IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received September 2, 2019, accepted October 9, 2019, date of publication October 14, 2019, date of current version October 31, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2947409

Query is GAN: Scene Retrieval With Attentional Text-to-Image Generative Adversarial Network

RINTARO YANAGI[1], (Student Member, IEEE), REN TOGO[2], (Member, IEEE), TAKAHIRO OGAWA[2], (Senior Member, IEEE), AND MIKI HASEYAMA[2], (Senior Member, IEEE)

[1]Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan
[2]Faculty of Information Science and Technology Division of Media and Network Technologies, Hokkaido University, Sapporo 060-0814, Japan

Corresponding author: Rintaro Yanagi (yanagi@lmd.ist.hokudai.ac.jp)

This work was supported in part by the MIC/SCOPE under Grant #181601001.

ABSTRACT Scene retrieval from input descriptions has been one of the most important applications with the increasing number of videos on the Web. However, this is still a challenging task since semantic gaps between features of texts and videos exist. In this paper, we try to solve this problem by utilizing a text-to-image Generative Adversarial Network (GAN), which has become one of the most attractive research topics in recent years. The text-to-image GAN is a deep learning model that can generate images from their corresponding descriptions. We propose a new retrieval framework, "*Query is GAN*", based on the text-to-image GAN that drastically improves scene retrieval performance by simple procedures. Our novel idea makes use of images generated by the text-to-image GAN as queries for the scene retrieval task. In addition, unlike many studies on text-to-image GANs that mainly focused on the generation of high-quality images, we reveal that the generated images have reasonable visual features suitable for the queries even though they are not visually pleasant. We show the effectiveness of the proposed framework through experimental evaluation in which scene retrieval is performed from real video datasets.

INDEX TERMS Scene retrieval, deep learning, generative adversarial network, text-to-image translation.

I. INTRODUCTION

With the increasing number of videos on the Web, methods of retrieval that provide users with scenes[1] corresponding to their descriptions have become important topics of study [1]–[5]. The scene retrieval task has been studied by many researchers, and there have been many reports that proposes text-based retrieval methods [6] and content-based retrieval methods [7]. With the rapid growth of deep learning technologies, studies on scene retrieval have moved to the next stage.

Realization of scene retrieval is difficult because several important challenges must be tackled simultaneously. First, videos and their corresponding descriptions are denoted as modalities that have different semantic spaces. Thus, it is necessary to match these two different modalities to retrieve scenes that are relevant to input text descriptions. The retrieval performance heavily depends on the matching accuracy of the two different modalities [8]–[10]. Moreover, descriptions of desired scenes are different for each user, and it is a difficult task to handle all of these descriptions. Therefore, it is essential for successful retrieval to learn high-level and robust feature representations of scenes and their corresponding descriptions [11], [12]. The above-described challenges were tackled in pioneering studies, and text-based and content-based retrieval methods have been widely adopted. Text-based methods [13]–[19], which perform annotations for target contents, retrieve contents by comparing input descriptions provided by users and the results of annotations given to the candidate contents. However, their retrieval performance depends on the quality of the annotations. Also, for realizing retrieval of new contents, it is necessary to prepare an enormous amount of training annotation data. Content-based methods [8]–[10], [20]–[23], which retrieve contents by computing similarities in content spaces, have recently attracted attention with the development of deep

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

[1] In this paper, we define a unit including continuous shots of the same time, the same place and the same action as a "scene" and denote images included in videos as "frames ".

learning techniques [24], [25]. Since content-based methods do not rely on annotated information but directly use content information, they tend to overcome the above-mentioned problems [26], [27]. However, input query contents must be provided to perform the retrieval in the content-based methods. Therefore, these methods cannot retrieve desired contents when users cannot prepare the query contents, and this restriction is not user friendly.

In this paper, we propose a new scene retrieval framework, *Query is GAN*, based on a text-to-image Generative Adversarial Network (GAN) [28]–[35] . As shown in Fig. 1, the proposed framework enables scene retrieval from input descriptions, *i.e.*, sentences, provided by users. The input descriptions are projected to the visual space through a multimodal neural network model, *i.e.*, the text-to-image GAN. Specifically, query images are generated from the input descriptions based on AttnGAN [35]. Then, by retrieving similar scenes based on their visual features, the proposed framework overcomes the remaining problems of existing methods.

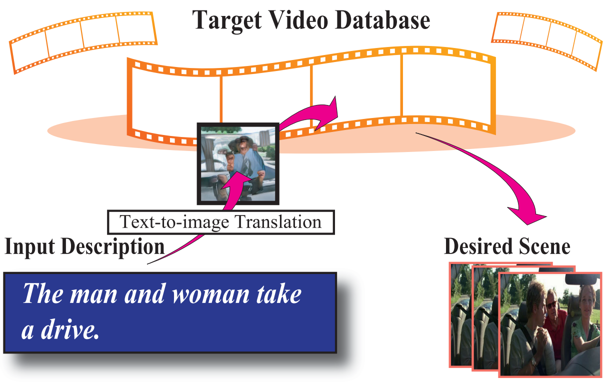


FIGURE 1. **Illustration of the scenario that we try to achieve. From the input description, our framework retrieves the corresponding desired scene.**

Our retrieval framework makes use of the hierarchical structure of AttnGAN. Specifically, in AttnGAN, higher-resolution images are recursively generated from their low-resolution versions. As the resolution becomes higher, attention is paid to words in descriptions. By focusing on the relationship between this hierarchical structure and the characteristic of the attention, retrieval that can gradually narrow down the retrieval candidates along the coarse-to-fine abstraction level direction can be realized.

In addition to the proposition of the above novel retrieval framework, new interesting results are also shown in this paper. It has been reported that the visual quality of images generated by GANs are still insufficient when the generation tasks become complicated [35]. For example, Fig. 2 (a) shows an image generated by AttnGAN trained on a bird dataset [36], and Fig. 2 (b) shows an image generated by AttnGAN trained on the Common Objects in Context (COCO) dataset [37]. Note that the COCO dataset is a large-scale captioning dataset that contains images of common scenes. Although AttnGAN trained on the COCO dataset can generate various images not limited to specific objects, its generation task becomes more complicated and difficult than that of the bird dataset. We can see that the generated image based on COCO dataset is not visually pleasant. However, in this study, we reveal that such visually non-pleasant images generated by the text-to-image GAN have reasonable visual features suitable for the queries, evidence of which was obtained in the experiments. Therefore, successful retrieval becomes feasible regardless of the visual quality of the generated images.

