confidence}_{i,j}, \quad (10)$$

where N is the count of the object. From each object score, we find the mean object score from all object scores mean_{obj} as:

$$\text{mean}_{\text{obj}} = \frac{1}{M} \sum_{i=0}^M \text{obj_score}_i, \quad (11)$$

where M is the number of the unique object. The mean object score is used for filtering out the object scores that are lower than the mean score.

Lastly, we collect the object pairs whose relation scores are greater than the mean score and employ PageRank to

Algorithm 1: Sub-Graph Score

Input: $\text{objs}_{\text{label}}, \text{objs}_{\text{score}}, \text{objs}_{\text{box}}, \text{pairs}_{\text{subj,obj}}, \text{rels}_{\text{label}}, \text{rels}_{\text{score}}$

Output: A summarized scene graph

Result: $\text{pagerank}_{\text{score}}$

$\text{objects} = \{\};$
 $\text{relations} = \{\};$

foreach
 $\text{obj, score, box} \in (\text{objs}_{\text{label}}, \text{objs}_{\text{score}}, \text{objs}_{\text{box}})$ **do**
 if $\text{obj} \in \text{objs}$ **then**
 $\text{objects}[\text{obj}] = (\text{objects}[\text{obj}][0] + \text{score}, \text{objects}[\text{obj}][1]);$
 else
 $\text{objects}[\text{obj}] = (\text{score}, \text{objects}[\text{obj}][1]);$
 end
end

$\text{mean}_{\text{obj}} = \text{mean}(\text{objs}_{\text{score}})$

foreach $\text{obj} \in \text{objects}$ **do**
 if $\text{obj.score} < \text{mean}_{\text{obj}}$ **then**
 $\text{objects.remove}(\text{obj})$
 end
end

$\text{triplets} = []$
 $\text{scores} = []$

foreach $\text{subj, obj, pred, score} \in (\text{pairs}_{\text{subj,obj}}, \text{rels}_{\text{label}}, \text{rels}_{\text{score}})$ **do**
 if $\text{sub, obj} \in \text{objects}$ **then**
 $\text{triplets.add}(\langle \text{subj}, \text{obj}, \text{pred} \rangle)$
 $\text{scores.add}(\text{score})$
 else
 $\text{triplets.add}(\langle \text{subj}, \text{obj}, \text{pred} \rangle)$
 $\text{scores.add}(\text{score})$
 end
end

$\text{pagerank}_{\text{score}} = \text{PageRank}(\text{triplets}, \text{scores})$

calculate the confidence score of each object whose process is detailed in Algorithm 1.

IV. EVALUATION PROCESS

Due to the lack of ground truth for this task, we use common metrics that are used in image collection scene-graph summarization tasks [2], [19]; similarity [16], [49], [50], coverage [28], [51], and diversity [52], [53] of a generated scene graph to the ground-truth scene graph of each image. However, most evaluation techniques focus on estimating the generating precision, in which the evaluation score tends to increase based on the quantity of the generated results. As such, we introduce an evaluation process which focuses on evaluating the quality of a summarized scene graph using F-score based on estimating the similarity between scene graphs. Since the estimation of the similarity between scene graphs has been attempted with various approaches, the technique of using word embedding shows a better qualitative estimation in scene-graph generation [50].

Given a ground-truth scene graph $G = \{t_1, \dots, t_n\}$ con-

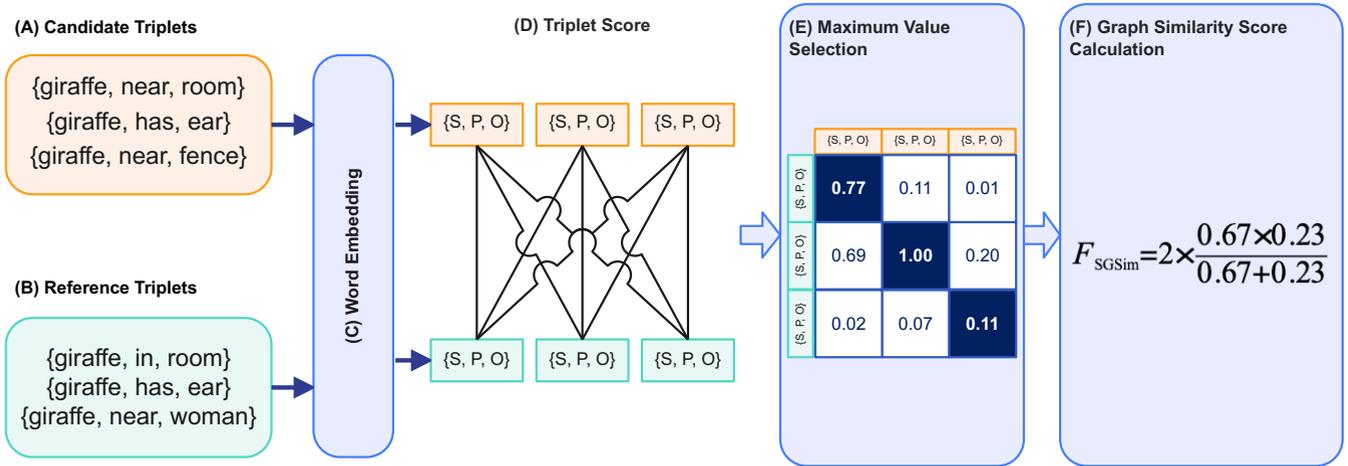


FIGURE 4. Overview of the evaluation process consisting of three components: From (A) *Candidate Triples* and (B) *Reference Triples*, (C) *Word Embedding* encodes both of them into a vector form, (D) *Triplet Score* calculates all triplet similarities between candidates and references, (E) *Maximum Value Selection* finds the maximum value of the similarity scores of each triplet pair, and (F) *Graph Similarity Score Calculation* calculates the final score.

