PAPER

Image Generative Semantic Communication with Multi-Modal Similarity Estimation for Resource-Limited Networks

Eri HOSONUMA[†a)], ***Student Member***, **Taku YAMAZAKI**[††], **Takumi MIYOSHI**[††,†], **Akihito TAYA**[†], **Yuuki NISHIYAMA**[†††], ***and*** **Kaoru SEZAKI**[†††,†], ***Members***

SUMMARY

To reduce network traffic and support environments with limited resources, a method for transmitting images with low amounts of transmission data is required. Several machine learning-based image compression methods, which compress the data size of images while maintaining their features, have been proposed. However, in certain situations, reconstructing the semantic information of images at the receiver end may be sufficient. To realize this concept, semantic-information-based communication, called semantic communication, has been proposed, along with an image transmission method using semantic communication. This method transmits only the semantic information of an image, and the receiver reconstructs the image using an image-generation model. This method utilizes one type of semantic information for image reconstruction, but reconstructing images similar to the original image using only it is challenging. This study proposes a multi-modal image transmission method that leverages diverse semantic information for efficient semantic communication. The proposed method extracts multi-modal semantic information from an original image and transmits only it to a receiver. Subsequently, the receiver generates multiple images using an image-generation model and selects an output image based on semantic similarity. The receiver must select the result based only on the received features; however, evaluating semantic similarity using conventional metrics is challenging. Therefore, this study explored new metrics to evaluate the similarity between semantic features of images and proposes two scoring procedures for evaluating semantic similarity between images based on multiple semantic features. The results indicate that the proposed procedures can compare semantic similarities, such as position and composition, between semantic features of the original and generated images. Thus, the proposed method can facilitate the transmission and utilization of photographs through mobile networks for various service applications, including monitoring, tracking, and detection.

key words: *semantic communication, image generation, image transmission, image captioning, semantic segmentation*

1. Introduction

It is expected to increase situations wherein photos taken at specific locations are transmitted through a mobile network and used for various services and applications such as monitoring, tracking, and detection. Additionally, recent surveys of undersea resources using underwater communication [1], [2] and environmental monitoring using satellite communication in severe locations [3]–[5] have also been conducted. Therefore, image transmission technologies will have to realize transmitting numerous images from multiple locations in the future.

However, it is important to reduce network traffic owing to data transmission including image transmission, because global network traffic is expected to continue increasing [6]–[8]. Additionally, developing periodic image transmission techniques for recent applications in severe environments wherein communication resources are extremely limited such as underwater remains challenging. Consequently, image transmission using fewer network resources has the potential for use in various applications.

A common approach for reducing the communication cost of image transmission is to use image compression methods [9]–[11]. Furthermore, machine-learning-based image compression methods that realize high compression rates have been proposed [12], [13]. However, these methods do not consider the transmission process. In contrast, deep joint source and channel coding (DeepJSCC) [14], [15] schemes have been developed to compress and transmit images efficiently by optimizing the joint coding scheme to suit wireless channels. DeepJSCC encodes images based on the semantic features in the data to be transmitted and realizes image transmission while maintaining their visual information. Communication technologies, such as DeepJSCC, that focus on the meaning and semantics in communication bits have recently attracted attention as semantic communication [16]–[18].

These methods aim to transmit images while maintaining the visual identity of the images. However, transmitting all features of the images is redundant in certain situations. For example, in an application aimed toward monitoring specific objects, the background information in a captured photo is less important than that related to the target object, such as its type and composition. Moreover, when viewing a live camera stream at an event or exhibition, the faces of people and background information are not important for viewers because their primary interests lie in the event and exhibits. In such situations, the transmitter does not need to transmit the all features contained in the image to the receiver.

Therefore, in such situations, if a transmitter can extract specific features from an original image based on the requirements of the receiver, the amount of data transmitted can be reduced by transmitting only the features extracted from the image. Additionally, owing to recent developments

Manuscript received January 1, 2015.
Manuscript revised January 1, 2015.

[†]The authors are with Institute of Industrial Science, The University of Tokyo, Meguro-ku, 153-8505, Japan.
[††]The authors are with Collage of Systems Engineering and Science, Shibaura Institute of Technology, Saitama-shi, 337-8570, Japan.
[†††]The authors are with Center for Spatial Information Science, The University of Tokyo, Kashiwa-shi, 277-8568, Japan.
a) E-mail: hosonuma@mcl.iis.u-tokyo.ac.jp

DOI: 10.1587/transcom.E0.B.1

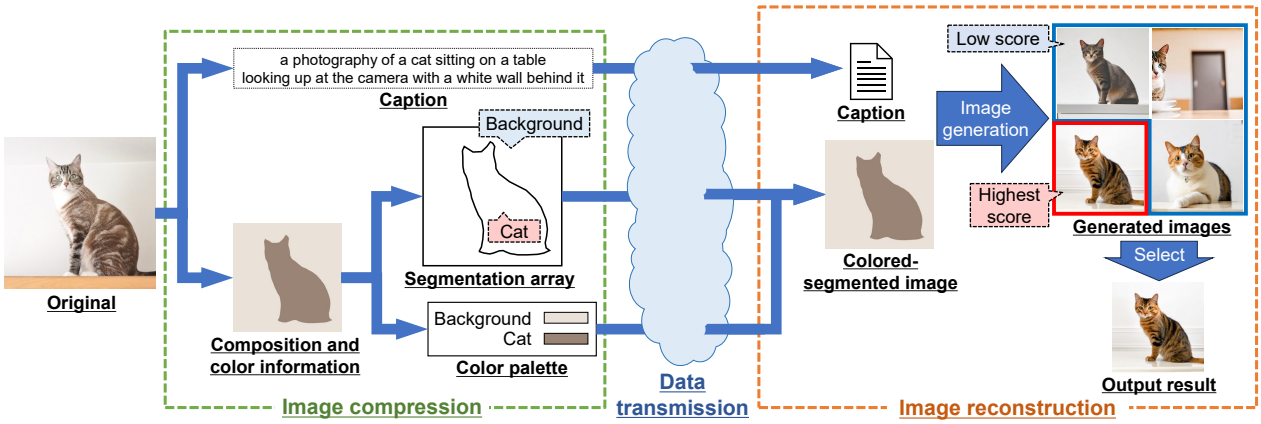


Fig. 1 Overview of the proposed image transmission method. A transmitter generates multiple features from an image for extracting semantic information and transmits only the features to achieve data reduction. A receiver receives the extracted features and reconstructs the image using an image-generation model.