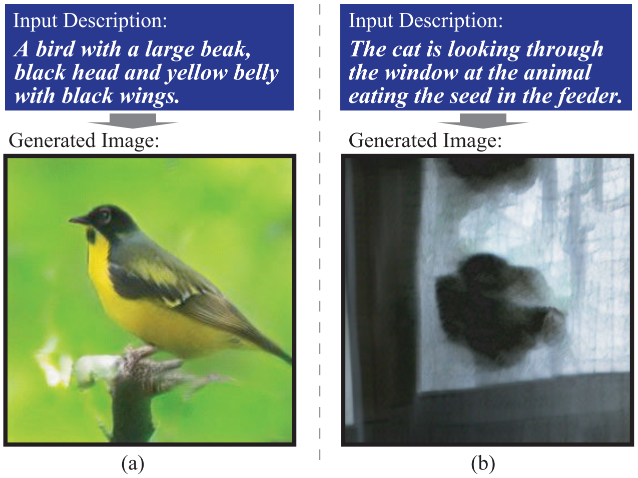

Input Description:
A bird with a large beak, black head and yellow belly with black wings.
Generated Image:
(a)
Input Description:
The cat is looking through the window at the animal eating the seed in the feeder.
Generated Image:
(b)

FIGURE 2. **Examples of generated images by different datasets: (a) generated image by AttnGAN trained on the bird dataset [36], (b) generated image by AttnGAN trained on COCO dataset [37].**

The contributions of this paper are summarized below.

Contribution 1

We propose a novel scene retrieval framework and achieve higher retrieval performance than the performances of state-of-the-art methods. The proposed framework utilizes multiple-resolution generated images that can pay attention to sentence-to-word characteristics based on the hierarchical structure of the text-to-image GAN.

Contribution 2

We demonstrate the usefulness and versatility of the generated images even if the images are not visually pleasant. By showing the effectiveness of the proposed retrieval framework through experimental evaluation using several input descriptions and their corresponding generated query images, the above characteristics are confirmed.

II. RELATED WORKS

A. MULTIMODAL RETRIEVAL USING DESCRIPTION QUERIES

A multimodal retrieval task from input descriptions can be considered as a matching task between contents included in a target database and the provided descriptions. This task can be broadly classified into two streams.

In the first stream, methods that prepare text labels for images included in a target database are widely utilized [15], [16], [38], [39], and they compare input query descriptions and the annotated text labels in the target database. In recent years, deep learning-based annotation has been used to automatically add labels to images in a database. Karpathy and Li [38] utilized a Convolutional Neural Network (CNN) and a bidirectional Recurrent Neural Network (bRNN) [40] to add labels to corresponding regions of the images. This enables retrieval of images even if they contain many and various objects. Vinyals *et al.* [39] utilized Long Short-Term Memory (LSTM) to estimate long descriptions by predicting the next words from their visual features and the preceding words. This model can generate natural text descriptions that express how objects included in the target images are related to each other. Then they realized retrieval that can consider the relationship between these objects. As a similar approach, Johnson *et al.* [15] generated graph-type text labels and realized retrieval that takes into account the relationship between objects.

As the second stream, there have been many proposed methods that retrieve images from input descriptions by embedding their text features into common semantic spaces with visual features for realizing their comparison [8], [10], [41]–[47]. These methods enable retrieval that does require text label annotation, which is necessary in the first stream. Kiros *et al.* [8] embedded visual and text features into a common semantic space on the basis of CNN and LSTM and enabled retrieval of relevant images. Vendrov *et al.* [10] proposed an embedding method that takes the order relationship between words into consideration. This method can retrieve images that are strongly related to structures of input descriptions based on CNN and Gated Recurrent Unit (GRU) even though they do not use text labels. Since the above-mentioned methods embed

input descriptions and images into common semantic spaces, they have robustness to input descriptions, and our framework also adopts an approach similar to these methods.

B. APPLICATIONS OF GENERATIVE ADVERSARIAL NETWORKS

The reality of images generated by GANs has been drastically improved in recent years. Accordingly, practical studies using the architecture of GANs have become popular. GAN models have been applied to image translation tasks [48], [49] in many studies, and many related works such as works on image super-resolution [50] and image inpainting [51] have been carried out. Studies not only on the generation of images but also other kinds of contents such as paintings [52], music pieces [53], computer graphics [54] and graphs [55] have also been increasing.

Text-to-image synthesis is one of the most attractive fields for application of GANs [29]–[35]. GAN-INT-CLS [29] is the first model in which concept of GAN was applied to a text-to-image translation task. Although this model can generate images from input texts, its resolution is limited to 64 × 64 pixels, and the generated images are not visually pleasant. AttnGAN [35] and HDGAN [32] have recently been proposed for improving the quality of generated images. By utilizing description information and its word information, AttnGAN can generate high-resolution images that can focus on details of the input description information. On the other hand, from its hierarchically-nested structure, HDGAN can generate images with higher resolution than that of images generated by AttnGAN. Even though these methods can generate visually pleasant images in simple tasks such as birds shown in Fig. 2 (a), it is still difficult to generate visually pleasant images in more complicated tasks as shown in Fig. 2 (b).

In the proposed framework, we utilize AttnGAN for a text-to-image synthesise task since its structure focusing on not only description information but also word information strongly matches the aim of our study.

III. OUR SCENE RETRIEVAL FRAMEWORK

The details of our proposed framework are presented in this section. Our goal is to retrieve scenes that contain particular semantic contents corresponding to input sentences and their words. An overview of our framework is shown in Fig. 3. Our framework consists of two phases, query image generation and estimation of relevant scenes. In the first phase, a hierarchical image generation network based on AttnGAN [35] is constructed and three different resolution query images that contain different abstraction levels are generated. In the second phase, a hierarchical retrieval architecture is constructed to find the most suitable scenes from the target video scene database by narrowing down the retrieval candidates along the coarse-to-fine abstraction level direction.

A. FIRST PHASE: QUERY IMAGE GENERATION

In the first phase, three different resolution query images are generated. By utilizing these generated images as queries, retrieval that takes into consideration the sentence structure is realized. To generate query images, a hierarchical network based on AttnGAN is constructed. AttnGAN consists of three generators, G_r ($r \in \{l, m, h\}$; l, m and h, representing resolutions, *i.e.*, low, middle and high, respectively), which take three hidden states $\boldsymbol{s}_r$ ($r \in \{l, m, h\}$) as inputs calculated by three neural networks F_r ($r \in \{l, m, h\}$) and generate three different resolution query images Q_r ($r \in \{l, m, h\}$).

First, we define a sentence feature vector and a word feature matrix extracted from an input sentence as $\boldsymbol{e}_{sen} \in \mathbb{R}^{D_{sen}}$ and $\boldsymbol{E}_{word} \in \mathbb{R}^{D_{word} \times T}$ [35]. D_{sen} and D_{word} denote the dimension of the extracted sentence features and that of the extracted word features, respectively, T denotes the number of words included in the input sentence. The features $\boldsymbol{e}_{sen}$ and $\boldsymbol{E}_{word}$ are calculated by a "sentence feature extractor that strongly focuses on the word relationship" and a "word attribute and feature extractor that strongly focuses on the detailed words", respectively.