sisting of ground-truth triplets, a generated scene graph $\hat{G} = \{\hat{t}_1, \dots, \hat{t}_m\}$ consisting of generated triplets, and each triplet in a scene graph denoted as $t = \langle s, p, o \rangle$, where s is subject, p is predicate, and o is object, we first employ GloVe [47] word embedding to transform all words in each triplet into token representation in a vector form. Then, we compute the similarity score of each triplet of a generated scene graph and each triplet of a ground-truth scene graph. Figure 4 illustrates the evaluation process.

The calculation process is adapted from BERTScore [54] to the evaluation process. In the BERTScore calculation, first, they implement Bidirectional Encoder Representations from Transformers (BERT) [55] embedding to tokenize all words of candidate and reference sentences into a vector form. Next, it calculates the similarity score between all words and then selects the maximum score of each word based on greedy matching. Lastly, it calculates the precision score, recall score, and F-score as evaluation metrics of a candidate sentence. From this process, we can also evaluate the false negative of a candidate scene graph, whereas other evaluation techniques mainly focus only on precision. Thus, in the proposed evaluation process, from all candidate triplets and reference triplets, we first encode all triplets into vector forms. Next, we calculate the similarity score between each triplet of all reference triplets and all candidate triplets. Then, we select the maximum similarity score of each candidate triplet calculation. Lastly, we calculate the precision score, recall score, and consecutively, F-score as a scene graph similarity score. Details of each process are described below.

A. TRIPLET SCORE

Given a generated triplet in a vector representation \hat{t} and a ground-truth triplet in a vector representation t , each triplet comprises tokens of a subject, a predicate, and an object. To calculate the similarity between token representations, we estimate the similarity between each ground-truth subject and

object and the generated subject and object by calculating the similarity S as follows:

$$S(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|}, \quad (12)$$

where \mathbf{a} and \mathbf{b} are the corresponding embeddings of each pair of subject or object, $\mathbf{a} \cdot \mathbf{b}$ is the dot product between vectors \mathbf{a} and \mathbf{b} , and $\|\mathbf{a}\|$ and $\|\mathbf{b}\|$ are the L2 norms of vectors \mathbf{a} and \mathbf{b} , respectively.

As the similarity between subjects or objects is calculated based on word similarity, the similarity between predicates is specifically estimated in the definition of entity-based similarity. The calculation of the scene-graph similarity focuses on the relationship between objects, which can reduce redundant information [49]. We employ the calculation of the similarity between predicate $S_{\text{pred}}(\mathbf{p}_i, \hat{\mathbf{p}}_j)$ as:

$$S_{\text{pred}}(\mathbf{p}, \hat{\mathbf{p}}) = \begin{cases} 1 & \mathbf{p} = \hat{\mathbf{p}}; \\ 0 & \mathbf{p} \neq \hat{\mathbf{p}}. \end{cases} \quad (13)$$

In the following, given all similarity scores of triplet t , consisting of subject similarity score S_{sub} , predicate similarity score S_{pred} , and object similarity score S_{obj} , we compute them into a single value by calculating the mean score M_{sim} as:

$$M_{\text{sim}}(t_i, \hat{t}_j) = \text{mean}(\{S_{\text{sub}}(\mathbf{s}_i, \hat{\mathbf{s}}_j), S_{\text{pred}}(\mathbf{p}_i, \hat{\mathbf{p}}_j), S_{\text{obj}}(\mathbf{o}_i, \hat{\mathbf{o}}_j)\}). \quad (14)$$

B. GRAPH SIMILARITY SCORE

To estimate the final score between scene graphs, we use the maximize function to find the maximum matching score, where each candidate triplet t is matched to a ground-truth triplet \hat{t} as:

$$R_{SGSim} = \frac{1}{|G|} \sum_{t_i \in G} \max_{\hat{t}_j \in \hat{T}} (M_{\text{sim}}(t_i, \hat{t}_j)). \quad (15)$$

