IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

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

Digital Object Identifier 10.1109/ACCESS.2023.0322000

An Approach to Generate a Caption for an Image Collection using Scene Graph Generation

ITTHISAK PHUEAKSRI[1,2], MARC A. KASTNER[3],(Member, IEEE), YASUTOMO KAWANISHI[2,1],(Member, IEEE), TAKAHIRO KOMAMIZU[4,1],(Member, IEEE), ICHIRO IDE[1],(Senior Member, IEEE)

[1]Graduate School of Informatics, Nagoya University, Nagoya, Aichi 464-8601 Japan (e-mail: phueaksrii@cs.is.i.nagoya-u.ac.jp, ide@i.nagoya-u.ac.jp)
[2]Guardian Robot Project, Information R&D and Strategy Headquarters, RIKEN, Seika, Kyoto 619-0288 Japan (e-mail: yasutomo.kawanishi@riken.jp)
[3]Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto-shi, Kyoto, 606-8501, Japan (e-mail: mkastner@i.kyoto-u.ac.jp)
[4]Mathematical and Data Science Center, Nagoya University, Nagoya, Aichi 464-8601 Japan (e-mail: taka-coma@acm.org)

Corresponding author: Itthisak Phueaksri (e-mail: phueaksrii@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 is a challenging task that aims to generate a summary by grasping common information of a given set of information. Text summarization is a popular task of determining the topic or generating a textual summary of documents. In contrast, image summarization aims to find a representative summary of a collection of images. However, current methods are still restricted to generating a visual scene graph, tags, and noun phrases, but cannot generate a fitting textual description of an image collection. Thus, we introduce a novel framework for generating a summarized caption of an image collection. Since scene graph generation shows advancement in describing objects and their relationships on a single image, we use it in the proposed method to generate a scene graph for each image in an image collection. Then, we find common objects and their relationships from all scene graphs and represent them as a summarized scene graph. For this, we merge all scene graphs and select part of it by estimating the most common objects and relationships. Finally, the summarized scene graph is input into a captioning model. In addition, we introduce a technique to generalize specific words in the final caption into common concept words incorporating external knowledge. To evaluate the proposed method, we construct a dataset for this task by extending the annotation of the MS-COCO dataset using an image retrieval method. The evaluation of the proposed method on this dataset showed promising performance compared to text summarization-based methods.

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

I. INTRODUCTION

Following the recent increase of the number of images in the real world, describing images has become one of the important image-to-text tasks. Specifically, the image captioning task [1], [2] is widely known as an approach for translating an image into a short description. However, since it is conceptually restricted to a single image, it is challenging to apply it to an image collection. The image collection summarization task focused in this paper aims for grasping common contents and understanding the overall visual features and relationships in an image collection. It could generate visual descriptions of concepts such as encyclopedia entries by generating a description for a large number of images for the same, e.g, animal species. Existing methods [3]–[5] working on this task can generate representative keywords, tags, or phrases. For example, it could automatically cluster and describe lots of images from Web pages [6]. However, they lack context awareness of the image collection. For example, they often select ''animal'' or ''white animal'' as a keyword for an image collection of animals whose context cannot be described. From a different perspective, image retrieval [7], [8] has been introduced as an approach to the image collection summarization task that selects an image from an image collection as a representative one. In recent years, scene graph generation has been introduced in various tasks, such as image captioning and image retrieval, which has shown the advancement in understanding objects and their relationships of objects in an image. Other works generate a caption for

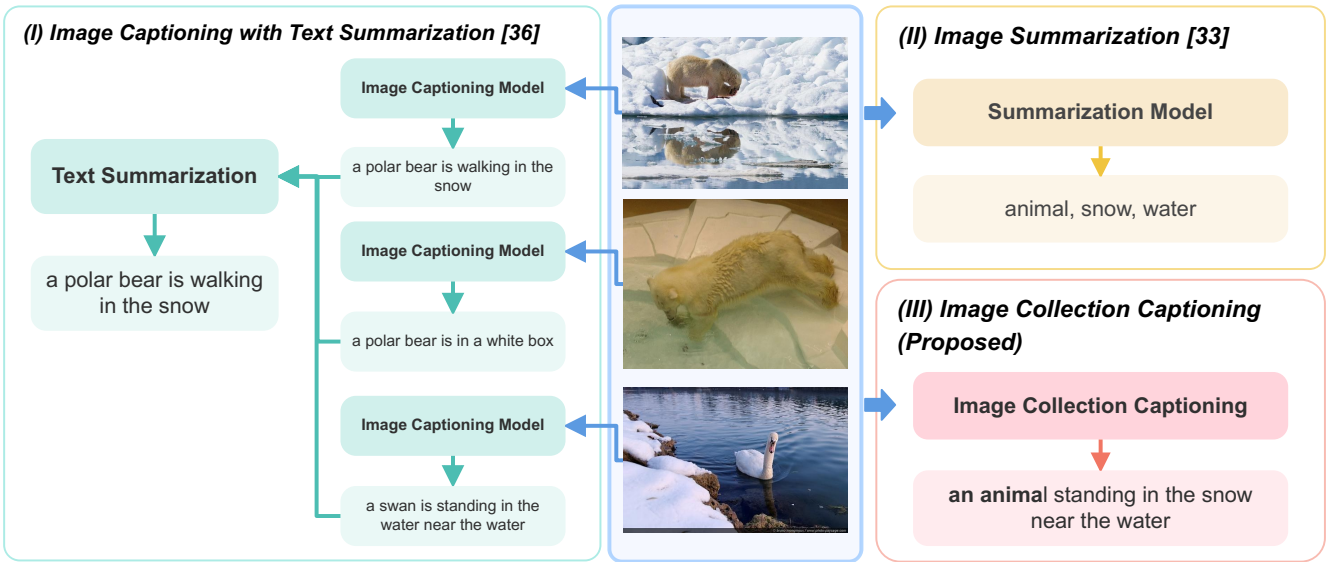


FIGURE 1. **Comparison of (I) image captioning with text summarization, (II) image summarization, and (III) image collection captioning (proposed): (I) generates a single caption from each image in an image collection, and then all the generated captions are passed into text summarization to generate a summary. (II) aims to grasp common visual features and then generate summarized information into keywords or tags. (III) generates a caption of an image collection and also generalizes specific words into common concept words.**

each image in an image collection and apply text summarization [9] to summarize all captions into a final caption. This approach tends to describe the most common occurring words over all images. As such, most recent works on this task still lack summarizing contexts (objects and relationships) of an image collection and focus on generating a summary based on the most commonly occurring contents instead of generalizing the concepts. In our preliminary work [10], we have introduced an image captioning task that aims to generate a caption to describe the overall objects and relationships of an image collection using the advantage of understanding all scene graphs in an image collection and incorporating external knowledge to find generalized concepts of an image collection. The idea of an image collection captioning task is visualized in Fig. 1, where we compare it to other related tasks.

In order to approach the image collection captioning task, there are three hurdles. First, we need to understand the overall visual features and relationships of each individual image. Second, we need to summarize the scene graphs of each image to an appropriate combined representation. Third, we need to caption this combined representation to generate a textual description, which is the fixed output. For the first step, we employ scene graph generation [11] on each image of an image collection. Then, we merge all image scene graphs into a combined scene graph and select the most co-occurring visual features and relationships. For the second step, we aim to generalize concepts across multiple images. For example, as visualized in Fig. 2, if one image contains the concept "man" and another the concept "woman", the combined representation should be generalized to "person". For this, we introduce a component called "*Sub-Graph Concept Generalization*". From the idea of word communities [12] from abstractive text summarization [13]–[15], we incorporate external knowledge. While merging scene graphs, we create intermediate representations of such generalized concepts. For the third step, we employ and compare two image captioning architectures, namely Graph Attention (GAT) model [16] and VinVL [17]. On these architectures, we first generate an initial image caption from the summarized scene graph. Next, we employ a step called "*Sentence Refinement*", where we update noun phrases based on the generalized concepts from the "*Sub-Graph Concept Generalization*" component. This allows us to create better-generalized captions for the whole image collection.

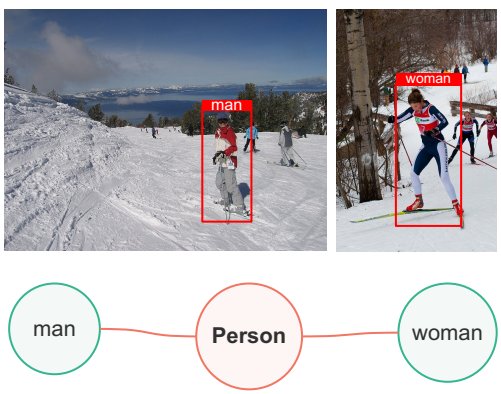


FIGURE 2. **Idea of the concept generalization. This example shows finding a common concept between "*man*" and "*woman*"; "*person*."**

This research is based on our previous work [10]. It showed promising results for an image collection captioning task with scene graph generation and GAT model incorporating

external knowledge, ConceptNet [18], on an image collection consisting of six images. As an improvement from this, in this paper, we modify the *Scene Graph Merging* component which also uses only frequent occurrences of objects (vertices) in the *Scene Graph Selection* component whereas the prior work used only the frequent occurrences of objects (vertices) and relations (edges). We also employ and compare GAT and VinVL as two different captioning backbones to generate a caption for each image collection. Based on the *Sentence Refinement* component of the previous work aiming to refine noun phrases considering determiners, we modify it to correct the determiner after replacing noun phrases with common concept words following grammatical rules. The initial work showed promising results limited to six images, while in this paper, we increase the number of images in testing image collections in the experiment with image collections containing 11 and 16 images. The *Scene Graph Generation* and *Sub-Graph Concept Generalization* components are similar to the previous work. The proposed method using the graph attention model backbone shows a significant improvement. The model which uses the pre-trained vision-language model captioning backbone achieves good results when increasing the number of images in an image collection, especially when evaluated by semantic text-based auto-evaluation metrics.