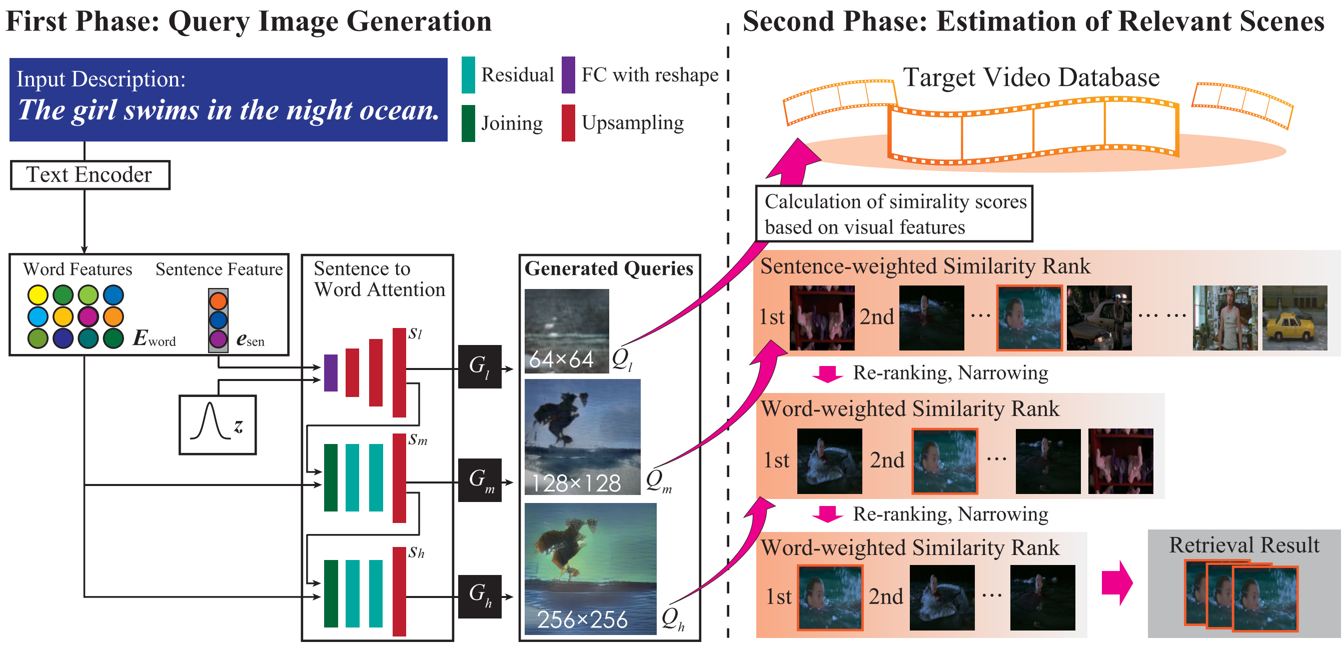


FIGURE 3. Overview of our scene retrieval framework. The proposed framework consists of two phases, and details of them are respectively explained in III-A and III-B respectively.

We obtain the hidden state $\boldsymbol{s}_l$ from the feature vector $\boldsymbol{e}_{sen}$ and the Gaussian noise z as follows:

$$s_l = F_l(z, F^{\mathrm{ca}}(\boldsymbol{e}_{sen})), \tag{1}$$

where F^{ca} is a function that stabilizes the training [34]. Specifically, it translates discontinuous features $\boldsymbol{e}_{sen}$ to continuous features by sampling $\boldsymbol{e}_{sen}$ on a normal distribution. Next, we calculate the hidden states $\boldsymbol{s}_m$ from the feature matrix $\boldsymbol{E}_{word}$ and the previously obtained hidden state $\boldsymbol{s}_l$ in Eq. (1). Then $\boldsymbol{s}_m$ becomes a feature vector that contains information on the sentence features and weakly contains information on the word features $\boldsymbol{E}_{word}$. Similarly, we obtain the hidden state $\boldsymbol{s}_h$ from the feature matrix $\boldsymbol{E}_{word}$ and the previous hidden state $\boldsymbol{s}_m$. Thus, $\boldsymbol{s}_h$ becomes a feature vector that contains information on both the sentence and word features. The relationship of $\boldsymbol{s}_l$, $\boldsymbol{s}_m$ and $\boldsymbol{s}_h$ can be calculated as follows:

$$\boldsymbol{s}_m = F_m(\boldsymbol{s}_l, F_m^{\mathrm{attn}}(\boldsymbol{E}_{word}, \boldsymbol{s}_l)), \tag{2}$$

$$\boldsymbol{s}_h = F_h(\boldsymbol{s}_m, F_h^{\mathrm{attn}}(\boldsymbol{E}_{word}, \boldsymbol{s}_m)), \tag{3}$$

where F_r^{attn} $(r \in \{m, h\})$ is a function that adds the word features $\boldsymbol{E}_{word}$ to the previous hidden states $\boldsymbol{s}_r$ $(r \in \{l, m\})$. AttnGAN can generate an image that focuses on each word of the input sentence through this function.

Finally, we generate the multiple resolution query image Q_r $(r \in \{l, m, h\})$ from each hidden state $\boldsymbol{s}_r$ as follows:

$$Q_r = G_r(\boldsymbol{s}_r) \quad (r \in l, m, h), \tag{4}$$

In the proposed framework, we utilize these three generated query images Q_l, Q_m and Q_h to retrieve relevant scenes in the following phase.

Here, we explain how to train the above hierarchical image generation network. To generate images that contain the content of the input sentence, the final objective function is defined as follows:

$$L = L_G + \lambda L_{DAMSM}, \tag{5}$$

where λ is a hyperparameter that balances L_G and L_{DAMSM}. In this final objective function, L_G is a loss function that approximates conditional and unconditional distributions, and L_{DAMSM} is a fine-grained image-text matching loss at the word level. In more detail, L_G in Eq. (5) is defined as follows:

$$L_G = L_{G_l} + L_{G_m} + L_{G_h}. \tag{6}$$

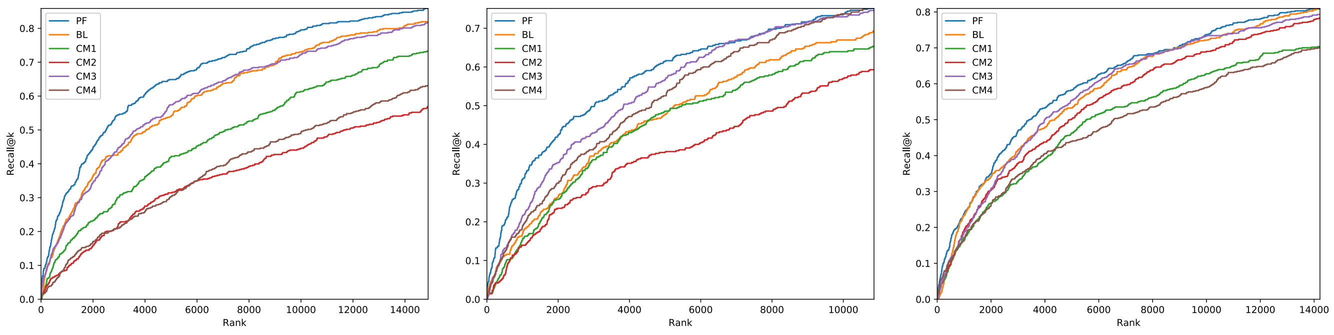
Each L_{G_r} $(r \in l, m, h)$ in Eq. (6) is defined as follows:

