Fig. 2. Overview of the proposed method, consisting of ve components: (a) Scene Graph Generation extracts a scene graph for each image. (b) Multiple-Scene Graph Processing combines all scene graphs and nds a representative graph. (c) Sub-Graph Concept Generalization nds word communities from all scene graphs and generates a common word. (d) Captioning Model generates the initial caption for the representative graph. (e) Sentence Re nement rephrases the caption with the common words.

select part of the combined graph, and Sub-Graph Concept Generalization to nd general concept words through the word communities. Then, the Captioning Model generates a sentence based on the representative scene graph. The generated sentence from the captioning model and the community word graphs are nally passed to the Sentence Re nement to output the nal caption.

3.1 Scene Graph Generation

Following the existing work on image captioning model leveraging scene graphs, we use the current state-of-the-art scene graph generation method with ResNet101 [11] + Neural Motif [29] as a scene graph parser in the proposed method. This model is retrained on the Visual Genome dataset [15], a popular practice for scene graph captioning. A recent image captioning work [2] shows that manually cleaning up duplicate labels of the Visual Genome dataset from 2,500/1,000/500 to 1,600/400/20 of objects/attributes/relations can improve image captioning performance. We also follow this idea. The result of scene graph generation is represented in a directed graph, which includes subjects, predicates, and objects.

3.2 Multiple-Scene Graph Processing

We build the multiple-scene graph processing module as a feature selection of all sub-graphs from the Scene Graph Generation Component. All the sub-graphs are merged into a single directed graph, as shown in Eq. 1, in which G is the merged graph and g_i is a directed graph represented as a set of triplets. In the merging process, we count the occurrence of each feature and the number of edges to be the weight in the selection step.

$$G = \bigcup_{i=1}^{n} g_i \tag{1}$$

Fig. 3. Example of constructing word communities. (a) Concept Expansion expands word concepts of each word in sub-graphs by incorporating with ConceptNet [23]. (b) Concept Joining joins all the same words together and selects the central word node as the representative.

3.4 Captioning Model

The captioning model consists of a Graph Convolutional Network (GCN) [19] and the Attention-based LSTM model [2]. We build the GCN to process the triplet of subject, predicate, and object features. Each feature is extracted from the scene graph generation process, which consists of 36 subject features, 36 object features, and 100 predicate features. Each feature dimension is 1,024, which is the feature size of theScene Graph Generationoutput. Next, we build the attention-based LSTM model following the top-down LSTM captioning with two layers of attention-based LSTM, both layers with a size of 512.

3.5 Sentence Re nement

To improve the caption, we modify the beam search of sentence generation to generalize the caption, mainly focusing on general noun words of the caption result. First, word tokens are extracted from the caption and labeled with NLTK POS tagging [18]. Next, a noun phrase is found and labeled with its object component. Finally, the object component of each noun phrase is mapped with the Sub-Graph Concept Generation Componentto replace the word with the representative of the word community in the sub-graph concept. In the following experiment, we select a beam size of ve for generating the nal caption.

4 Dataset Construction

The proposed method aims to generate a caption for an image collection, i.e., a sentence describing multiple images. Due to this task being novel, there is no existing dataset for this. The MS-COCO dataset [17] is a popular image captioning dataset which is closest to this task. However, it is typically used only for the single-image captioning task in which each image is captioned with one or more sentences.

In our work, we build upon the MS-COCO dataset by estimating the semantic contents of images and captions and use this to augment the dataset towards

Fig. 5. Examples of the proposed method for image collections built with the image-caption retrieval approach. The results show that the proposed method can extract important features from the image collection and normalize the word concepts in captioning results. Images with a border indicate that they fit the image collection caption, while those without indicate outliers.

collections, in Fig. 4. It shows that the proposed method detects the most frequently occurring contents, ignoring less appearing contents. Then, it generates a general caption to describe most image contents in an image collection. The evaluation results are shown in Table 1, which shows that the proposed method achieved the best result on overall automatic evaluations.