Our contributions can be summarized as follows:

- We introduce a new challenging task, an image collection captioning task, which aims to generate a caption of an image collection.
- We propose a baseline framework for the novel task using a combined scene graph captioning approach, which handles the combined scene graph as a representative of all images and then generates a caption from it.
- We show that the proposed scene graph summarization approach with the GAT model and the pre-trained VinVL model is proper in generating a caption with the current state-of-the-art methods.
- We construct a dataset for this task by extending the annotation of the MS-COCO dataset incorporating an image-retrieval task.

II. RELATED WORKS

A. MULTIPLE-IMAGE SUMMARIZATION

The multiple-image summarization task [3], [4], [6] was recently introduced. It aims to find semantic concepts or general concepts that represent an image collection. Finding images in the image collection as representatives of the image collection is also introduced in album summarization tasks [19], [20]. Scene graph generation [21] is used to generate a summary of an image collection by estimating the similarity between vertices and edge features. Samani et al. [3] proposed a method to find a semantic summarization of an image collection integrating specific domains of external knowledge by estimating the semantic information of each image. Zhang et al. [4] proposed a novel approach to analyze the visual (images) and textual (topic) information of an image collection to generate a summary. Trieu et al. [6] proposed the multiple-image summarization task, which aimed to generate a descriptive textual summary for an image collection in noun phrases. They also introduced a dataset by gathering images from Web pages consisting of 2.1 million images and building collections of images from them, with each collection containing at least five images. Gencer et al. [9] introduced a caption summarization method for remote-sensing scene images. In our previous work [10], we introduced a novel image collection captioning method on an image collection comprising six images using the Graph Attention (GAT) model [16].

B. CAPTIONING MODELS

Image captioning is an image-to-text translation task that aims to describe an image in a sentence, aided by object detection. The MS-COCO [22] dataset is popular for this task. Various models for the image captioning task commonly consist of an encoder and a decoder that respond to extract the meaning of visual features and generate a caption, respectively.

a: Graph Attention for Captioning Model

Since the visual scene graph has been introduced to describe an image with graphs, the Graph Attention (GAT) model is introduced as a captioning model for the image captioning task taking advantage of detecting an image in a scene graph. Nguyen et al. [23] proposed a framework named SG2Caps that utilized spatial locations of visual scene graph nodes and human-object interaction information to generate a caption from a visual scene graph. Milewski et al. [16] proposed a conditional GAT network that aimed to investigate the difference of generated visual scene graphs to fuse the object and relation information for caption generation. Zhong et al. [24] proposed a novel image captioning model by exploring the important sub-graphs of an image scene graph to be the input of the decoder in caption generation.

b: Pre-trained Vision Language Models

Generating a caption using pre-trained vision-language models achieves good results in caption generation. BEiT-3 [25] is the current state-of-the-art in the vision-language task with multi-way transformers. mPlug [1] is a new asymmetric vision-language architecture that addresses the problems of information asymmetry and computational efficiency. Zhang et al. [17] presents a vision-language model to show the importance of realizing visual features and pre-trained on a large training corpus consisting of many annotated object detection datasets. Based on the performance of VinVL, which is trained on both Visual Genome [26] and MS-COCO [22] datasets, we fine-tune it as a captioning backbone.

C. SCENE GRAPH GENERATION

Scene graph generation is a method of describing an image using objects and their relationships in a graph form. The scene graph generation architecture is designed with an object detector to find the objects and an edge detector to detect their relationships. Faster R-CNN [27] is a popular

technique, used as a backbone of the object detector. Neural Motif [11] is a model constructed from Motif Network [28] by estimating the local and global contexts of images using bi-directional Long Short-Term Memory (LSTM) [29] on the Visual Genome dataset and further evaluated on the MS-COCO dataset. Zhang et al. [30] proposed Graphical Contractive losses which effectively utilize semantic information of scene graphs to overcome the issues of object entity confusion and relationship ambiguity. Graph Recognition Convolutional Neural Network [31] was introduced by implementing a convolutional network on the graph to predict edge contexts. Since a long-tail distribution on the Visual Genome dataset is mentioned, Tang et al. [32] proposed a novel framework named Internal and External Data Transfer to automatically transfer data from general predicates to informative ones and relabel relations that are not correctly annotated by annotators. Cong et al. [33] proposed a novel on-stage end-to-end framework for scene graph generation by giving a fixed number of coupled subjects and objects, and a fixed size of relationships using attention mechanisms.

D. TEXT SUMMARIZATION

Text Summarization is known as a text-to-text generation task that aims to generate a topic or a summary of documents. In recent paradigms, text summarization is introduced in two main types: extractive text summarization and abstractive text summarization. Extractive text summarization aims to obtain salient information from the documents. T5 [34] is a strong baseline of the supervised text summarization model, which is pre-trained using the Wikipedia dataset[1]. SUPERT [35] is an unsupervised learning model for multi-document summarization that evaluates all sentences in multiple documents and selects one of them as a topic of the documents. However, extractive text summarization is limited to generating a summary based on document content. To overcome the limitation of extractive text summarization, abstractive text summarization [36] is introduced, which focuses on generating a summary with common words. It aims to generate a summary by rewriting the summarized sentence using the semantic content found in the documents. XL-Sum [37] is proposed as a multilingual abstractive text summarization method on a large multilingual article dataset by fine-tuning the T5 model.

III. PROPOSED METHOD: IMAGE COLLECTION CAPTIONING

The proposed framework starts with generating a scene graph for each image in an image collection, then summarizing them into a single scene graph as a representative scene graph of an image collection. The scene graph is used to generate a caption by using a captioning model. We further find common concept words from all scene graphs and then use them to refine the initial caption.

[1]https://www.tensorflow.org/datasets/catalog/wikipedia/ (Accessed: May 30, 2023)

Our previous method [10] also addresses the task of generating a caption of an image collection, making use of scene graph generation. However, we found its limitation in finding common visual features of an image collection and in refining the final caption. Additionally, the previous work was implemented only on a small image collection containing six images. For this, in this follow-up work, we have modified the process of generating a summarized scene graph considering only the common vertices instead of the common vertices and edges, by fine-tuning a pre-trained vision-language Model; VinVL [17].

Based on the idea, we propose a framework of image collection captioning consisting of five main components as shown in Fig. 3: (A) The first component is *Scene Graph Generation*, which generates a scene graph of each image in a semantic image collection. (B) Then, all scene graphs are passed into the *Scene Graph Merging* component to combine them into a combined scene graph. (C) From a combined scene graph, the *Sub-Graph Selection* component selects a summarized scene graph as a representation. (D) The *Sub-Graph Concept Generalization* component finds common concept words from a combined scene graph. (E) The *Captioning Model* generates an initial caption from the summarized scene graph. (F) Finally, the initial caption is refined by the *Sentence Refinement* component that generalizes specific words of the initial caption to enable the capability of describing the caption for all images in an image collection.

In this chapter, Sec. III-A and III-E revisit parts of our previous method [10], while Sec. III-B, III-C, and III-F describe the newly introduced *Sub-Graph Merging* component that merges all scene graphs into a combined scene graph and counts the occurrence node, and the *Sentence Refinement* component that refines the initial caption.

A. SCENE GRAPH GENERATION

As scene graph generation shows the advantage of understanding image contexts on various tasks, especially an image captioning task, we implement scene graph generation of ResNet101-FPN [38] as a backbone and Neural Motif [11] as a relation predictor. This scene graph generation model is trained by the Visual Genome dataset [26], which is a popular visual scene graph dataset used for image captioning tasks. Due to ambiguous labels, a recent work in image captioning [39] cleans up duplicate labels from 2,500 object labels, 1,000 attribute labels, and 500 relation labels to 1,600 object labels, 400 attribute labels, and 20 relation labels, which can improve the captioning performance. We then build a directed scene graph in which scene graphs of each image are represented in triplets, including subject, predicate, and object. Following the scene graph generation benchmark [21], we determine the maximum detection number to 100 for both detecting objects and relationships.