$$\begin{aligned} L_{G_r} = &-\frac{1}{2}\mathbb{E}_{Q_r \sim p_{G_r}}[log(D_r(Q_r))] \\ &-\frac{1}{2}\mathbb{E}_{Q_r \sim p_{G_r}}[log(D_r(Q_r, \boldsymbol{e}_{sen}))], \end{aligned} \tag{7}$$

where Q_r is from the generation model distribution p_{G_r} at scale r. In Eq. (7), the first term determines whether the image is real or fake, while the second term determines whether the image and the sentence match or not. Next, the second term of Eq. (5), L_{DAMSM}, is calculated by a Deep Attentional Multimodal Similarity Model (DAMSM) described in [35]. For a batch of B image-sentence pairs, L_{DAMSM} is defined as follows:

$$L_{DAMSM} = -\sum_{i=1}^{B} logP_{word_1}^{i} - \sum_{i=1}^{B} logP_{word_2}^{i} \tag{8}$$

$$-\sum_{i=1}^{B} logP_{sen_1}^{i} - \sum_{i=1}^{B} logP_{sen_2}^{i}, \tag{9}$$



(a) (b) (c)

FIGURE 4. Recall@*k* obtained in the quantitative evaluation for each movie. These results represent the proportion of the scenes relevant to the input sentence at rank of *k*. The horizontal axis represents the rank of frames, and the vertical axis represents Recall@*k* defined in Eq. (12). A higher value indicates a better result. (a), (b) and (c) respectively represent the results of "Bad Santa", "As Good As it Gets" and "Harry Potter and the Prisoner of azkaban".

where $P^i_{word_1}$ is a posterior probability that measures how the generated images are matched with their corresponding text descriptions at the word level, and $P^i_{word_2}$ is a posterior probability that measures how the sentences are matched with their corresponding generated images at the word level. $P^i_{sen_1}$ and $P^i_{sen_2}$, which are similar to $P^i_{word_1}$ and $P^i_{word_2}$, are posterior probabilities at the sentence level.

At scale r, the generator G_r has a corresponding discriminator D_r. Alternately to the training of G_r, each discriminator D_r is trained to classify whether the input image is real or fake by minimizing the following loss:

$$\begin{aligned} L_{D_r} = &-\frac{1}{2}\mathbb{E}_{\hat{Q}_r\sim p_{data_r}}[log(D_r(\hat{Q}_r))] \\ &-\frac{1}{2}\mathbb{E}_{Q_r\sim p_{G_r}}[log(1-D_r(Q_r))] \\ &-\frac{1}{2}\mathbb{E}_{\hat{Q}_r\sim p_{data_r}}[log(D_r(\hat{Q}_r, \boldsymbol{e}_{sen}))] \\ &-\frac{1}{2}\mathbb{E}_{Q_r\sim p_{G_r}}[log(1-D_r(Q_r, \boldsymbol{e}_{sen}))], \end{aligned} \tag{10}$$

where $\hat{Q}_r$ is from the true image distribution p_{data_r} at scale r.

B. SECOND PHASE: ESTIMATION OF RELEVANT SCENES

In the second phase, scene retrieval is performed by using the three generated query images Q_r $(r \in \{l, m, h\})$. First, we define candidate frames as f_{n_l} $(n_l = 1, 2, \ldots, N; N$ being the number of frames included in all candidate scenes, *i.e.*, retrieval targets) and calculate the visual features $\boldsymbol{v}_l$ and $\boldsymbol{v}_{n_l}$ from Q_l and f_{n_l}. In the proposed framework, we utilize outputs of the third pooling layer of Inception-v3 [56] pre-trained on ImageNet [57] as the visual features. We utilize Inception-v3 since the loss function L_{DAMSM} in the generation network utilizes Inception-v3 as the image feature extractor for calculating the image-text matching loss.

Then we simply calculate the following cosine similarities w_{n_l} between $\boldsymbol{v}_l$ and $\boldsymbol{v}_{n_l}$:

$$w_{n_l} = \frac{\boldsymbol{v}_l \cdot \boldsymbol{v}_{n_l}}{|\boldsymbol{v}_l||\boldsymbol{v}_{n_l}|} \quad (n_l = 1, 2, \ldots, N). \tag{11}$$

This value indicates the similarity between the query image Q_l and the retrieval candidate frames f_{n_l} $(n_l = 1, 2, \ldots, N)$. From the obtained similarities, we can calculate the rankings of the candidate frames. As described above, the low-resolution query image Q_l focuses on the whole information of the input sentence. Therefore, Q_l has the role of screening of large-scale retrieval candidates.

Next, we select the frames that are included in the top $100P_m$ percent of the retrieval candidates. In the same manner as Eq. (11), we respectively calculate the visual features $\boldsymbol{v}_m$ and $\boldsymbol{v}_{n_m}$ from Q_m and f_{n_m} $(n_m = 1, 2, \ldots, \lfloor P_m N \rfloor)$ and calculate their cosine similarities w_{n_m} to extract the top $100P_h$ percent candidates, where f_{n_m} are the top $100P_m$ percent selected frames according to the similarities w_{n_l}. These procedures are also performed for the highest resolution query image Q_h and the further screened $\lfloor P_m P_h N \rfloor$ candidates. Finally, we can obtain the scenes for which frames have higher similarities than those of the other candidate frames. It should be noted that since the query images Q_m and Q_h focus on the information of the input sentence and its words, they have roles in narrowing down the retrieval candidates with consideration of the object relationship. By introducing the mechanism that hierarchically selects candidate scenes that are similar to the generated images Q_l and Q_m mainly reflecting the contents of the object relationship, we can screen only the scenes that are similar to objects of the input sentence. Although our scene retrieval framework is quite simple, it can successfully retrieve relevant scenes based on the hierarchical structure of AttnGAN.

IV. EXPERIMENTAL RESULTS

In this section, we quantitatively and qualitatively evaluate our framework by comparing it with some state-of-the-art retrieval methods. We first describe the details of datasets in IV-A. Results of quantitative and qualitative evaluations are presented in IV-B and IV-C, respectively.

TABLE 1. Results of subjective evaluation. These results show the average scores obtained from 25 subjects. The score of 1 represents "Not Relevant", and the score of 5 represents "Relevant".