Algorithm 2: Graph Similarity

Input: G, \hat{G}
Output: F-score of reference and candidate graphs
Result: F_{SGSim}
 $F_{SGSim} = 0$
 $scores_{max} = []$
foreach $\langle s, p, o \rangle \in G$ **do**
 $scores = []$
 foreach $\langle \hat{s}, \hat{p}, \hat{o} \rangle \in \hat{G}$ **do**
 $S_{sub}(s, \hat{s}) = \frac{s \cdot \hat{s}}{\|s\| \cdot \|\hat{s}\|}$;
 $S_{pred}(p, \hat{p}) = \begin{cases} 1 & \text{if } p = \hat{p}; \\ 0 & \text{otherwise } p \neq \hat{p}; \end{cases}$
 $S_{obj}(o, \hat{o}) = \frac{o \cdot \hat{o}}{\|o\| \cdot \|\hat{o}\|}$;
 $score = \text{mean}(\{S_{sub}(s, \hat{s}), S_{pred}(p, \hat{p}), S_{obj}(o, \hat{o})\})$;
 $scores.insert(score)$;
 end
 $scores_{max}.insert(\max(scores))$
end
 $R_{SGSim} = \frac{1}{|G|} \text{sum}(scores_{max})$
 $P_{SGSim} = \frac{1}{|\hat{G}|} \text{sum}(scores_{max})$
 $F_{SGSim} = 2 \frac{P_{SGSim} R_{SGSim}}{P_{SGSim} + R_{SGSim}}$

Then, we calculate the ratio of the sum of the maximum similarity scores to the norm of a generated graph as:

$$P_{SGSim} = \frac{1}{|\hat{G}|} \sum_{t_i \in \hat{T}} \max_{t_j \in \hat{T}} (M_{sim}(t_i, t_j)). \quad (16)$$

Lastly, the mean of R_{SGSim} and P_{SGSim} is calculated as:

$$F_{SGSim} = 2 \frac{P_{SGSim} R_{SGSim}}{P_{SGSim} + R_{SGSim}}. \quad (17)$$

We demonstrate in Algorithm 2 the calculation process for all triplets of a summarized scene graph and a ground-truth scene graph.

V. EXPERIMENTALS

A. DATASET

Due to the lack of image summarization datasets, we adapt two datasets, including an image captioning dataset, MS-COCO [24], and a visual scene-graph dataset, Visual Genome [33], for the experiments. For the training process and the preliminary evaluation, we use the VG200 dataset [46], which is based on the Visual Genome dataset consisting of 50 relationships and is balanced in category frequency. It contains 101,174 images from the MS-COCO dataset. To experiment on a scene-graph image-collection summarization task, we build a testing set of an image collection with annotation by grouping images in the MS-COCO testing dataset using VSE++ [56], which is an image retrieval task by estimating the similarity of image contexts and image captions. Following the Karpathy split [57] on the MS-COCO

dataset, the initial testing set was selected from 5,000 images of the MS-COCO testing set. Then, we retrieved 5 images, with each image annotated with 5 captions, to build a collection, making our testing set contain 5,000 collections with 6 images each. Lastly, we build the ground truth of each image collection for the evaluation process in a scene-graph form.

As image summarization aims to generate summarized information that can describe the overall contexts of an image collection and the limitation of the ground truth in the proposed method, we use Neural Motif [35] pre-trained on the VG200 dataset and evaluated on Scene-Graph Detection Recall (SGDet R@100), to generate a scene graph of each image in a collection for evaluation. Then, we consider them as the ground truth of each collection for evaluating the proposed method which makes each collection consisting of 6 ground-truth scene graphs.

B. TRAINING STRATEGY

With the lack of ground truth in scene-graph summarization datasets, we first train and evaluate the scene-graph generation on a single-image from the VG200 dataset. In the training phase, we train the model following the VG200 dataset where the number of labels and predicates are 150 and 50, respectively. The learning rate is initiated to 0.12. We use Adam [58] for optimization and cross-entropy loss as the loss function. To pre-evaluate the model for multiple-image scene-graph summarization, we observed a SGDet recall to select the best checkpoint for the proposed method.

C. EVALUATION

As the proposed method is modified from a single scene-graph generation approach, we consider evaluating the proposed method in two aspects. First, *Multiple-Images Scene-Graph Summarization* evaluates the proposed method for an image-collection scene-graph summarization. Second, *Single-Image Scene-Graph Generation* evaluates the proposed method to confirm that it is still sustainable for a single-image scene-graph generation. Lastly, we benchmark the evaluation process in *Benchmark for the Evaluation Process* to show the accountability for scene-graph generation.

1) Multiple-Images Scene-Graph Summarization

For multiple-image scene-graph summarization, we evaluate the proposed method for image-collection scene-graph summarization on the MS-COCO dataset. Due to the lack of ground truth, we follow the common practice in the evaluation of scene graph generation in three perspectives; “Coverage” [28], [51], “Diversity” [52], [53], and “Similarity” [49], [50]. For the Coverage evaluation, we follow the graph theory to estimate the coverage of a generated scene graph to ground-truth scene graphs. For the Diversity evaluation, we implement two evaluation processes comprising graph diversity and Graph Edit Distance (GED) [59]. For the Similarity evaluation, we adopt a simple contrastive learning framework for connecting scene-graphs and images (GICON) [60] which is the evaluation technique by learning the

similarity between an image and a scene graph with bounding boxes or without bounding boxes. Since the proposed method focuses on image collection summarization, we evaluate the proposed method only without bounding boxes. Lastly, we employ an evaluation process that evaluates the similarity of a summarized scene graph to the ground truth by SGSim proposed in Section IV.