B. SCENE GRAPH MERGING

Following our previous work [10], this component aims to estimate the common occurrence of objects and relationships

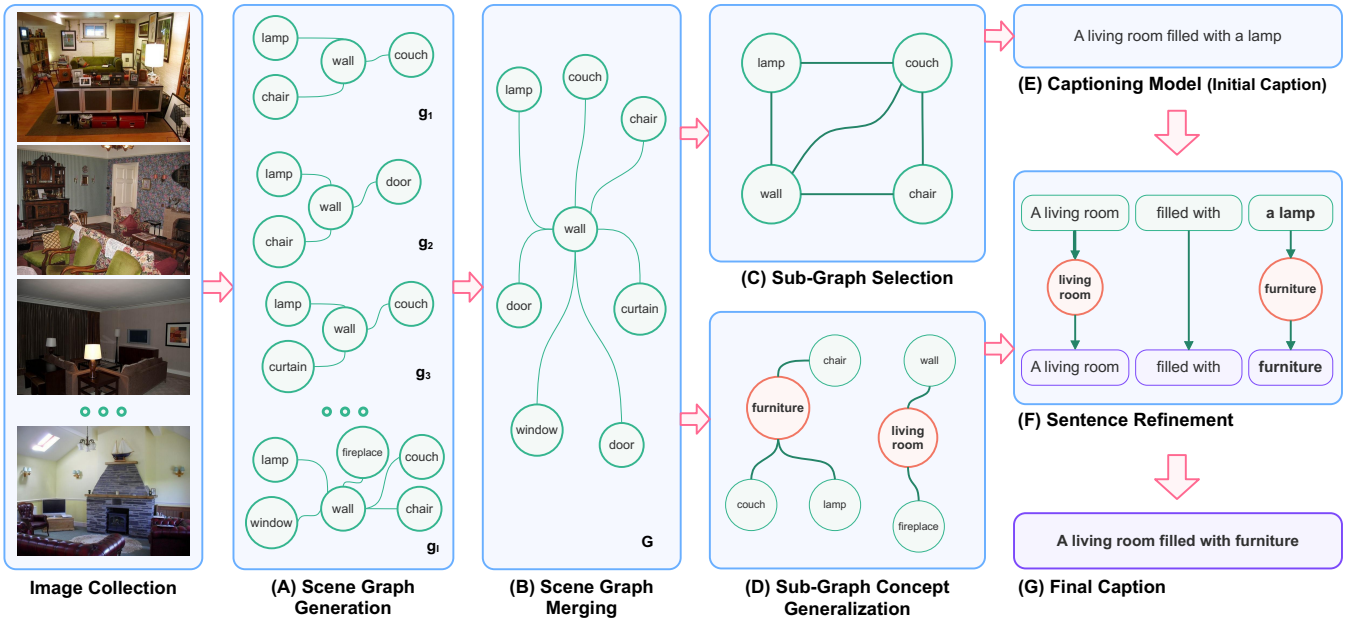


FIGURE 3. **Overview of the proposed method; An image collection captioning method, consisting of five major processes: (A) *Scene Graph Generation* extracts a scene graph from each image in an image collection. (B) *Scene Graph Merging* merges all scene graphs into a combined scene graph. (C) *Sub-Graph Selection* finds the representative graph from a combined scene graph. (D) *Sub-Graph Concept Generalization* builds word communities of a combined scene graph incorporating external knowledge and then finds common concept words as representation. (E) *Captioning Model* generates an initial caption from the representative graph. (F) *Sentence Refinement* refines the initial caption with common words from the *Sub-Graph Concept Generalization* process as (G) the *Final Caption*.**

(vertices and edges in a set of graphs). Since increasing the number of images in an image collection causes ambiguity in the relationships of visual features across the images. We hence focus on estimating only the common occurrence of objects in this work.

To summarize all scene graphs from scene graph generation into a combined scene graph, we first build a scene graph summarization component. From all scene graphs generated in the scene graph generation component, we merge all directed scene graphs by union vertices and edges of all image scene graphs, in which each image scene graph is represented as $g_i = (V_i, E_i)$, where V_i is a set of vertices and $E_i \subseteq V_i \times V_i$ is a set of edges of the i-th scene graph, into a combined scene graph $G(V, E)$ as:

$$G(V,E) = \left(\bigcup_{i}^{I} V_i, \bigcup_{i}^{I} E_i\right), \tag{1}$$

where I is the number of images in the image collection.

C. SUB-GRAPH SELECTION

To find a summarized scene graph, in this section, we discuss how to select the most occurred context. Following measuring the centrality based on graph theory, we adopt betweenness centrality [40] to measure graph centrality based on the shortest paths for every pair of vertices to estimate the most occurrences of vertices, whereas degree centrality focuses on estimating the highest degree vertices and closeness centrality focuses on the average shortest path vertices. In our experiment, we found betweenness centrality is the most effective method for finding common contexts from a combined scene graph.

As a combined scene graph from *Scene Graph Merging* and focusing on estimating the centrality, we select the top occurred vertices from a combined scene graph as a summarized scene graph. Following the implementation of betweenness centrally, we also count the occurrences of each vertex and use them as the weight in the estimation as:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{2}$$

where C_B is the sum of the fraction of the shortest paths of all pairs that pass through v, $\sigma_{st}(v)$ is the number of paths from s to t passing through v, and σ_{st} is the total number of the shortest paths from s to t.

In the following, we build a summarized sub-graph using all the betweenness centrality values. First, we sort all betweenness centrality values in descending order. Then, we select the vertices with the highest value and all connected vertices, in which we limit the number to 100 vertices in the following experiment, which is the same number as the scene graph generation. Lastly, a summarized scene graph is constructed as a subgraph from all selected vertices and edges.

D. CAPTIONING MODEL

Since we aim at generating a caption of an image collection but there is no dataset for training, we follow a single image captioning approach. We transfer a captioning model to

an image collection captioning framework for the inference phrase. Based on the idea of finding common information using scene graph generation, we first implement the *Graph Attention (GAT)* model, which is a popular method in generating a caption based on scene graph generation. However, state-of-the-art image captioning using a pre-trained vision-language model has shown advancement in results [17], [25]. Thus, in the following experiments, we will use *VinVL* [17], a pre-trained vision language model, as a captioning model.

As the proposed method focuses on processing sub-graphs from scene graph generation, we first implement a graph attention model consisting of a Graph Convolutional Network (GCN) and an Attention-based Long Short-Term Memory (LSTM) model [16]. We build the GCN layer to encode triplets of subject, predicate, and object from a summarized scene graph. Each triplet feature is represented in 2,048 dimensions and is encoded into 1,024 dimensions. We next build two layers of the attention-based LSTM model following the top-down captioning model to generate a caption.

E. SUB-GRAPH CONCEPT GENERALIZATION

In this section, we discuss how to find a general concept from specific contexts of an image collection. Since a generated caption using an image captioning model is based on training visual features and vocabularies, we introduce the *Sub-Graph Concept Generalization* component which aims to improve the ability in describing an image collection overall by finding concept vocabularies incorporating external knowledge. For example, an image collection of people in which each image contains "man" and "woman," or a collection of animals which contains "bear," "elephant," and "bird." These concept words will be used for refining a caption generated from the captioning model. In the following, our idea is to find common words among the specific words instead of generating a caption based on the limitation of training visual features and vocabularies. Inspired by abstractive text summarization, the *Sub Graph Concept Generalization* component finds common words of an image collection. To find the common words, we build word communities incorporating ConceptNet [18], which is a popular text-based semantic network. We extend the related words in word communities and then find the representative of each word community to refine the final caption.

With text analysis based on word relationships from abstractive text summarization, this component aims to find the representative word in each word community. First, we lemmatize all words. Next, we build word embedding graphs from the *Scene Graph Generation* component. Then, we expand the word embedding graph with related words incorporating ConceptNet [18], which provides various relationships (e.g., *related terms*, *synonyms*, and *antonyms*). Since our objective focuses on finding the semantic words as a representation, we focus on *related terms* and *synonyms* in the expanding process.

Next, we join all the expanded word embedding graphs and drop all non-degree vertices. Consequently, each sub-graph is

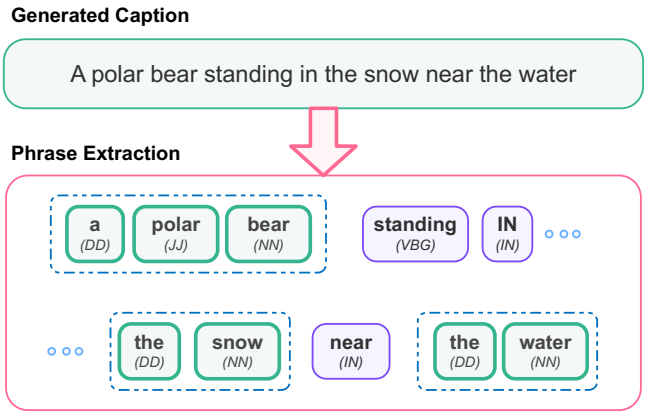


FIGURE 4. Example of phrase extraction from a generated caption.

determined as a word community. To find the representative word in each community, we employ Glove word embedding [41] to encode all words in the word embedding graph into word embedding features. We then define the similarity between vertices as edge weight by calculating the similarity, S as follows:

$$S(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\cdot\|\mathbf{b}\|}, \tag{3}$$

where $\mathbf{a}$ and $\mathbf{b}$ are the corresponding embeddings of given two vertices, $\mathbf{a}\cdot\mathbf{b}$ is dot product between vectors $\mathbf{a}$ and $\mathbf{b}$, and $\|\mathbf{a}\|$ is the L_2 norm of vector $\mathbf{a}$.