	Scene1	Scene2	Scene3	Scene4	Scene5	Scene6	Scene7	Scene8	Scene9	Scene10	
PF	**4.12**	**3.76**	**3.92**	**4.40**	**3.84**	**4.16**	**4.84**	4.44	**4.72**	**3.32**	
BL	2.80	3.60	1.60	3.20	2.88	3.80	3.68	**4.76**	4.20	**3.32**	
CM1	1.04	1.04	1.04	1.02	2.04	1.84	1.08	1.08	1.68	1.04	
CM2	1.04	1.48	1.60	1.72	2.44	1.08	1.08	1.32	3.68	2.48	
CM3	3.48	3.60	1.04	3.04	2.04	**4.16**	3.68	4.44	4.44	2.96	
CM4	1.36	1.40	1.20	1.72	2.12	3.80	1.40	1.48	1.08	1.24	
	Scene11	Scene12	Scene13	Scene14	Scene15	Scene16	Scene17	Scene18	Scene19	Scene20	Average
PF	**4.12**	**4.64**	**2.68**	**4.48**	**4.96**	**4.96**	**3.24**	**4.88**	**4.52**	**3.64**	**4.18**
BL	1.12	1.76	1.36	3.88	3.96	2.84	2.04	3.48	3.44	2.96	3.03
CM1	1.12	1.40	1.08	3.20	1.52	3.08	1.92	1.08	1.28	1.20	1.49
CM2	1.24	1.96	1.12	3.88	3.16	4.48	1.92	1.08	4.00	1.60	2.12
CM3	3.24	1.96	2.12	1.64	3.08	2.80	3.16	2.40	4.20	**3.64**	3.06
CM4	1.04	3.68	2.12	3.20	4.88	2.73	2.08	3.48	1.32	1.20	2.13

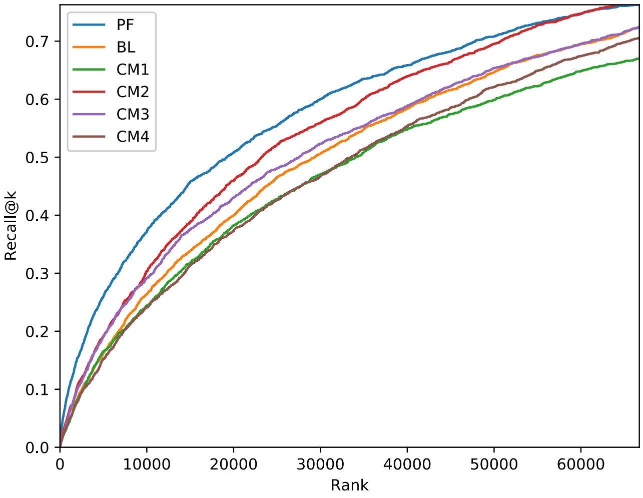


FIGURE 5. Recall@*k* obtained for the integrated dataset including five movies, "Bad Santa", "As Good as it Gets", "Halloween", "Rendezvous mit Joe Black" and "Harry Potter and the Prisoner of Azkaban".

A. DATASETS

We used the following two datasets in the experiment.

COCO dataset [37]

The COCO dataset consists of daily scene images and their description annotations. The dataset contains 82,783 training images, each of which is associated with 5 descriptions. In the proposed framework, we trained AttnGAN for text-to-image translation by using the COCO dataset. We used the COCO dataset since it contains various words and various daily scene images, and it has been widely used for text-to-image translation tasks. By evaluating the retrieval performance with this common dataset, we confirmed the capability of the proposed framework without fine-tuning for objective retrieval dataset.

MP-II MD dataset [58]

The MP-II MD dataset contains 68,000 scenes of 94 HD movies. This dataset contains a large number of scenes extracted from one movie, and each scene is associated with one description. In the experiment, we defined scenes corresponding to their descriptions, which were utilized as input descriptions for generating the query images, as the ground truth. We used this dataset for considering actual applications such as retrieval from one video.

B. QUANTITATIVE EVALUATION

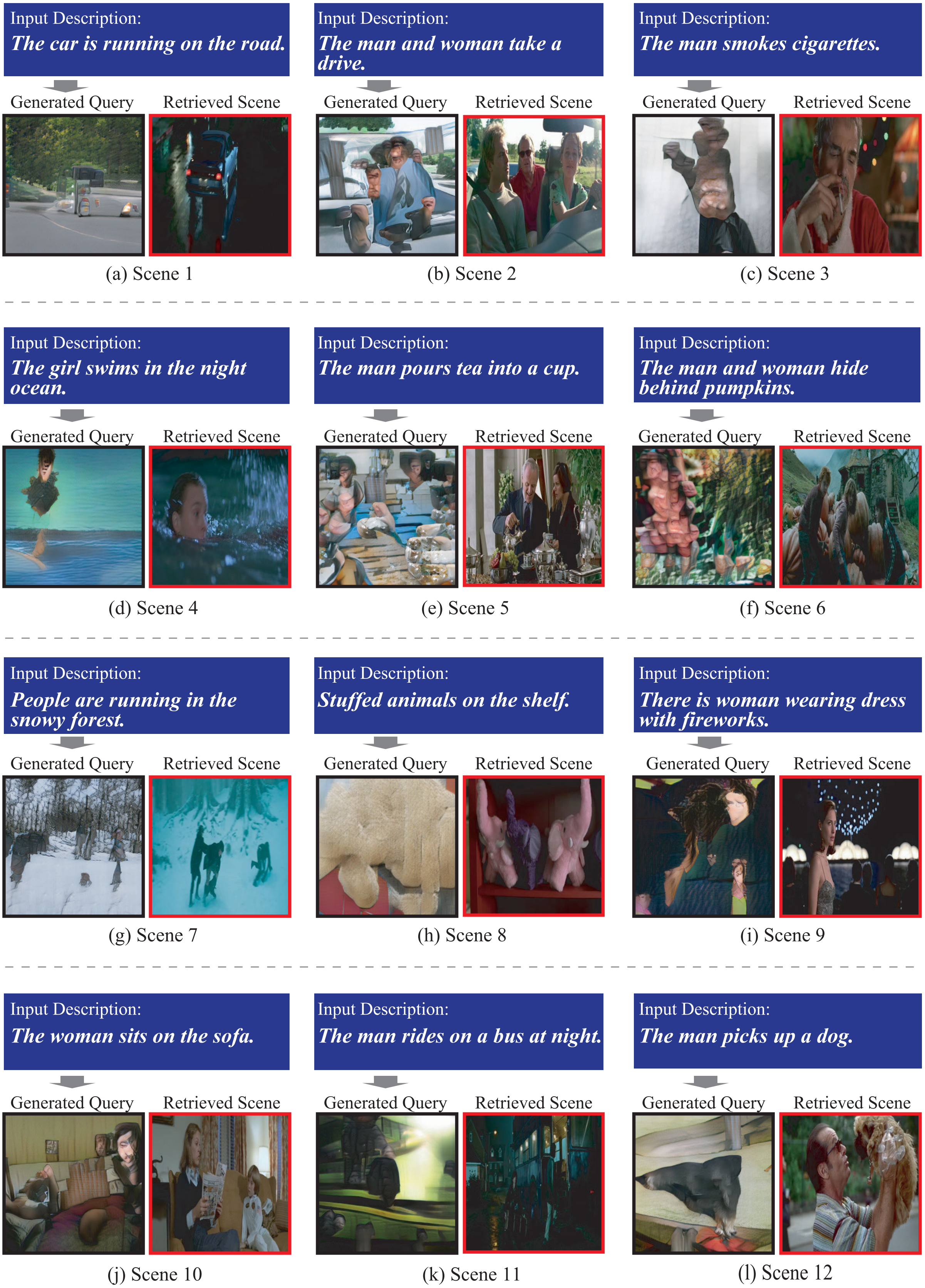
From the MP-II MD dataset, we selected three movies, "Bad Santa", "As Good As it Gets" and "Harry Potter and the Prisoner of Azkaban" which consist of 430, 538 and 592 scenes, respectively, with each scene having an average of 100 frames and with a total of 153,320 frames. By inputting the description of one scene to our framework, we performed retrieval and iterated these procedures for all scenes included in each movie. We defined frames included in the target scene as our ground truth and utilized the following Recall@k for the quantitative evaluation criterion:

$$\text{Recall@}k = \frac{r_k}{M} \quad (k = 1, 2, \ldots, N), \tag{12}$$

where r_k is the number of correctly retrieved scenes in the top-k retrieval results. In this experiment, we sorted all N frames in M candidate scenes according to their similarity ranks.