2) Single-Image Scene-Graph Generation

For single-image scene-graph generation, we evaluate the performance on VG200 compared with the baseline to ensure that the proposed method still sustains good results. We evaluate three scene graph evaluation metrics; *Scene Graph Classification Recall* (SGCls Recall), which measures subjects, objects, and predicates using ground-truth bounding boxes, *Predicate Classification Recall* (PredCls Recall), which is the relationships prediction using ground-truth bounding boxes, subjects, and objects, and *Scene Graph Detection Recall* (SGDet Recall), which is the prediction of subjects, objects, and predicates without using the ground truth.

3) Benchmark for the Evaluation Process

Here, we discuss the evaluation metric to ablate the evaluation process. As it is proposed for evaluating scene-graph generation, we benchmark it on single-image scene-graph generation with the VG200 dataset by comparing it with other scene-graph generation baselines. As we aim to evaluate based on the false negative generation, we assess it with two evaluation metrics. First, *Scene Graph Detection Recall* (SGDet Recall) is a popular scene graph evaluation metric. Next, GICON is an evaluation metric from learning the similarity between a generated scene graph and an image with bounding boxes or without bounding boxes.

D. BASELINES

As discussed in the previous section, the evaluation is divided into three tasks; multiple-image scene-graph summarization, single-image scene-graph generation, and the ablation study on the evaluation process. In this section, we introduce baseline methods corresponding to each of them.

1) Baseline for Multiple-Images Scene-Graph Summarization

To evaluate the proposed method in the multiple-images summarization setting, we choose three baseline methods; Semantic Image Summarization (SIImS) [2], Image Collection Captioning (ICC) [8], [9], and k -Medoids [17]. SIImS is a scene graph summarization method on the MS-COCO dataset by finding frequent sub-graphs. ICC is a scene graph summarization method precisely proposed by us for generating a caption based on graph theory. k -Medoids is a clustering method in which the implementation of the summarization is the same as the SIImS [2]. All of these baselines are evaluated on the testing set of the MS-COCO dataset which consists of 6 images per image collection.

2) Baseline for Single-Image Scene-Graph Generation

To evaluate the proposed method for single-image scene graph generation, we choose four baseline methods; Iterative Message Passing (IMP) [41] which uses the standard Recurrent Neural Network (RNN) via message-passing process, Neural Motif [35] which is implemented based on Stacked Motifs architecture, Transformer [42] which is based on causal inference, and VCTree [40] which takes advantage of the structured object representations. All of the baseline models are trained on the VG200 dataset. Then, we observe the best checkpoint on SGCls Recall, PredCls Recall, and SGDet Recall for evaluating the proposed method on the VG200 dataset.

3) Baseline for Ablation Study on the Evaluation Process

For the ablation study on the evaluation method, we aim to benchmark the evaluation process compared to other evaluation metrics. We choose four state-of-the-art scene-graph generation methods on the Visual Genome dataset; Neural Motif, Transformer, ReIDN [61], and ReITR [43].

E. RESULTS

We report the results of the proposed method in the three evaluation tasks; multiple-image scene-graph summarization, single-image scene-graph generation, and the ablation study.

1) Multiple-Images Scene-Graph Summarization

In this section, we discuss the results of the proposed method for an image collection summarization task. For comparison, we select top-10 scores in three aspects; Coverage, Diversity, and Similarity. The results are shown in Table 1.

a: Coverage

For Coverage evaluation, the coverage of objects and subjects (nodes), and predicates (edges) are evaluated based on graph theory [30]. The results in Table 1 show that k -Medoids achieves the best score in generating a summarized scene graph, whereas the proposed method achieves the second.

b: Diversity

For Diversity evaluation, we use two evaluation metrics which consist of Diversity and GED. Diversity evaluation refers to the similarity distance between a summarized scene graph and a ground-truth scene graph. The results in Table 1 show that the proposed method achieves the best scores in both Diversity and GED, whereas SIImS achieves the second for Diversity and k -Medoids achieves the second in GED.

c: Similarity

For Similarity evaluation, Table 1 shows that the proposed method achieves the best score compared to the other methods on GICON which estimates the Similarity between scene graph and images, and the proposed evaluation process, SGSim which is evaluated based on scene-graph contents. Meanwhile, k -Medoids achieves the second in both GICON and SGSim.

TABLE 1. Evaluation of an image collection summarization compared with SimS [2], *k*-Medoids [17], ICC [9], and the proposed method by estimating Coverage [30], Diversity [30], Graph Edit Distance (GED) [59], GICON [60], and the proposed evaluation process (SGSim). Results in bold indicate the highest scores whereas those underlined indicate the second highest scores.

Models	Coverage Evaluation	Diversity Evaluation		Similarity Evaluation	
	Coverage [30] ↑	Density [30] ↓	GED [59] ↓	GICON [60] (Location-Free) ↑	Proposed (SGSim) ↑
SimS [2]	18.1	<u>33.3</u>	6.6	10.7	11.4
<i>k</i> -Medoids [17]	42.3	41.7	<u>3.9</u>	<u>20.8</u>	<u>14.3</u>
ICC [9]	18.1	66.9	17.9	16.7	10.4
Proposed	<u>22.2</u>	27.1	2.6	33.3	14.8

TABLE 2. Evaluation of single-image scene-graph generation on an image from the VG200 dataset compared with baseline methods; IMP [41], Neural Motif [35], Transformer [42], VCTree [40], ReIDN [61] and the proposed method by observing recall scores of Scene Graph Detection (SGDet), Scene Graph Classification (SGCls), and Predicate Classification (PredCls). Results in bold indicate the highest scores whereas those underlined indicate the second highest scores.