In finding the representative word of each word community, we aim to find the highest degree of vertex considering the average shortest path. We hence implement the improved closeness centrality [42], which is good in estimating the centrality of the graph with many connections to estimate the centrality vertex of each community as:

$$C_C(\mathbf{u}) = \frac{n-1}{N-1}\frac{n-1}{\sum_{v=1}^{n-1} S(\mathbf{u},\mathbf{v})}, \tag{4}$$

where $C_C(\mathbf{u})$ is the closeness centrality of vertex $\mathbf{u}$, n is the number of all reachable vertices, $n-1$ is the number of vertices reachable from $\mathbf{u}$, N is the number of vertices in the graph, and $\mathbf{v}$ is a vertex.

F. SENTENCE REFINEMENT

As the objective of the proposed method is to generate a caption that can describe the overall contents of an image collection, we build a sentence refinement component to generalize a final caption for an image collection from specific words with concept words.

In order to generalize nouns in a sentence, we first construct noun phrases from a generated caption of an image collection as shown in Fig 4. For this, we implement NLTK POS Tagging [43] to define the part-of-speech tag for each word in a sentence. Next, we determine a sequence of nouns, proper nouns, premodifiers, postmodifiers, and determiners, in which the part of speech tag is NN, NNP, JJ, JJ, and DT,

respectively, as noun phrases. Lastly, we define a noun, in which the part-of-speech tag is NN, as a main noun to be a representative of each community.

To reconstruct a generated caption with common concept words, we first find the relationships of concept words from the *Sub-graph Concept Generalization* component and noun phrases. From noun phrases, all main nouns are used to search in the sub-graph concept. Then, all main nouns, found in the sub-graph concept, are replaced with representative words of word communities. Lastly, we construct the final caption from a sentence word graph that is mapped with a sub-graph concept. Furthermore, we also correct the determiner of main nouns that are replaced with a representative in the final caption following the grammatical rule, for example, the specific word of "bear" or "bird" in which the determiner, is "a" whereas a concept word is "animal" in which the determiner is "an" as shown in Fig. 5. Additionally, we demonstrate the refinement process in Algorithm 1.

Algorithm 1: Sentence Refinement

Input: $caption_{generated}$, $Sub\_Graph\_Concept$
Output: A summarized scene graph
Result: $caption_{final}$
$tokens$ = word_tokenize($caption_{generated}$);
$tagged$ = pos_tag($tokens$);
$words$ = [];
foreach $i \in range(length(tokens))$ **do**
  if $type(tokens[i]) ==$ NN **and** $tokens[i] \in Sub\_Graph\_Concept$ **then**
    $modifiers$ = []
    while $length(words) > 0$ **do**
      $\_word = words[i-1]$;
      if $type(\_word) \in \{DT, JJ, NNP\}$ **then**
        $modifiers.push(\_word)$;
        $words.pop()$;
      else
        $words.push(modifiers)$;
        $words.push(Sub\_Graph\_Concept[word])$;
        $break$;
      end
    end
  else
    $words.push(tokens[i])$;
  end
end
$caption_{final}$ = " ".join($words$)

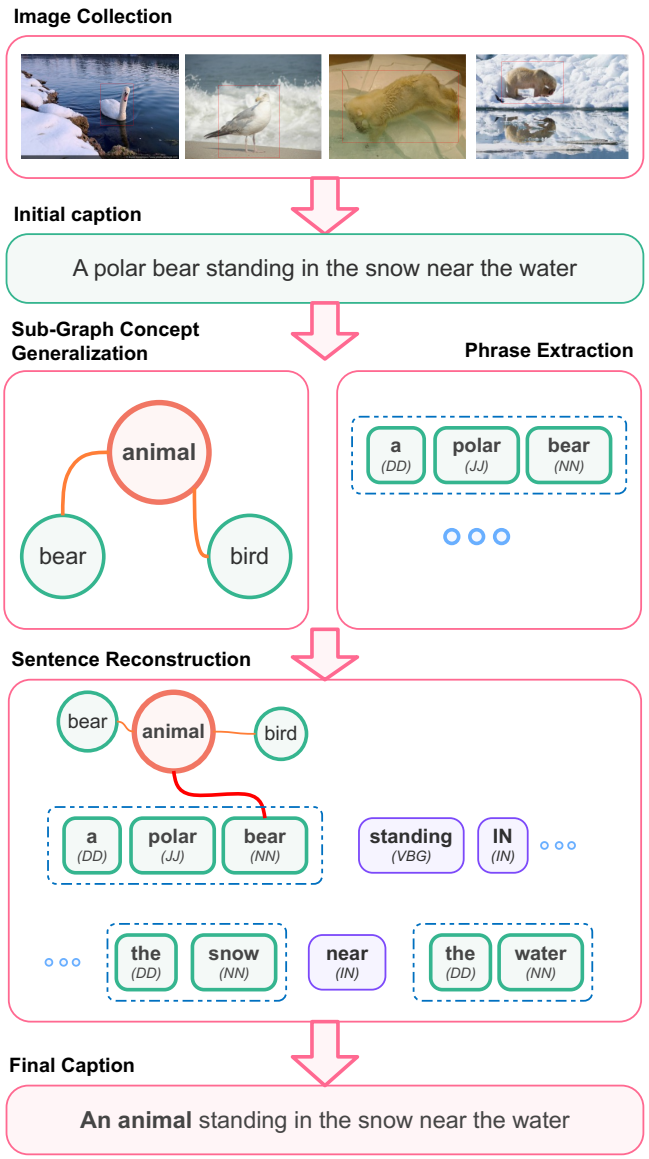


FIGURE 5. Process of *Sentence Refinement* to construct a final sentence mapped between sub-graph concept and phrase extraction.

IV. EXPERIMENTALS

A. DATASET

There is no existing dataset for the proposed task of generating a caption for an image collection, but the most related one is the MS-COCO [22] dataset which is popular for the single image captioning task.

In the following experiments, the MS-COCO dataset is used in this task for training, fine-tuning, and evaluation. We follow the Karpathy split [44], where 123k images and is split into 118k for training, 5k for validation, and 5k for testing. Each image has 5 captions for training and fine-tuning. To evaluate the proposed method, we build the testing set from 5k images of the MS-COCO testing set. We implement VSE++ [45], which is a technique to query similar images by learning from visual-semantic embedding based on estimating visual features of images and textual features of captions. Following the retrieval technique, we generate an image collection from 5k testing images. Each image in a testing set is defined as an initial image, and then we query the top-k semantic images of the testing set. The number of k is set to 5, 10, and 15, in which 5 and 10 are reported in the experimental results [45], while 15 is to validate the proposed method that is still able to describe the overall content of an

image collection. Hence, the evaluation dataset would consist of an initial image and the top-k query images, resulting in three different sizes of image collections comprised of 6, 11, and 16 images as shown in Figure 6. In the following, the ground truth captions for the evaluation process are built from all images in each image collection, in which each image consists of 5 captions. Therefore, an image collection comprised of 6 images has 30 captions, an image collection comprised of 11 images has 55 captions, and an image collection comprised of 16 images has 80 captions.

B. TRAINING STRATEGY ON GRAPH ATTENTION MODEL

Due to the limitation of datasets as explained above, we adapt the MS-COCO [22] dataset for training and validating the Graph Attention (GAT) captioning model [16].

From the MS-COCO dataset, we generate scene graphs for all images by the scene graph generation model. For training, we implement an initial learning rate of 0.0008, a decay rate of 0.8 every 8 epochs, and Adam [52] as an optimizer. Cross-entropy loss and multi-label margin loss are implemented as loss functions. To decide the best model for the proposed method, we evaluate the training process with the CIDEr metric [47].

C. PRE-TRAINED VISION LANGUAGE MODEL

Since large pre-trained language models show advancement in text generation, we adapt VinVL [17], a pre-trained vision-language model, as the captioning model which is designed and trained by large corpora, such as the Visual Genome dataset [26] and the MS-COCO [22] dataset. Following the experimental report of pre-trained large-scale object-detection of C4 and FPN models of VinVL, we use RestNext152-C4 [53] as a backbone object detector which shows the best result compared with others [17]. Whereas the scene graph generation in the proposed method is pre-trained with the Visual Genome dataset which uses ResNet101-FPN [54] as a backbone object detector and the image collection is built from the MS-COCO dataset. As the configurations of VinVL are similar to the proposed method, we implement following the practice of image captioning using the VinVL model, which uses the seq2seq objective. Since VinVL is not built for generating a caption from a scene graph, we use only the object features in a scene graph to generate a caption. We also follow 15% of the random masking out of the caption tokens, which is the same as the VinVL configuration.

To fine-tune the VinVL model, we choose a checkpoint of image captioning trained on the MS-COCO dataset initially trained with cross-entropy loss. In fine-tuning, we implement a learning rate of 0.00003 and a decay rate of 0.05. To make a decision on the best model, we evaluate the generated caption on the evaluation set with the CIDEr metric.

D. EVALUATION

Due to the limitation of ground-truth captions of an image collection, we annotate it by collecting all 5 captions annotated to each image with 5 captions in the MS-COCO [22] dataset. Therefore, the ground truth of 6, 11, and 16 images would have 30, 55, and 80 ground-truth captions, respectively.