Furthermore, when the frames of the target scene were included in the top-k retrieval results, we regarded the target scene to have been correctly retrieved. We utilized Recall@k at the frame level because it can evaluate the performance in more detail compared with utilization of Recall@k at the scene level. In our framework, we simply set P_m and P_h to 50, and the sizes of the low-, middle- and high-resolution images (Q_l, Q_m and Q_h) were 64 × 64 pixels, 128 × 128 pixels and 256 × 256 pixels, respectively.

We compared the performance of the proposed framework (**PF**) with the performances of some state-of-the-art methods. We selected the following comparative methods.



(a) Scene 1 (b) Scene 2 (c) Scene 3

(d) Scene 4 (e) Scene 5 (f) Scene 6

(g) Scene 7 (h) Scene 8 (i) Scene 9

(j) Scene 10 (k) Scene 11 (l) Scene 12

FIGURE 6. Examples of the first retrieved frames by the proposed framework.

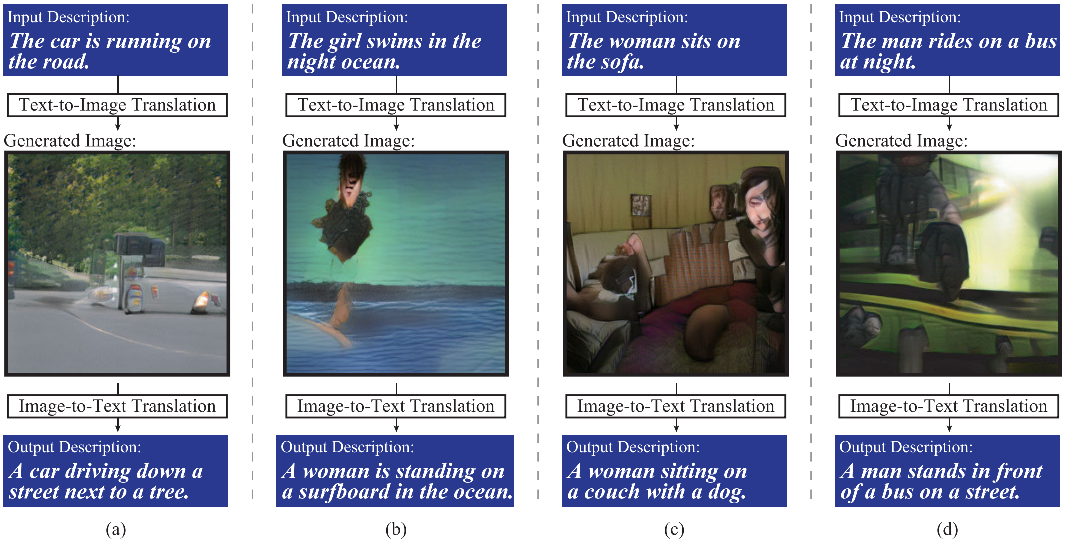

Input Description:
The car is running on the road.
Text-to-Image Translation
Generated Image:
Image-to-Text Translation
Output Description:
A car driving down a street next to a tree.
(a)
Input Description:
The girl swims in the night ocean.
Text-to-Image Translation
Generated Image:
Image-to-Text Translation
Output Description:
A woman is standing on a surfboard in the ocean.
(b)
Input Description:
The woman sits on the sofa.
Text-to-Image Translation
Generated Image:
Image-to-Text Translation
Output Description:
A woman sitting on a couch with a dog.
(c)
Input Description:
The man rides on a bus at night.
Text-to-Image Translation
Generated Image:
Image-to-Text Translation
Output Description:
A man stands in front of a bus on a street.
(d)

FIGURE 7. Examples of the image-to-text results obtained from images generated by the text-to-image GAN: (a)-(d) results respectively corresponding to Figs. 6 (a), (d), (j) and (k).

- **Baseline method (BL)**
  This is our baseline method that utilizes only high-resolution images generated by AttnGAN. By comparing with this method, we evaluated whether the use of the hierarchical structure in our framework is effective.
- **Comparative method 1 (CM1)** [8]
  This is a simple embedding method that utilizes deep learning-based techniques. It utilizes LSTM and CNN to compare visual and sentence features by embedding them into a common visual semantic space. We used this method as the baseline method using deep learning-based techniques.
- **Comparative method 2 (CM2)** [10]
  This is a method that takes the order relationship between words into consideration in addition to the mechanism of CM1. By comparing with this method, we evaluated whether the proposed framework can effectively use the sentence structure.
- **Comparative method 3 (CM3)** [9]
  This is a method that only utilizes the visual feature space. It embeds sentence features into the image feature space based on deep learning-based techniques. By comparing with this method, we evaluated whether the use of query images generated by the text-to-image GAN is effective.
- **Comparative method 4 (CM4)** [59]
  This is a method that adds a loss function that reduces the number of negative samples between a query and an objective sample in addition to the mechanism of CM1. We used this method since it was one of the most recent state-of-the-art methods.

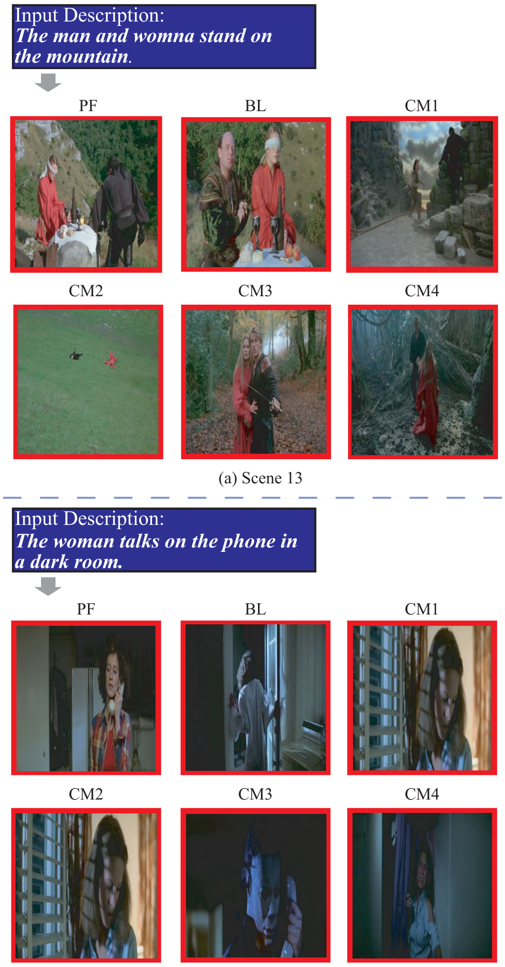
It should be noted that all of the comparative methods are constructed on the basis of open source codes provided by each author.