Models	SGDet			SGCls			PredCls		
	R@20 ↑	R@50 ↑	R@100 ↑	R@20 ↑	R@50 ↑	R@100 ↑	R@20 ↑	R@50 ↑	R@100 ↑
IMP [41]	18.1	25.9	31.2	34.0	37.5	38.5	54.3	61.1	63.1
Neural Motif [35]	<u>25.5</u>	<u>32.8</u>	37.2	35.6	38.9	39.8	58.5	65.2	67.0
Transformer [42]	25.6	33.0	<u>37.4</u>	<u>36.9</u>	<u>40.2</u>	<u>41.0</u>	59.1	65.6	67.3
VCTree [40]	24.5	31.9	36.2	42.8	46.7	47.6	<u>59.0</u>	<u>65.4</u>	<u>67.2</u>
ReIDN [61]	24.0	32.4	37.8	31.9	35.7	36.6	54.0	60.9	61.8
Proposed	22.8	28.6	36.6	35.9	38.6	39.4	55.4	64.7	66.7

d: Qualitative Results

Qualitative results are shown in Fig. 5. It demonstrates that the proposed method performs good in finding the relationship and estimating the commonly occurring information. For example, the first example shows how the proposed method can find all common object information, such as *sheep* and *cow*, and further estimate the commonsense relationship between them based on the location, *hill*. In contrast, SimS and *k*-Medoids generate a summarized scene graph based on the most frequent object, *cow*, neglecting the other common object, *sheep*. ICC can generate information of *sheep* but cannot infer common relationships between *sheep* and *cow*. The second example shows how the proposed method can handle the overall context location of an image collection, while SimS, *k*-Medoids, and ICC lose some of the overall location information in their results. The third example shows the performance in finding summarized information of *bus*, and their common environmental characteristics *street*, *building*, and *people* are connected. In contrast, SimS, *k*-Medoids, and ICC fail to include the object *people*.

From the overall evaluation scores, the proposed method achieves better scores in Diversity and Similarity perspectives, whereas *k*-Medoids achieves the best score in Coverage. Most *k*-Medoids scores achieved the second place except for Diversity where SimS [2] achieved the second place. Meanwhile, the qualitative results showed the performance of finding common context being beneficial for, e.g., summarization tasks such as photo album summarization. However,

the limitation of the proposed method in generating relationships between images is their reliance on the commonsense knowledge-graph. As the inference relationships are reasoned from external knowledge and not grounded in visual information, the generated result might be unrelated to the actual image collection.

2) Single-Image Scene-Graph Generation

The single-image scene-graph generation results are shown in Table 2. Since the objective of the proposed method mainly observes Scene Graph Detection (SGDet), we focus on its result when assessing a single-image scene-graph generation task. The result of SGDet R@100 shows that Neural Motifs [35] and Transformer [42] achieve better results compared with the proposed method while the proposed method achieves better results compared with IMP [41] and VCTree [40]. In contrast, the results of R@20 and R@50 show that the proposed method achieves better results only compared with IMP [41]. As the proposed method aims to enhance the relation prediction toward unseen relationships, it is not restricted to the ground truth in a single scene-graph generation as shown in the result. As the proposed method shows better scores compared to IMP in SGDet evaluation, ReIDN in SGCls, and IMP and ReIDN in PredCls, it is still sustainable for a single-image scene-graph generation even if it cannot overcome other scene-graph generation baselines.

However, since the proposed method targets multiple-images summarization, this out-of-task evaluation was purely



FIGURE 5. Comparison of the proposed method with baseline methods in three examples. (A) *Proposed* shows a summarized scene graph generated by the proposed method. (B) *SimS* [2] demonstrates that using Semantic Image Collection Summarization. (C) *k-Medoids* [17] demonstrates that generated by the clustering technique. (D) *ICC* demonstrates that generated by our previous work [8], [9] using graph theory.

TABLE 3. Benchmark of the evaluation methodology compared with Scene Graph Detection (SGDet) with R@20, R@50, and R@100 and GICON [60] for both location free and with bounding box. SGSim with k is the number of triplets that is used in calculating the similarity score. Results in bold indicate the highest scores whereas those underlined indicate the second highest scores.

Models	SGDet			GICON [60]		SGSim		
	R@20 ↑	R@50 ↑	R@100 ↑	Location-Free ↑	W/ Bounding box ↑	$k=10$ ↑	$k=30$ ↑	$k=50$ ↑
Neural Motif [35]	<u>25.5</u>	27.2	37.2	84.5	90.2	9.5	18.9	25.2
Transformer [42]	25.6	33.0	<u>37.4</u>	92.9	<u>96.5</u>	8.1	11.5	15.1
RelDN [61]	24.0	<u>32.4</u>	37.8	93.7	89.9	8.1	<u>16.6</u>	<u>22.5</u>
RelTR [43]	21.2	27.8	33.7	<u>93.1</u>	97.0	<u>9.2</u>	11.9	11.9

performed to understand the limitations of this approach.