Image-Caption Retrieval Dataset We next show results on the dataset constructed with the image-caption retrieval approach in Fig. 5. It shows that the prediction relates to the most frequent content in the image collection, and unrelated contents are ignored. They also keep the main specific content if it can describe the image collection. Table 2 shows the evaluation results, which show that the proposed method beats text summarization methods in this novel task.

The experiments above show a novel image collection captioning baseline compared with text summarization methods. However, we also found a limitation of automatic evaluation metrics when the captioning results are refined. The proposed method without refining the captioning results, achieved better evaluation scores in ROUGE, BERTScore, and CIDerBtw, due to the nature of these metrics focusing on text similarity. However, WEEM4TS is a novel metric for abstractive summarization evaluation and shows a promising direction for the image collection captioning task being more suitable for our task of generalizing across images.

Table 2. Evaluation of the results of image collections built with the image-caption retrieval approach compared to SUPERT [5], T5 [21], and XL-Sum [10].

Method		R-1 ↑	R-2 ↑	R-L ↑	BERTScore ↑	CIDErBtw ↑	WEEM4TS ↑
SUPERT [5]		0.3756	0.1105	0.3231	0.6166	0.7016	0.1083
T5 [21]		0.3441	0.1037	0.3025	0.6057	0.5524	0.1031
XL-Sum [10]		0.2148	0.0367	0.1833	0.4678	0.1023	0.0860
Proposed	w/o CG	0.3782	0.1265	0.3409	0.6270	0.7955	0.1062
	w/ CG	0.3517	0.1105	0.3140	0.6093	0.7156	0.1096

6 Conclusion

We introduced a challenging new task which aims to produce a single tting caption for a collection of images. The key idea was to generate a scene graph for each image in the collection, combine them to generate a generalized combined representation, and then generate a caption. The proposed method showed potential for transferring a single-image captioning model to image collection captioning. Inspired by text summarization methods generating a summary, the proposed method improves the abstractiveness of the summarized image collection caption by nding generalized words using graph theory and word communities. We additionally introduced the prospect of using an augmented version of the MS-COCO dataset, a popular image captioning dataset, in the image collection captioning task. The results are promising and pioneering steps toward captioning an image collection with a shared description. In the future, we plan to work also on a more challenging dataset and also improve the captioning model focusing on estimating the overall semantic context of an image collection incorporating external knowledge. Our project can be found at https://www.cs.is.i.nagoya-u.ac.jp/opensource/nu-icc/

Acknowledgements Parts of this work were supported by JSPS Grant-in-aid for Scienti c Research (21H03519) and a joint research project with National Institute of Informatics.

References

1. Alrasheed, H.: Word synonym relationships for text analysis: A graph-based approach. PloS One 16(7), e0255127 (2021)
2. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: 2018 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 6077 6086 (2018)
3. Deng, J., et al.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Comput. Society Conf. Comput. Vis. Pattern Recognit. pp. 248 255 (2009)
4. Faghri, F., et al.: VSE++: Improving visual-semantic embeddings with hard negatives. In: 29th Brit. Mach. Vis. Conf. (2018)
5. Gao, Y., et al.: SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In: 58th Annu. Meeting of the Assoc. for Computat. Linguist. pp. 1347 1354 (2020)
6. Girshick, R.: Fast R-CNN. In: 16th IEEE Int. Conf. Comput. Vis. pp. 1440 1448 (2015)
7. Gupta, S., et al.: Abstractive summarization: An overview of the state of the art. Expert Syst. Appl. 121, 49 65 (2019)
8. Hailu, T.T., et al.: A framework for word embedding based automatic text summarization and evaluation. Information 11(2), 78 100 (2020)
9. Han, X., et al.: Image scene graph generation (SGG) benchmark. Comput. Res. Reposit. arXiv preprint arXiv:2107.12604 (2021)
10. Hasan, T., et al.: XL-Sum: Large-scale multilingual abstractive summarization for 44 languages. In: Findings Assoc. Comput. Linguist.: ACL-IJCNLP 2021. pp. 4693 4703 (2021)