Figure 4 shows the results of Recall@k obtained for each movie. As shown in Fig. 4, the proposed framework outperforms the comparative methods (CM1, CM2, CM3 and CM4). In addition, since the proposed framework outperforms BL, which only utilizes high-resolution images, it can be seen that we can obtain better results by utilizing different resolution query images for reflecting the whole input description and their detailed words. Specifically, we can improve the retrieval performance by narrowing down the candidate scenes hierarchically utilizing the low-resolution image Q_l and middle-resolution image Q_m. Since each generated image contains different semantic information along the coarse-to-fine abstraction levels, the above screening is effective.

Furthermore, in order to verify the robustness of the proposed framework for a larger scale dataset that contains various scenes, we evaluated the retrieval performance for five movies selected from the MP-II MD dataset. We constructed an integration dataset including five movies, "Bad Santa", "As Good As it Gets", "Halloween", "Rendezvous mit Joe Black" and "Harry Potter and the Prisoner of Azkaban", including $430 + 538 + 676 + 296 + 592 (= 2,532)$ scenes with 247, 320 frames. Figure 5 shows the results of Recall@k obtained from this integrated dataset. As shown in this figure, the proposed framework also outperforms the comparative methods. It can be seen that the proposed framework can retrieve desired scenes more successfully.

C. QUALITATIVE EVALUATION

In this experiment, 25 subjects (5 females and 20 males, 20-27 years old) watched input descriptions and their corresponding retrieved first results obtained by our framework and the comparative methods. The subjects evaluated the relevance of the retrieved results in 5 grades ("1 Not Relevant",


Input Description:
The man and womna stand on the mountain.
PF
BL
CM1
CM2
CM3
CM4
Input Description:
The woman talks on the phone in a dark room.
PF
BL
CM1
CM2
CM3
CM4

(a) Scene 13

(b) Scene 17

FIGURE 8. The first retrieved frames of Scenes 13 and 17 that had low scores in the subjective evaluation.

"2 Not So Relevant", " 3 Neither Agree Nor Disagree", "4 A Little Relevant" and "5 Relevant"). We randomly selected 20 scenes from the MP-II MD dataset and gave their corresponding descriptions in this experiment. Examples of the retrieval results obtained by the proposed framework are shown in Fig. 6, and the results of qualitative evaluation are shown in Table 1. Each example in Fig. 6 respectively corresponds to the results in Table 1. In Table 1, the values of "PF" represent the average scores of all subjects obtained by the proposed framework. The values of "BL, CM1, CM2, CM3 and CM4" represent the average scores obtained by the comparative methods that are shown in the quantitative evaluation.

It can be seen that the scores of our framework averagely exceed "A Little Relevant". Therefore, the proposed framework can retrieve scenes related to the input descriptions. Also, the scores of our framework are better than those of "BL, CM1, CM2, CM3 and CM4". Furthermore, the differences are statistically significant in Welch's t-test with $p < 0.01$ given a significance level Îś $= 0.01$.

In Fig. 6, we can see that the proposed framework can retrieve relevant scenes even if the generated query images are not visually pleasant. From this fact, we can verify that the deep learning-based features obtained from the generated images have semantic information even if they are not visually pleasant. As additional evidence, we also show results of image-to-text translation from the images generated by our framework in Fig. 7. In this experiment, we utilized AttnGAN for the text-to-image translation and show and tell [39] for the image-to-text translation, and these two methods were completely independent. From this figure, we can see that the generated images seem to be translated to reasonable descriptions.

Although the effectiveness of the proposed framework was confirmed by the results of evaluations described in this paper, there are scenes with low scores (in particular, Scenes 13 and 17 in the qualitative evaluation). The retrieval results of Scenes 13 and 17 are shown in Fig. 8. In Fig. 8 and Table 1, we can see that the proposed framework and comparative methods obtain not too high scores and results even though the results obtained by the proposed framework were better than the results obtained by the other methods. There is therefore room for improvement of the proposed framework.

V. CONCLUSION

In this paper, we have proposed *Query is GAN*, a novel scene retrieval framework that utilizes images generated by AttnGAN as query images. Experimental results have shown that the proposed framework can accurately retrieve scenes and enables users to find their desired scenes. Furthermore, by showing the effectiveness of the proposed framework, the usefulness of the generated images, which are not visually pleasant, can also be confirmed. In a future work, we will introduce temporal processing to the proposed framework for realizing ideal scene retrieval.

REFERENCES

[1] Z. Wang, L. Sun, W. Zhu, S. Yang, H. Li, and D. Wu, "Joint social and content recommendation for user-generated videos in Online social network," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 698–709, Apr. 2013.

[2] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Netw. Appl.*, vol. 19, no. 2, pp. 171–209, Apr. 2014.

[3] Z. Wang, L. Sun, C. Wu, W. Zhu, Q. Zhuang, and S. Yang, "A joint online transcoding and delivery approach for dynamic adaptive streaming," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 867–879, Jun. 2015.

[4] D. Wu, S. Ji, P. Zhou, J. Xu, and Y. Feng, "Video big data retrieval over media cloud: A context-aware online learning approach," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1762–1777, Jul. 2019.

[5] B. Prabhakaran, Y.-G. Jiang, H. Kalva, and S.-F. Chang, "Editorial IEEE transactions on multimedia special section on video analytics: Challenges, algorithms, and applications," *IEEE Trans. Multimedia*, vol. 20, no. 5, p. 1037, May 2018.

[6] Y. Rui, T. S. Huang, and S.-T. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *Vis. Commun. Image Represent.*, vol. 10, no. 1, pp. 39–62, 1999.

[7] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognit.*, vol. 40, no. 1, pp. 262–282, 2007.

[8] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," Nov. 2014, *arXiv:1411.2539*. [Online]. Available: https://arxiv.org/abs/1411.2539

[9] J. Dong, X. Li, and C. G. M. Snoek, "Word2VisualVec: Image and video to sentence matching by visual feature prediction," Apr. 2016, *arXiv:1604.06838*. [Online]. Available: https://arxiv.org/abs/1604.06838

[10] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–12.

[11] A. Anjulan and N. Canagarajah, "Video scene retrieval based on local region features," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2006, pp. 3177–3180.

[12] H.-W. Yoo and S.-B. Cho, "Video scene retrieval with interactive genetic algorithm," *Multimedia Tools Appl.*, vol. 34, no. 3, pp. 317–336, 2007.