3) Ablation Study on the Evaluation Methods on Scene-Graph Generation Models

We benchmark our evaluation methodology with existing methods that evaluate graph-oriented (graph structure) and graph similarity-oriented (similarity vertices and edges) compared to other evaluation methods. In the benchmark process, we construct a benchmark for single scene graph generation for the VG200 dataset and perform analysis on four models; Neural Motif, Transformer, VCTree, and RelTR. For graph-oriented evaluation, we use Scene Graph Detection Recall (R@20, R@50, R@100) as a metric. For the graph similarity-oriented evaluation, we use GICON which is a learnable graph similarity metric for evaluating with bounding boxes (W/ Bounding Box) and without bounding boxes (Location Free). For each evaluation, we find the top- k triplets that are used in the process in which k are 10, 30, and 50 triplets. In the triplet selection, we observe the relationship scores to find the top- k triplets for the benchmark.

The benchmark results in Table 3 report the results based on the number of retrieved triplets with confidence scores which shows the relevance to the rise of the scores. However, the high number of triplets does not always increase the similarity in the evaluation process, as shown in Transformer and RelTR. As RelTR is provided for inferring a fixed-size set of triplets, even if we increase the number from 30 to 50 triplets, the accuracy is still not significantly improved. Meanwhile, Transformer shows little improvement when increasing the retrieving number of triplets, and the other methods show significant improvement when increasing the retrieving number of triplets. Consequently, the other evaluation metrics, GICON and SGDet, focus on evaluating precision and recall, so the high retrieving number of triplets tends to result in high scores.

VI. CONCLUSION

We introduced a scene-graph summarization method following the idea that aims to enhance the relation predictor in the training process for an image collection incorporating external knowledge. The results show that the proposed method can generate a summarized scene-graph that is good in diversity and similarity perspectives compared with other baseline

methods while it still lacks accuracy in terms of the coverage information. Additionally, the experimental results showed the advantage of using external knowledge in grasping the overall context of an image collection for finding the common relationships across images which is beneficial for a summarization task, especially, photo album summarization. However, the limitation is the lack of actual ground truth in the evaluation process. In the future, we plan to build a more suitable dataset for an image-collection scene-graph summarization task.

ACKNOWLEDGMENT

The computation was carried out using the General Projects on the supercomputer ‘‘Flow’’ at Information Technology Center, Nagoya University.

REFERENCES

- [1] Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alex Hauptmann. A comprehensive survey of scene graphs: Generation and application. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):1–26, Dec. 2021.
- [2] Andrea Pasini, Flavio Giobergia, Eliana Pastor, and Elena Baralis. Semantic image collection summarization with frequent subgraph mining. *IEEE Access*, 10:131747–131764, Dec. 2022.
- [3] Brigit Schroeder and Subarna Tripathi. Structured query-based image retrieval using scene graphs. In *Proc. CVPR2020*, pages 680–684, [Online], Jun. 2020.
- [4] Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. Webly supervised concept expansion for general purpose vision models. In *Proc. ECCV2022*, volume 36, pages 662–681, Tel Aviv, Israel, Oct. 2022.
- [5] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graphs. In *Proc. AAAI2021*, pages 3208–3216, [Online], Feb. 2021.
- [6] Zhuowan Li, Quan Tran, Long Mai, Zhe Lin, and Alan L. Yuille. Context-aware group captioning via self-attention and contrastive features. In *Proc. CVPR2020*, pages 3440–3450, [Online], Jun. 2020.
- [7] Nicholas Trieu, Sebastian Goodman, Pradyumna Narayana, Kazoo Sone, and Radu Soricut. Multi-image summarization: Textual summary from a set of cohesive images. *Comput. Res. Reposit. arXiv Preprint*, arXiv:2006.08686, [Online], Jun. 2020.
- [8] Itthisak Phueaksri, Marc A. Kastner, Yasutomo Kawanishi, Takahiro Komamizu, and Ichiro Ide. Towards captioning an image collection from a combined scene graph representation approach. In *Proc. MMM2023*, volume 1, pages 178–190, Bergen, Norway, Mar. 2023.
- [9] Itthisak Phueaksri, Marc A. Kastner, Yasutomo Kawanishi, Takahiro Komamizu, and Ichiro Ide. An approach to generate a caption for an image collection using scene graph generation. *IEEE Access*, Nov. 2023.
- [10] Licheng Yu, Mohit Bansal, and Tamara L. Berg. Hierarchically-attentive RNN for album summarization and storytelling. In *Proc. EMNLP2017*, pages 966–971, Copenhagen, Denmark, Sep. 2017.