With the objective of an image collection captioning task which aims to generate a caption that can describe the overall contexts of an image collection, we compare a generated caption of the image collection to each caption of the ground-truth captions using automatic evaluation metrics.

a: Baseline

To evaluate the proposed methods using text summarization models, we select three baseline text summarization models; T5 [34], XL-Sum [37], and SUPERT [35]. The T5 model is a strong baseline of extractive text summarization using supervised learning. XL-Sum is an abstractive summarization model that fine-tunes the T5 model with 1.35 million articles. SUPERT is an unsupervised text summarization model. To apply text summarization models to the image collection captioning task, we first generate a single caption for each image with a captioning model. Then, we generate a summarized caption of all captions with the text summarization models.

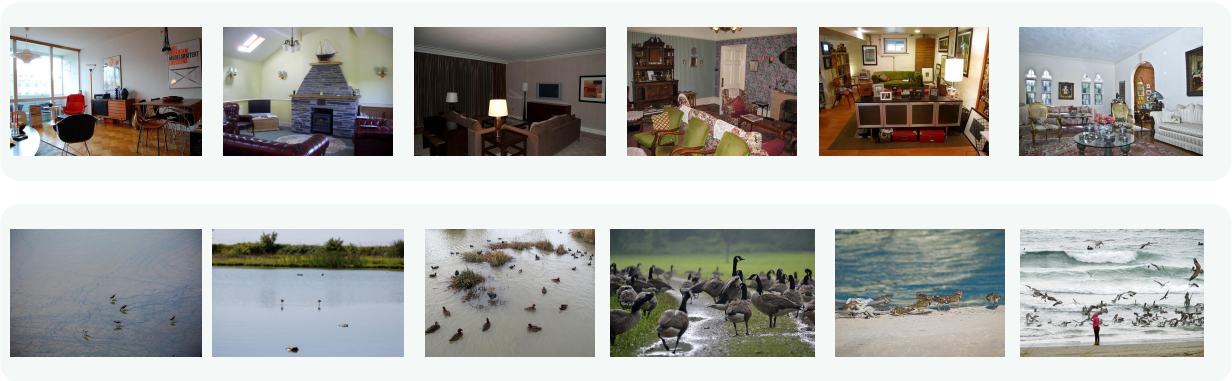
In addition, we compare with our previous work, Image Collection Captioning (ICC) [10], which focuses on summarizing common relationships of all image scene graphs and refining noun phrases with common words without considering phrase modification.

b: Metrics

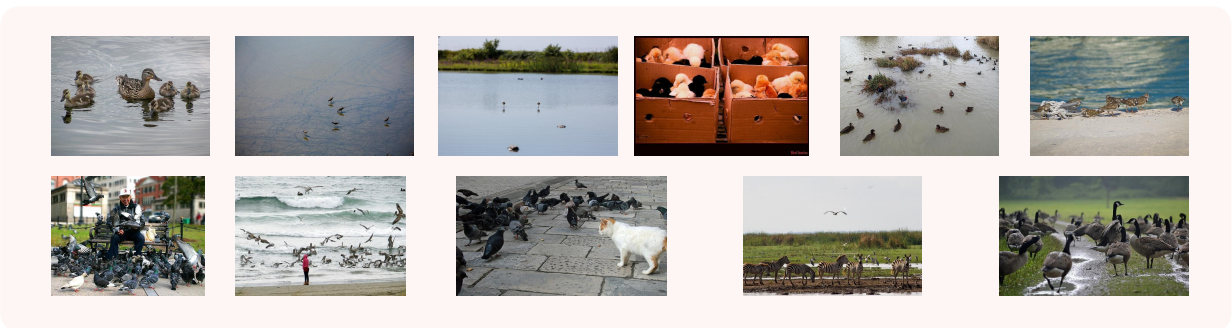
To evaluate the generated captions for an image collection, we used text-based evaluation metrics and machine learning metrics. For text-based evaluation, we use BLEU (2/4) [46], CIDEr [47] and ROUGE (1/2/L) [55]. As the proposed method aims to generalize specific words of a generated caption into concept words, we employ machine learning-based evaluation metrics that focus on estimating the semantic similarity between word tokens. First, we use BERTScore [49], an evaluation metric that calculates the similarity between word tokens based on Bidirectional Encoder Representations from Transformers (BERT) [56] contextual embeddings. Then, we use MoverScore, an unsupervised evaluation metric that combines contextual embeddings and Earth Mover's Distance (EMD) [57]. Lastly, since the proposed method is inspired by abstractive text summarization [58], we also use ROUGE-WE [48] and WEEM4TS [51], which are introduced to evaluate abstractive text summarization.

E. RESULTS

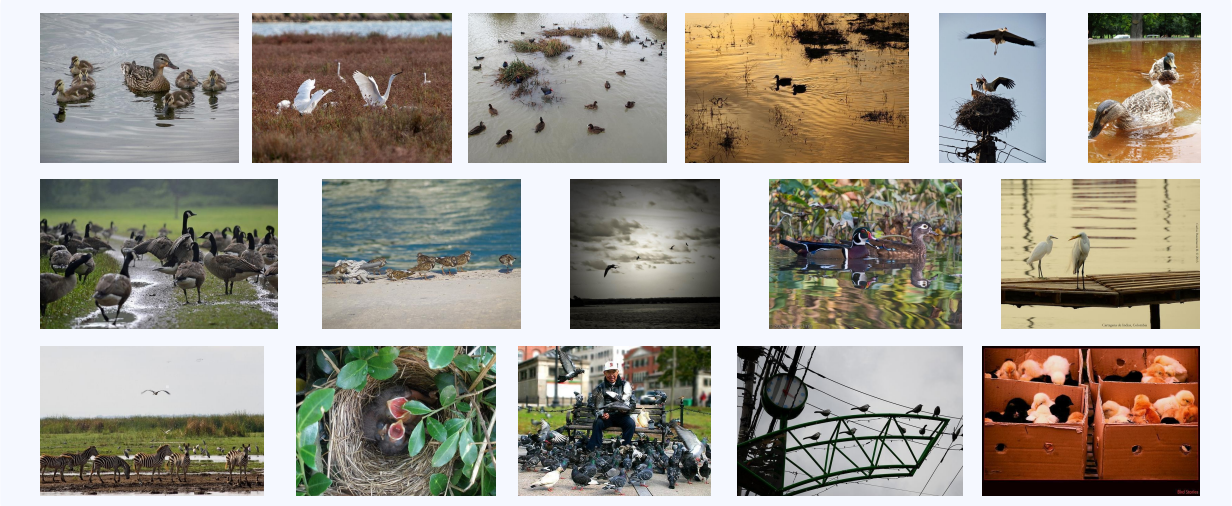
We report the experimental result of the proposed method on 5k test image collections in Tables 1, 2, and 3 for image collection sizes of 6, 11, and 16 images, respectively. To compare the experimental results, the results of the use of text-summarization models are ablated with two methods; Graph Attention (GAT) [16] and VinVL [17]. Meanwhile, the results of our previous work, ICC [10] and the proposed method are ablated with two methods; with and without the *Sub-Graph Concept Generalization (CG)* component.



(A) 6 images in a collection



(B) 11 images in a collection



(C) 16 images in a collection

FIGURE 6. Examples of image collections.

TABLE 1. Evaluation of the summarization results of image collections which contain six images in which all sentences are limited in length to twenty words, compared to ICC (our previous work) [10], SUPERT [35], T5 [34], and XL-Sum [37]. "w/ GAT" is the result generated by the Graph Attention (GAT) model. "w/ LM" is the result generated by VinVL, "CG" is the *Sub-Graph Concept Generalization* component. "+" is the result generated by implementing the CG and "–" is the result generated without implementing the CG. "B-*n*" is BLEU [46], "C" is CIDEr [47], "R-*n*" is ROUGE Score [48], "R-WE" is ROUGE-WE [48], "BERT" is BERTScore [49], Mover is MoverScore [50], and WEE is WEEM4TS [51]. Results in bold indicate the highest scores and those underlined indicate the second highest scores.