[13] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 41, no. 6, pp. 797–819, Nov. 2011.

[14] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning, "Generating semantically precise scene graphs from textual descriptions for improved image retrieval," in *Proc. 4th Workshop Vis. Lang.*, Sep. 2015, pp. 70–80.

[15] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3668–3678.

[16] D. Lu, X. Liu, and X. Qian, "Tag-based image search by social re-ranking," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1628–1639, Aug. 2016.

[17] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based LSTM and semantic consistency," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2045–2055, Sep. 2017.

[18] S. Chen, Q. Jin, J. Chen, and A. G. Hauptmann, "Generating video descriptions with latent topic guidance," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2407–2418, Sep. 2019.

[19] X. Ke, J. Zou, and Y. Niu, "End-to-End automatic image annotation based on deep CNN and multi-label Data augmentation," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 2093–2106, Aug. 2019.

[20] A. Alzu'bi, A. Amira, and N. Ramzan, "Semantic content-based image retrieval: A comprehensive study," *Vis. Commun. Image Represent.*, vol. 32, pp. 20–54, Oct. 2015.

[21] S. Ercoli, M. Bertini, and A. Del Bimbo, "Compact hash codes for efficient visual descriptors retrieval in large scale databases," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2521–2532, Nov. 2017.

[22] S. Pang, J. Ma, J. Xue, J. Zhu, and V. Ordonez, "Deep feature aggregation and image re-ranking with heat diffusion for image retrieval," *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1513–1523, Jun. 2019.

[23] G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, and I. Kompatsiaris, "FIVR: Fine-grained Incident video retrieval," *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2638–2652, Oct. 2018.

[24] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: A comprehensive study," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 157–166.

[25] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1224–1244, May 2018.

[26] S. Park, H. Park, and C. D. Yoo, "Complex video scene analysis using kernelized-collaborative behavior pattern learning based on hierarchical representative object behaviors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 6, pp. 1275–1289, Jun. 2017.

[27] Q. Abbas, M. E. A. Ibrahim, and M. A. Jaffar, "Video scene analysis: An overview and challenges on deep learning algorithms," *Multimedia Tools Appl.*, vol. 77, no. 16, pp. 20415–20453, 2018.

[28] I. J. Goodfellow, J. Pouget-abadie, M. Mirza, B. Xu, D. Warde-farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[29] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," May 2016, *arXiv:1605.05396*. [Online]. Available: https://arxiv.org/abs/1605.05396

[30] N. Bodla, G. Hua, and R. Chellappa, "Semi-supervised FusedGAN for conditional image generation," Jan. 2018, *arXiv:1801.05551*. [Online]. Available: https://arxiv.org/abs/1801.05551

[31] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 217–225.

[32] Z. Zhang, Y. Xie, and L. Yang, "Photographic text-to-image synthesis with a hierarchically-nested adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6199–6208.

[33] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Liwicki, and M. Z. Afzal, "TAC-GAN—Text conditioned auxiliary classifier generative adversarial network," Mar. 2017, *arXiv:1703.06412*. [Online]. Available: https://arxiv.org/abs/1703.06412

[34] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis.*, Oct. 2017, pp. 5907–5915.

[35] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1316–1324.

[36] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.

[37] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[38] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3128–3137.

[39] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3156–3164.

[40] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[41] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2623–2631.

[42] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. ACM Int. Conf. Multimedia*, Nov. 2014, pp. 7–16.

[43] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 394–407, Feb. 2018.

[44] Y. Huang, W. Wang, and L. Wang, "Instance-aware image and sentence matching with selective multimodal LSTM," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2310–2318.

[45] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image-sentence embeddings using large weakly annotated photo collections," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 529–545.

[46] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3441–3450.

[47] A. Karpathy, A. Joulin, and L. F. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1889–1897.

[48] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," Nov. 2017, *arXiv:1711.09020*. [Online]. Available: https://arxiv.org/abs/1711.09020

[49] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5967–5976.

[50] E. L. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1486–1494.

[51] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6882–6890.

[52] W. R. Tan, "ArtGan: Artwork synthesis with conditional categorical GANs," in *Proc. IEEE Conf. Image Process.*, Sep. 2017, pp. 3760–3764.

[53] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, "MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," in *Proc. AIII Conf. Artif. Intell.*, 2018, pp. 34–41.

[54] M. Hou, B. Chaib-draa, C. Li, Q. Zhao, and S. Engineering, "Generative adversarial positive-unlabeled learning," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 2255–2261.

[55] K. Chen, C. B. Choy, M. Savva, A. X. Chang, T. Funkhouser, and S. Savarese, "Text2Shape: Generating shapes from natural language by learning joint embeddings," Mar. 2018, *arXiv:1803.08495*. [Online]. Available: https://arxiv.org/abs/1803.08495

[56] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

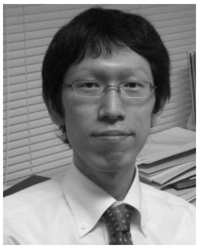
[57] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[58] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3202–3212.

[59] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," Jul. 2017, *arXiv:1707.05612*. [Online]. Available: https://arxiv.org/abs/1707.05612

RINTARO YANAGI received the B.S. degree in electronics and information engineering from Hokkaido University, Japan, in 2019, where he is currently pursuing the M.S. degree with the Graduate School of Information Science and Technology. His research interest includes machine learning and its applications.

REN TOGO received the B.S. degree in health sciences from Hokkaido University, Japan, in 2015, and the M.S. and Ph.D. degrees from the Graduate School of Information Science and Technology, Hokkaido University, in 2017 and 2019, respectively. He is a Radiological Technologist. His research interest includes machine learning and its applications.



TAKAHIRO OGAWA received the B.S., M.S., and Ph.D. degrees in electronics and information engineering from Hokkaido University, Japan, in 2003, 2005, and 2007, respectively, where he is currently an Assistant Professor with the Faculty of Information Science and Technology Division of Media and Network Technologies. His research interest includes multimedia signal processing and its applications. He is a member of EURASIP, IEICE, and the Institute of Image Information and Television Engineers (ITE). He has been an Associate Editor of *ITE Transactions on Media Technology and Applications*.

MIKI HASEYAMA received the B.S., M.S., and Ph.D. degrees in electronics from Hokkaido University, Japan, in 1986, 1988, and 1993, respectively, where she joined the Graduate School of Information Science and Technology as an Associate Professor, in 1994. She was a Visiting Associate Professor with Washington University, USA, from 1995 to 1996. She is currently a Professor with the Faculty of Information Science and Technology Division of Media and Network Technologies, Hokkaido University. Her research interest includes image and video processing and its development into semantic analysis. She is a member of IEICE, Institute of Image Information and Television Engineers (ITE), and Acoustical Society of Japan (ASJ). She has been a Vice-President of the Institute of Image Information and Television Engineers, Japan (ITE), an Editor-in-Chief of *ITE Transactions on Media Technology and Applications*, and a Director of International Coordination and Publicity of The Institute of Electronics, Information and Communication Engineers (IEICE).