- [11] Ruize Wang, Zhongyu Wei, Piji Li, Qi Zhang, and Xuanjing Huang. Storytelling from an image stream using scene graphs. In *Proc. AAAI2020*, pages 9185–9192, New York, NY, USA, Feb. 2020.
- [12] Chong-Wah Ngo, Yu-Fei Ma, and Hong-Jiang Zhang. Video summarization and scene detection by graph modeling. *IEEE Trans. Circuits Syst. Video Technol.*, 15(2):296–305, 2005.
- [13] Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. Comprehensive image captioning via scene graph decomposition. In *Proc. ECCV2020*, volume 14, pages 211–229, [Online], Aug. 2020.
- [14] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. CVPR2018*, pages 6077–6086, Salt Lake City, UT, USA, Jun. 2018.
- [15] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proc. CVPR2015*, pages 3668–3678, Boston, MA, USA, Jun. 2015.
- [16] Sangwoong Yoon, Woo Young Kang, Sungwook Jeon, SeongEun Lee, Changjin Han, Jonghun Park, and Eun-Sol Kim. Image-to-image retrieval by learning similarity between scene graphs. In *Proc. AAAI2021*, pages 10718–10726, [Online], Feb. 2021.
- [17] Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for k-medoids clustering. *Expert Syst. Appl.*, 36(2):3336–3341, Mar. 2009.
- [18] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. In *Proc. ECCV2020*, volume 23, pages 606–623, [Online], Aug. 2020.
- [19] Yike Liu, Tara Safavi, Abhilash Dighe, and Danai Koutra. Graph summarization methods and applications: A survey. *ACM Comput. Surv.*, 51(3):1–34, Jun. 2018.
- [20] Minghao Xu, Meng Qu, Bingbing Ni, and Jian Tang. Joint modeling of visual objects and relations for scene graph generation. *Adv. Neural Inf. Process. Syst.*, 34:7689–7702, Dec. 2021.
- [21] Xuan Kan, Hejie Cui, and Carl Yang. Zero-shot scene graph relation prediction through commonsense knowledge integration. In *Proc. ECML-PKDD2021*, pages 466–482, Sep. 2021.
- [22] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proc. AAAI2017*, pages 4444–4451, San Francisco, CA, USA, Feb. 2017.
- [23] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the Web. In *Proc. WWW1999*, pages 1–17, Amsterdam, the Netherlands, Nov. 1999.
- [24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. ECCV2014*, volume 5, pages 740–755, Zurich, Switzerland, Sep. 2014.
- [25] Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, and Feng Zhang. Hierarchical photo-scene encoder for album storytelling. In *Proc. AAAI2019*, pages 8909–8916, Honolulu, HI, USA, Jan. 2019.
- [26] Teuvo Kohonen. The self-organizing map. *Proc. IEEE*, 78(9):1464–1480, Sep. 1990.
- [27] Da Deng. Content-based image collection summarization and comparison using self-organizing maps. *Pattern Recognit.*, 40(2):718–727, Feb. 2007.
- [28] Zahra Riahi Samani and Mohsen Ebrahimi Moghaddam. A knowledge-based semantic approach for image collection summarization. *Multimed. Tools Appl.*, 76(9):11917–11939, May 2017.
- [29] Wenkai Zhang, Kun Fu, Xian Sun, Yuhang Zhang, Hao Sun, and Hongqi Wang. Joint optimisation convex-negative matrix factorisation for multimodal image collection summarisation based on images and tags. *IET Comput. Vis.*, 13(2):125–130, May 2018.
- [30] Richard J. Trudeau. *Introduction to Graph Theory*. Dover, New York, NY, USA, 1993.
- [31] Xiaotian Han, Jianwei Yang, Houdong Hu, Lei Zhang, Jianfeng Gao, and Pengchuan Zhang. Image Scene Graph Generation (SGG) benchmark. *Comput. Res. Reposit. arXiv Preprint*, arXiv:2107.12604, [Online], 2021.
- [32] Jinggang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, and Ziwei Liu. Panoptic video scene graph generation. In *Proc. CVPR2023*, pages 18675–18685, Vancouver, BC, Canada, Jun. 2023.
- [33] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannic Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, May 2017.
- [34] Victor Milewski, Marie-Francine Moens, and Iacer Calixto. Are scene graphs good enough to improve image captioning? In *Proc. ACL-IJCNLP2020*, pages 504–515, Suzhou, Jiangsu, China, Sep. 2020.
- [35] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural Motifs: Scene graph parsing with global context. In *Proc. CVPR2018*, pages 5831–5840, Salt Lake City, UT, USA, Jun. 2018.
- [36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Processing Syst.*, 28:91–99, Dec. 2015.
- [37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR2016*, pages 770–778, Las Vegas, NV, USA, Jun. 2016.
- [38] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proc. CVPR2017*, pages 1492–1500, Honolulu, HI, USA, Jun. 2017.
- [39] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Bidirectional long short-term memory networks for relation classification. In *Proc. EMNLP2014*, pages 1–11, Doha, Qatar, Oct. 2014.
- [40] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proc. CVPR2019*, pages 6619–6628, Long Beach, CA, USA, Jun. 2019.
- [41] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proc. CVPR2017*, pages 5410–5419, Honolulu, HI, USA, Jun. 2017.
- [42] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proc. CVPR2020*, pages 3716–3725, [Online], Jun. 2020.
- [43] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. RelTR: Relation transformer for scene graph generation. *IEEE Trans. Pattern Anal. Mach. Intell.*, Apr. 2023.
- [44] Wenhao Yu, Meng Jiang, Zhiting Hu, Qingyun Wang, Heng Ji, and Nazneen Rajani. Knowledge-enriched natural language generation. In *Proc. EMNLP2021*, pages 11–16, [Online], Nov. 2021.
- [45] Ross Girshick. Fast R-CNN. In *Proc. ICCV2015*, pages 1440–1448, Santiago, Chile, Dec. 2015.
- [46] Lin Cheng and Zijing J. Yang. GRCNN: Graph Recognition Convolutional Neural Network for synthesizing programs from flow charts. *Comput. Res. Reposit. arXiv Preprint*, arXiv:2011.05980, [Online], Nov. 2020.
- [47] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for word representation. In *Proc. EMNLP2014*, pages 1532–1543, Doha, Qatar, Oct. 2014.
- [48] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. An end-to-end deep learning architecture for graph classification. In *Proc. AAAI2018*, pages 4438–4445, New Orleans, LA, USA, 2018.
- [49] Jinghui Peng, Zhen Wang, and Shizhe Wang. Similarity calculation method for images based on the scene graph. *Signal Image Video Process.*, 17(5):2395–2403, Jan. 2023.
- [50] Paridhi Maheshwari, Ritwick Chaudhry, and Vishwa Vinay. Scene graph embeddings using relative similarity supervision. In *Proc. AAAI2021*, pages 2328–2336, [Online], Feb. 2021.
- [51] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Proc. ACL2004*, pages 74–81, Barcelona, Catalunya, Spain, Jul. 2004.
- [52] Jorge E. Camargo and Fabio A. González. Multimodal latent topic analysis for image collection summarization. *Inf. Sci.*, 328:270–287, Jan. 2016.
- [53] Sebastian Tschiatschek, Rishabh K. Iyer, Haochen Wei, and Jeff A. Bilmes. Learning mixtures of submodular functions for image collection summarization. *Adv. Neural Inf. Process. Syst.*, 27:1413–1421, Dec. 2014.
- [54] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *Proc. ICLR2020*, [Online], Apr. 2020.
- [55] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT 2019*, volume 1, pages 4171–4186, Minneapolis, MN, USA, 2018.
- [56] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proc. BMVC2018*, pages 12:1–12:13, Newcastle upon Tyne, England, UK, Sep. 2018.
- [57] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. CVPR2015*, pages 3128–3137, Boston, MA, USA, Jun. 2015.
- [58] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. ICLR2014*, Banff, AB, Canada, Apr. 2014.