Models	Captioning backbone	CG	B-2 ↑	B-4 ↑	C ↑	R-1 ↑	R-2 ↑	R-L ↑	R-WE ↑	BERT ↑	Mover ↑	WEE ↑
SUPERT [35]	w/ GAT	N/A	0.559	0.365	0.702	0.376	0.111	0.323	0.074	0.612	0.608	0.108
	w/ LM		0.724	0.461	0.613	0.615	0.434	0.586	0.080	0.834	0.592	0.101
T5 [34]	w/ GAT	N/A	0.740	0.500	0.552	0.344	0.104	0.303	0.073	0.606	0.565	0.103
	w/ LM		**0.755**	0.491	0.626	0.686	0.487	0.657	0.067	0.923	**0.610**	0.097
XL-SUM [37]	w/ GAT	N/A	0.271	0.099	0.102	0.215	0.037	0.183	0.042	0.468	0.521	0.086
	w/ LM		0.268	0.107	0.140	0.384	0.171	0.040	0.081	0.873	0.523	0.091
ICC [10]	w/ GAT	–	0.742	**0.508**	0.796	0.378	0.127	0.341	0.081	0.627	0.570	0.106
		+	0.701	0.475	0.716	0.352	0.111	0.314	0.072	0.609	0.565	**0.110**
Proposed (Scene-graph)	w/ GAT	–	0.744	0.507	**0.810**	**0.722**	**0.532**	**0.699**	0.081	**0.927**	0.571	0.106
		+	0.729	0.488	0.768	0.712	0.515	0.685	**0.082**	0.924	0.569	**0.110**
Proposed (Scene-graph)	w/ LM	–	0.567	0.324	0.475	0.558	0.356	0.529	0.067	0.851	**0.610**	0.107
		+	0.557	0.313	0.456	0.553	0.348	0.525	0.067	0.850	0.609	0.108

a: Text Summarization w/ GAT

From the results of using the captioning backbone, the T5 text summarization model applied to the image collection comprising 16 images achieved the second-highest scores as shown in Table 3, whereas the ICC and the proposed method showed better results on image collections comprising 6 and 11 images as shown in Tables 1 and 2. In addition, the SUPERT model showed better results than those of the T5 model when evaluated with CIDEr, ROUGE, BERTScore, MoverScore, and WEEM4TS only on the image collection comprising 6 images. Finally, the results by the XL-Sum, abstractive text summarization model, showed the worst results on all image collections.

b: Text Summarization w/ LM

From the results of using the VinVL model as a captioning backbone, Table 1 shows that T5 achieved the best score on BLEU-2 and the second-highest score on MoverScore. Tables 2 and 3 show that the T5 model achieved the highest score when evaluated with BLEU scores and the highest scores on BERTScore. Meanwhile, the SUPERT and XL-Sum models could not achieve good results compared to the T5 model. In addition, when comparing the use of text summarization with ICC and the proposed method, most of the results achieved better scores, especially the evaluation score on CIDEr, ROUGE, BERTScore, and WEEM4TS.

c: Proposed method w/ GAT and w/ LM

The results of the proposed method with VinVL showed lower scores than with GAT on an image collection containing 6 images when evaluated with CIDEr, ROUGE, BERTScore, and WEEM4TS as shown in Table 1. However, when compared using the GAT model with ICC, the overall scores of the proposed method implemented with CG and without CG overcame the other methods except when evaluated by BLEU-2 and BLEU-4 where ICC achieved the better scores. In contrast, when increasing the number of images in an image collection, the scores of the proposed method with the VinVL model significantly rose essentially on the testing set of an image collection containing 11 or 16 images as shown in Tables 2 and 3. Following the results, only the MoverScore achieved the best scores for all image collection sizes, and for the testing set of an image collection containing 16 images, only Rouge-WE showed the best score with CG compared with the proposed method with the GAT model.

d: Proposed method w/ CG and w/o CG

Following the inspiration of abstractive text summarization in generalizing a generated caption, we report the experimental result of the proposed method with CG and without CG to show the effectiveness of the proposed method. The results that are generated without CG achieved better results on the traditional automatic evaluation metrics; BLEU, CIDEr and ROUGE, as shown in Tables 1, 2, 3. On the contrary, for the evaluation with abstractive text summarization metrics, the results showed that the proposed method achieved the best results on WEEM4TS on all image collection sizes. Also, when increasing the number of images in the image collection, the scores tend to increase on Rouge-WE and MoverScore, which evaluate the captions with the similarity scores based on word embedding. Furthermore, we also show the improvement in generating a caption of an image collection of the proposed

TABLE 2. Evaluation of the summarization results of image collections which contain eleven images in which all sentences are limited in length to twenty words, compared to ICC (our previous work) [10], SUPERT [35], T5 [34], and XL-Sum [37]. "w/ GAT" is the result generated by the Graph Attention (GAT) model. "w/ LM" is the result generated by the pre-trained vision-language model, "CG" is the *Sub-Graph Concept Generalization* component. "+" is the result generated by implementing the CG and "–" is the result generated without implementing the CG. "B-*n*" is BLEU [46], "C" is CIDEr [47], "R-*n*" is ROUGE Score [48], "R-WE" is ROUGE-WE [48], "BERT" is BERTScore [49], Mover is MoverScore [50], and WEE is WEEM4TS [51]. Results in bold indicate the highest scores and those underlined indicate the second highest scores.

Models	Captioning backbone	CG	B-2 ↑	B-4 ↑	C ↑	R-1 ↑	R-2 ↑	R-L ↑	R-WE ↑	BERT ↑	Mover ↑	WEE ↑
SUPERT [35]	w/ GAT	N/A	0.629	0.437	0.516	0.607	0.448	0.589	0.074	0.788	0.508	0.193
	w/ LM		0.775	0.524	0.526	0.643	0.471	0.616	0.080	0.851	0.568	0.179
T5 [34]	w/ GAT	N/A	0.794	0.569	0.587	0.730	0.550	0.706	0.075	0.927	0.560	0.147
	w/ LM		**0.835**	**0.581**	0.620	0.725	0.539	0.699	0.077	0.928	0.556	0.169
XL-SUM [37]	w/ GAT	N/A	0.279	0.107	0.100	0.366	0.152	0.326	0.041	0.870	0.517	0.194
	w/ LM		0.298	0.128	0.100	0.384	0.171	0.342	0.038	0.081	0.520	0.151
ICC [10]	w/ GAT	–	0.789	0.554	0.683	0.734	0.550	0.707	0.082	0.857	0.554	0.195
		+	0.757	0.509	0.608	0.716	0.522	0.689	0.080	0.852	0.561	0.195
Proposed (Scene-graph)	w/ GAT	–	0.802	0.578	**0.711**	**0.748**	**0.570**	**0.723**	**0.083**	**0.930**	0.564	**0.197**
		+	0.761	0.542	0.653	0.722	0.544	0.695	0.074	0.924	0.560	**0.197**
Proposed (Scene-graph)	w/ LM	–	0.712	0.449	0.524	0.655	0.458	0.623	0.079	0.903	**0.576**	0.192
		+	0.696	0.429	0.488	0.645	0.443	0.613	0.079	0.901	0.575	0.192

method compared to our previous work, ICC. The overall results of the proposed method achieved better results. Only when it was evaluated with BLEU-4 in the image collection comprising 6 images, ICC achieved the best results as shown in Table 1.

The overall results show that the proposed method with the GAT and VinVL models achieved better scores in generating a caption of an image collection compared with our previous work, ICC and baseline text summarization models. Moreover, the idea of generalizing a caption by finding concept words using word community improved the ability in captioning an image collection and achieved the best scores compared with other methods. However, we found the limitation in automatic evaluation metrics on a final caption that was refined by *Sub-Graph Concept Generalization (CG)*. Only the automatic evaluation metric that was implemented based on calculating word similarity in the proposed method with *Sub-Graph Concept Generalization* achieved better scores. Automatic evaluation metrics showed that the proposed method with the GAT and VinVL models achieved the best overall result compared with other text summarization baselines. However, we found that the result of the testing set of an image collection containing 11 or 16 images achieved good results when evaluated with BLEU.

F. QUALITATIVE RESULTS

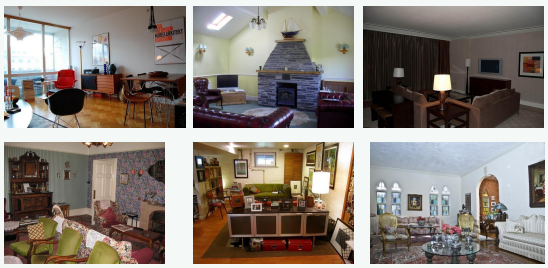
We show examples of the generated captions for three different image collection sizes with the proposed method in Fig. 7. We can see that the proposed method with GAT and VinVL models can describe the overall contexts of an image collection by grasping the common visual features and relationships. Overall results show the improvement of the proposed method after refining compared with our previous method, ICC. In addition, the results of the proposed method with GAT and VinVL models on image collections comprising 6 and 11 images show not much difference in generating a caption. However, for the image collection comprising 16 images, the generated caption of the proposed method with the VinVL model shows generating more accurate vocabularies compared with the proposed method with the GAT model and ICC.

Additionally, we demonstrate the comparison of the results between generating with *Sub-Graph Concept Generalization* (w/ CG) and without *Sub-Graph Concept Generalization* (w/o CG) in Fig. 8. The results show that *Sub-Graph Concept Generalization* can generalize specific words into concept words. However, the limitation of the proposed method with *Sub-Graph Concept Generalization* component (w/ CG) relies on the external knowledge that is provided in the proposed method. A generated caption might not be generalized if the concept words are not provided in the external knowledge and over-generalized words as shown in Figure 8(b). Moreover, over-generation results in a performance decrease when evaluated by text-based evaluation metrics, according to BLEU, etc., because it definitely increases performance compared to the generated ground truth.

V. CONCLUSION

We introduced a new challenging task to generate a fitting caption that represents an image collection. For this, the proposed method aimed to find a summarized scene graph of an image collection by combining each scene graph into a

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3332098



Results of image collection captioning [30]:

room filled with furniture and fire place.

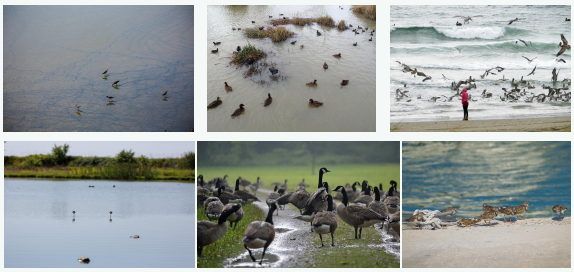
Results of the proposed method with GAT:

a living room filled with furniture and a fire place.

Results of the proposed method with VinVL:

a living room filled with furniture and lamps.

(a) Example 1 (6 imaages)



Results of image collection captioning [30]:

birds are in the water.

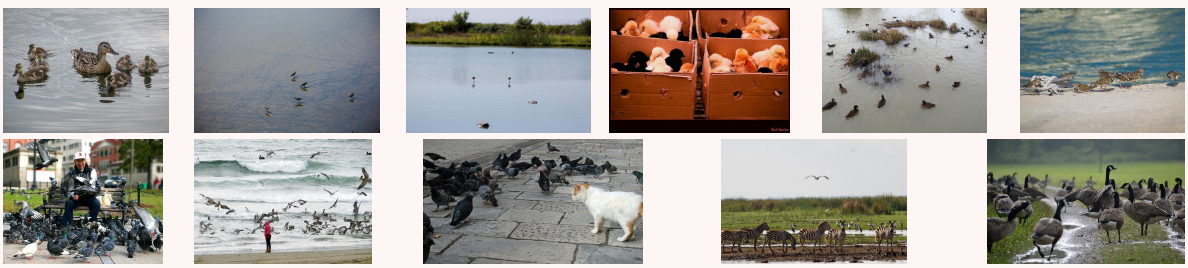
Results of the proposed method with GAT:

a couple of birds are in the water.

Results of the proposed method with VinVL:

a group of birds standing in the water.

(b) Example 2 (6 images)

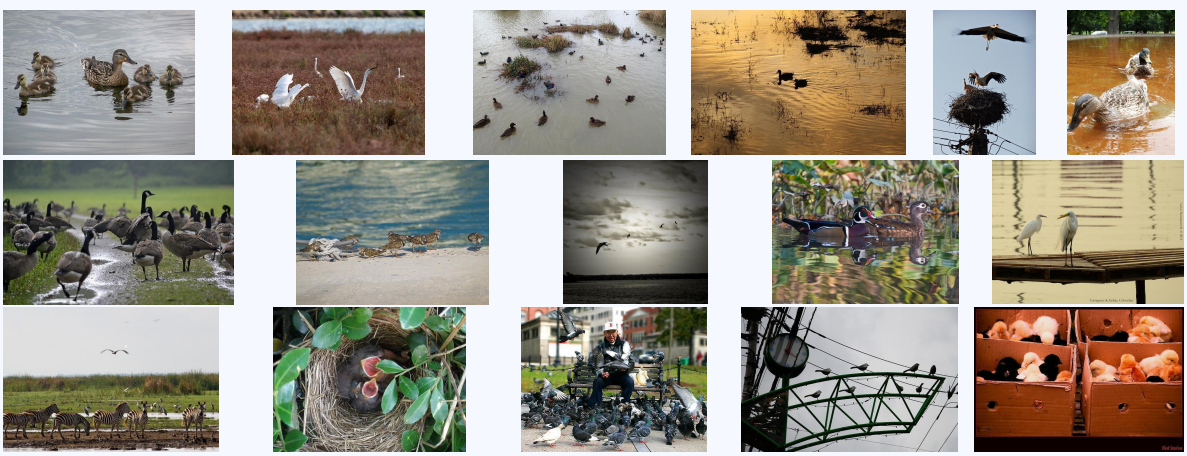


Results of image collection captioning [30]: ducks swimming in a lake.

Results of the proposed method with GAT: a group of ducks swimming in a lake.

Results of the proposed method with VinVL: a flock of ducks swimming on top of a lake.

(c) Example 3 (11 images)



Results of image collection caption [30]: birds flying over water.

Results of the proposed method with GAT: a flock of birds flying over a body of water.

Results of the proposed method with VinVL: a flock of ducks swimming on top of a lake.

(d) Example 4 (16 images)

FIGURE 7. **Examples of captions generated for different sizes of image collections on the *Graph Attention (GAT)* model [16] and the *Pre-trained Vision Language Model* (VinVL) [17].**

With Sub-Graph Concept Generalization

a bedroom with **furniture** and television

Without Sub-Graph Concept Generalization

a bedroom with **a bed desk** and television

(a) Example 1 (6 images)

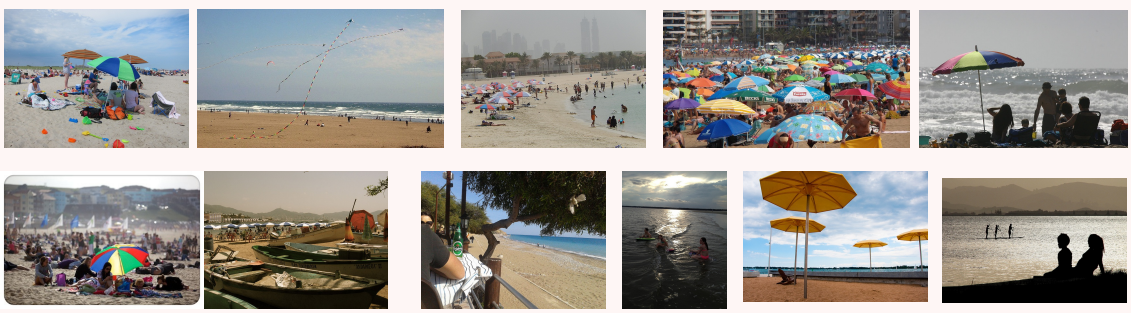
With Sub-Graph Concept Generalization

a herd of **animal** standing on top of a lush green field

Without Sub-Graph Concept Generalization

a herd of **sheep** standing on top of a lush green field

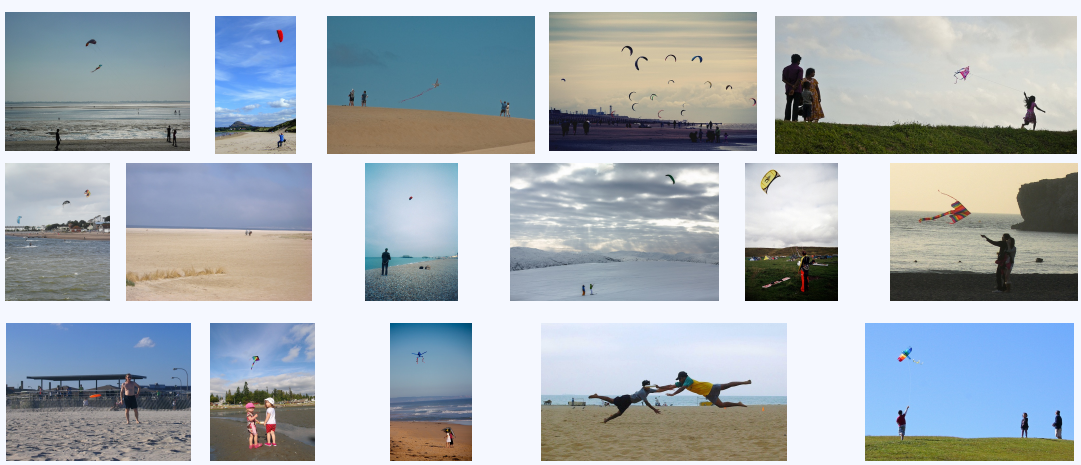
(b) Example 2 (6 images)



With Sub-Graph Concept Generalization **person** standing on a beach next to an umbrella

Without Sub-Graph Concept Generalization a man standing on a beach next to an umbrella

(c) Example 3 (11 images)



With Sub-Graph Concept Generalization **person** is flying a kite on the beach

Without Sub-Graph Concept Generalization a man is flying a kite on the beach

(d) Example 4 (16 images)

FIGURE 8. **Comparison of captions generated with and without *Sub-Graph Concept Generalization* component.**

TABLE 3. Evaluation of the summarization results of image collections which contain sixteen images in which all sentences are limited in length to twenty words, compared to ICC (our previous work) [10], SUPERT [35], T5 [34], and XL-Sum [37]. "w/ GAT" is the result generated by the Graph Attention (GAT) captioning backbone. "w/ LM" is the result generated by VinVL captioning backbone, "CG" is the *Sub-Graph Concept Generalization* component. "+" is the result generated by implementing the CG and "–" is the result generated without implementing the CG. "B-*n*" is BLEU [46], "C" is CIDEr [47], "R-*n*" is ROUGE Score [48], "R-WE" is ROUGE-WE [48], "BERT" is BERTScore [49], Mover is MoverScore [50], and WEE is WEEM4TS [51]. Results in bold indicate the highest scores and those underlined indicate the second highest scores.