- [59] Zeina Abu-Aisheh, Romain Raveaux, Jean-Yves Ramel, and Patrick Martineau. An exact graph edit distance algorithm for solving pattern recognition problems. In *Proc. ICPRAM2015*, pages 271–278, Lisbon, Portugal, Jan. 2015.
- [60] Yuren Cong, Wentong Liao, Bodo Rosenhahn, and Michael Ying Yang. Learning similarity between scene graphs and images with transformers. *Comput. Res. Reposit. arXiv Preprint*, arXiv:2304.00590, [Online], Apr. 2023.
- [61] Ji Zhang, Kevin J. Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proc. CVPR2019*, pages 11535–11543, Long Beach, CA, USA, Jun. 2019.



NLP, and JSAL.

TAKAHIRO KOMAMIZU (M'18) received the B.Eng degree in computer science, the M.Eng degree, and the PhD degree in engineering from University of Tsukuba, Japan, in 2009, 2011, and 2015, respectively. He is currently an Associate Professor in the Mathematical and Data Science Center at Nagoya University, Japan. His research interests include database, data analysis, multimedia data managements, and Linked Open Data. He is a member of ACM, IPS Japan, IEICE, DBSJ,



tion and image captioning.

PHUEAKSRI ITTHISAK received the B.S. in computer engineering from Rajamangala University of Technology Phra Nakhon, Thailand, in 2012, M.S degrees in computer science, in 2020 from Chulalongkorn University, Thailand. He is currently a doctoral student at the Graduate School of Informatics, Nagoya University, Japan. His research interests are in computer vision and natural language processing, focusing on scene graph generation and the use of scene graphs in image summariza-



man with multimedia, covering vision and language and affective computing related tasks. Dr. Kastner is a member of IEICE, IPS Japan, and ACM.

MARC A. KASTNER (M'22) received his BSc and MSc in Computer Science from Braunschweig University of Technology, Germany, in 2013 and 2016, respectively. He received his PhD in Informatics in 2020 at the Graduate School of Informatics of Nagoya University, Japan. In 2020, he moved to the National Institute of Informatics, Japan as a Postdoctoral Researcher. Since 2022, he has been an Assistant Professor at Kyoto University, Japan. His research focuses on the connection of the human



et Systèmes Aléatoires (IRISA), France in 2005, 2006, and 2007, a Senior Visiting Researcher at ISLA, Instituut voor Informatica, Universiteit van Amsterdam, the Netherlands from 2010 to 2011. His research interests range from the analysis and indexing to authoring and generation of multimedia contents, especially in large-scale broadcast video archives and social media, mostly on news, cooking, and sports contents. He is a senior member of IEICE and IPS Japan, and a member of ACM, JSAL, and ITE.



Recognition Research Team, RIKEN Guardian Robot Project in Kyoto, Japan. 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 IEEEJ and IEICE. He received the Best Paper Award from SPC2009 and the Young Researcher Award from the IEEE ITS Society Nagoya Chapter.

YASUTOMO KAWANISHI (M'16) 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. In 2012, he became a Postdoctoral Fellow with Kyoto University. 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