Models	Captioning backbone	CG	B-2 ↑	B-4 ↑	C ↑	R-1 ↑	R-2 ↑	R-L ↑	R-WE ↑	BERT ↑	Mover ↑	WEE ↑
SUPERT [35]	w/ GAT	N/A	0.659	0.472	0.479	0.623	0.470	0.606	0.074	0.790	0.505	0.283
	w/ LM		0.800	0.559	0.471	0.655	0.491	0.630	0.080	0.858	0.561	0.255
T5 [34]	w/ GAT	N/A	0.822	0.606	0.588	0.748	0.578	0.726	0.077	0.929	0.556	0.280
	w/ LM		**0.870**	**0.625**	0.588	0.742	0.563	0.717	0.080	0.929	0.556	0.239
XL-SUM [37]	w/ GAT	N/A	0.294	0.120	0.062	0.367	0.162	0.325	0.039	0.868	0.514	0.202
	w/ LM		0.321	0.143	0.075	0.387	0.182	0.345	0.038	0.872	0.519	0.206
ICC [10]	w/ GAT	–	0.758	0.542	0.618	0.554	0.420	0.532	0.085	0.717	0.560	0.263
		+	0.726	0.495	0.541	0.539	0.396	0.518	0.082	0.712	0.556	0.263
Proposed (Scene-graph)	w/ GAT	–	0.832	0.614	**0.646**	**0.757**	**0.588**	**0.734**	0.085	**0.931**	0.560	0.283
		+	0.781	0.567	0.578	0.728	0.555	0.702	0.075	0.924	0.555	**0.285**
Proposed (Scene-graph)	w/ LM	–	0.761	0.498	0.527	0.692	0.499	0.656	**0.088**	0.921	**0.564**	0.273
		+	0.740	0.469	0.481	0.679	0.478	0.644	**0.088**	0.917	0.562	0.273

single scene graph and selecting a representative sub-graph, which is used to generate a caption by a captioning model. With inspiration from abstractive text summarization, we showed that building word communities using graph theory to generalize the final caption by finding concept words and refining the generated caption, contributes to describing the overall contexts of an image collection. To evaluate the proposed method of an image collection captioning task, we built an image collection dataset, annotated from the MS-COCO dataset, a popular captioning dataset. Experiments showed promising results for the image collection captioning task. The proposed method can be applied to other related tasks, such as image album summarization or video summarization. In the future, we plan to work on a more challenging dataset. Our project can be found at https://www.cs.is.i.nagoya-u.ac.jp/opensource/nu-icc/ [2].

ACKNOWLEDGMENT

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

REFERENCES

[1] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In *Proc. EMNLP2022*, pages 7241–7259, Abu Dhabi, United Arab Emirates, Dec. 2022.

[2]The dataset used in this paper is currently available. The sources and pre-trained models are planned to be made available upon acceptance of the paper.

[2] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Comput. Res. Reposit. arXiv Preprint*, arXiv:2301.12597, [Online], Jan. 2023.

[3] 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.

[4] Wenkai Zhang, Kun Fu, Xian Sun, Yuhang Zhang, Hao Sun, and Hongqi Wang. Joint optimisation convex-negative matrix factorisation for multi-modal image collection summarisation based on images and tags. *IET Comput. Vis.*, 13(2):125–130, May 2018.

[5] Andrea Pasini, Flavio Giobergia, Eliana Pastor, and Elena Baralis. Semantic image collection summarization with frequent subgraph mining. *IEEE Access*, 10:131747–131764, Dec. 2022.

[6] 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.

[7] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proc. CVPR2015*, pages 3668–3678, Boston, MA, USA, Jun. 2015.

[8] Brigit Schroeder and Subarna Tripathi. Structured query-based image retrieval using scene graphs. In *Proc. CVPR2020*, pages 178–179, [Online], Jun. 2020.

[9] Gencer Sumbul, Sonali Nayak, and Begüm Demir. SD-RSIC: Summarization-driven deep remote sensing image captioning. *IEEE Trans. Geosci. Remote Sens.*, 59:6922–6934, Oct. 2020.

[10] 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.

[11] 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.

[12] Hend Alrasheed. Word synonym relationships for text analysis: A graph-based approach. *PloS One*, 16(7):e0255127, Jul. 2021.

[13] Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. Bottom-up abstractive summarization. In *Proc. EMNLP2018*, pages 4098–4109, Brussels, Belgium, Oct.–Nov. 2018.

[14] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model

pre-training for natural language understanding and generation. *Adv. Neural Inf. Process. Syst.*, 32:13063–13075, Dec. 2019.

[15] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proc. ICML2020*, pages 11328–11339, [Online], Jul. 2020.

[16] Victor Milewski, Marie-Francine Moens, and Iacer Calixto. Are scene graphs good enough to improve image captioning? In *Proc. ACL-IJCNLP2020*, Suzhou, Jiangsu, China, Sep. 2020.

[17] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. VinVL: Revisiting visual representations in vision-language models. In *Proc. CVPR2021*, pages 5579–5588, Nashville, TN, USA, Jun. 2021.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Kien Nguyen, Subarna Tripathi, Bang Du, Tanaya Guha, and Truong Q Nguyen. In defense of scene graphs for image captioning. In *Proc. ICCV2021*, pages 1407–1416, Montreal, QC, Canada, Oct. 2021.

[24] 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.

[25] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEiT pretraining for all vision and vision-language tasks. *Compt. Res. Reposit. arXiv Preprint*, arXiv:2208.10442, [Online], Aug. 2022.

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S Bernstein, and Li Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, May 2017.

[27] 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.

[28] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network Motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[29] Alex Graves. *Long Short-Term Memory*, pages 37–45. Springer Berlin Heidelberg, Berlin, Heidelberg, Jan. 2012.

[30] 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.

[31] Lin Cheng and Zijiang 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.

[32] 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.

[33] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. RelTR: Relation transformer for scene graph generation. *IEEE Trans. Pattern Anal. Mach. Intell.*, Apr. 2023.

[34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, Jan. 2020.

[35] Yang Gao, Wei Zhao, and Steffen Eger. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proc. ACL2023*, pages 1347–1354, [Online], 2020.

[36] Som Gupta and Sanjai Kumar Gupta. Abstractive summarization: An overview of the state of the art. *Expert Syst. Appl.*, 121:49–65, May 2019.

[37] Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. XL-Sum: Large-scale multilingual abstractive summarization for 44 languages. In *Proc. ACL-IJCNLP2021*, pages 4693–4703, [Online], Jun. 2021.

[38] 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.

[39] 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.

[40] Ulrik Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Netw.*, 30(2):136–145, May 2008.

[41] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global vectors for word representation. In *Proc. EMNLP2014*, pages 1532–1543, Doha, Qatar, Oct. 2014.

[42] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*, volume 8. Cambridge University Press, Cambridge, England, UK, 1994.

[43] Edward Loper and Steven Bird. NLTK: The natural language toolkit. In *Proc. ACL2002*, volume 1, pages 63–70, Barcelona, Catalunya, Spain, May 2002.

[44] 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.

[45] 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.

[46] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proc. ACL2002*, pages 311–318, Philadelphia, PA, USA, Jul. 2002.

[47] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proc. CVPR2015*, pages 4566–4575, Boston, MA, USA, Jun. 2015.

[48] Jun-Ping Ng and Viktoria Abrecht. Better summarization evaluation with word embeddings for ROUGE. In *Proc. EMNLP2015*, pages 1925–1930, Lisbon, Portugal, Sep. 2015.

[49] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *Proc. ICLR2020*, [Online], Apr. 2020.

[50] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. MoverScore: Text generation evaluating with contextualized embeddings and Earth Mover Distance. In *Proc. EMNLP2019*, pages 563–578, Hong Kong, China, Nov. 2019.

[51] Tulu Tilahun Hailu, Junqing Yu, and Tessfu Geteye Fantaye. A framework for word embedding based automatic text summarization and evaluation. *Information*, 11(2):78–100, Jan. 2020.

[52] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. ICLR2014*, Banff, AB, Canada, Apr. 2014.

[53] 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.

[54] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proc. CVPR2017*, pages 936–944, Honolulu, HI, USA, Jun. 2017.

[55] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Proc. ACL2004*, pages 74–81, Barcelona, Catalunya, Spain, Jul. 2004.

[56] 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*, pages 4171–4186, Minneapolis, MN, USA, June 2019.

[57] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The Earth Mover's Distance as a metric for image retrieval. *Int. J. Comput. Vis.*, 40:99–121, 2000.

[58] Mahmood Yousefi-Azar and Len Hamey. Text summarization using unsupervised deep learning. *Expert Sys. Appl.*, 68:93–105, Feb. 2017.

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 summarization and image captioning.

MARC A. KASTNER (M'22) received his BSc and MSc in Computer Science from Braunschweig University of Technology in Braunschweig, 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 with multimedia, covering vision and language and affective computing related tasks. Dr. Kastner is a member of IEICE, IPS Japan, and ACM.

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 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.

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. His research interests include database, data analysis, multimedia data managements, and Linked Open Data. He is a member of ACM, IPS Japan, IEICE, DBSJ, NLP, and JSAI.

ICHIRO IDE (M'12–SM'21) received his BEng, MEng, and PhD from The University of Tokyo in 1994, 1996, and 2000, respectively. He became an Assistant Professor at the National Institute of Informatics, Japan in 2000, and an Associate Professor at Nagoya University, Japan in 2004. Since 2020, he has been a Professor there. He was a Visiting Associate Professor at National Institute of Informatics from 2004 to 2010, an Invited Professor at Institut de Recherche en Informartique 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.

Dr. Ide is a senior member of IEICE and IPS Japan, and a member of ACM, JSAI, and ITE.