in the field of artificial intelligence, image-generation models [19]–[21] that allow generating images using descriptive information or sketches have become widely available. Moreover, recent real-time image-generation models [22] can apply image-generation mechanisms to time-sensitive applications. Consequently, there is potential for the development of image-generation-based image transmission, which can be considered a type of semantic communication in the future.

As an example of such a kind of communications, the authors of [23] have proposed an image transmission method based on the descriptive information contained in an image. In this method, the transmitter transmits only the descriptive text information extracted from the original image using an image-to-text algorithm to the receiver, and the receiver then reconstructs the image based on the received information through an image-generation model. Although this method can significantly reduce the communication cost compared to existing image transmission methods, some important semantic features, such as object position and composition, might be lost because it only relies on a single semantic feature from the reconstructed image. Therefore, it is necessary to extract and transmit multi-modal semantic features from the original image to realize image reconstruction with high semantic similarity and satisfy the requirements of the receiver.

To address this issue, we previously proposed an image transmission concept that employs multiple types of semantic information [24]. In this method, a transmitter extracts multiple semantic features from an original image, such as composition and color, in addition to captions, and transmits them to a receiver, which reconstructs the image using an image-generation model. The images generated using multiple semantic features are expected to have a higher similarity to the original images than those generated using only a single semantic feature. However, a quantitative analysis of the semantic similarity between the original and generated images was not conducted in [24].

In this study, we propose a highly efficient image transmission method based on the concept proposed in [24] and employ newly designed evaluation metrics to assess the semantic similarity between the original and generated images. In the proposed method, an image-generation method at the receiver reconstructs images with the semantic information required by its application using the multi-modal semantic information extracted from the original image. Fig. 1 shows an overview of the proposed method. To achieve image transmission with a small amount of data, the transmitter transmits multiple features of the original image, and the receiver reconstructs an image with the same semantic information as the original image. First, the transmitter extracts multiple features representing composition, description, and color information from the original image as diverse semantic information using existing machine learning algorithms. Thereafter, the transmitter transmits only the extracted semantic features to the receiver, with the aim of minimizing the amount of data transmitted. The receiver generates multiple images by inputting the received features into an image-generation model. As the image-generation model generates various images, the receiver must select the best image as an output image. To evaluate the reconstruction result, this study developed quality assessment procedures, which is introduced later. In the proposed method, the receiver reconstructs images by the received semantic information to be easily recognized by humans. Therefore, the proposed method can be potentially used in various applications.

To select an appropriate image using the proposed method, the quality of the generated images must be assessed based on the semantic similarity between the original and generated images on the receiver. Although some metrics assess the image quality by comparing two images [25], the proposed method cannot compare the original and generated images directly because the receiver does not possess the original image. Therefore, the quality of the generated images must be assessed based only on the received semantic features of the original image. To address this issue, this

study investigates various metrics to evaluate the semantic similarity between the original and generated images, with a particular focus on multi-modal semantic information. This study proposes two scoring procedures for evaluating the similarity between semantic features, such as descriptive and segment information, of the original and generated images. Additionally, this study also proposes a background recoloring method for enhancing the contours of objects in images to improve image generation.

The contributions of this study can be summarized as follows:

- We propose an image construction method that reconstructs images at the receiver end using multi-modal semantic communication. Therefore, the proposed method can transmit images with higher semantic similarity information than conventional transmission methods. Additionally, the proposed method only uses open-source machine learning algorithms, making it easy to deploy and underscoring its potential for use in various applications.
- The proposed method demonstrates better reproducibility than [24] because it evaluates the generated images based on their similarity between the semantic features of the original and generated images. To realize this, we propose two scoring procedures for evaluating the semantic similarity between images. These procedures can evaluate the semantic similarity between images based only on the semantic features received by the receiver.
- We propose and investigate new metrics for evaluating the semantic similarity between the original and generated images through an experiment. This is an important first step toward establishing evaluation metrics for image transmission methods using semantic communication. The results of the experiment indicate that the proposed scoring procedures can evaluate semantic features of the target object, such as position and composition, and the entire of background information of the image. Additionally, the results demonstrate that the proposed method with the background recoloring technique can effectively reconstruct the composition and position of the target object in the generated images.

The remainder of this paper is organized as follows: Section 2 presents the existing image compression and transmission methods. Section 3 introduces the proposed image-generation-based transmission methods. Section 4 investigates evaluation metrics for evaluating semantic similarity between the original images and those generated using the proposed methods. Section 5 elucidates the experiments performed to evaluate the proposed methods and the semantic similarity between images using the proposed metrics mentioned previous section. Section 6 presents and discusses the experimental results. Section 7 concludes the paper and presents future research directions.

2. Related Work

Many studies have investigated various machine-learning-based image compression methods [12], [13], [26]–[30]. These methods compress the image data using various machine learning algorithms such as convolutional neural networks (CNNs) [26]–[28] and autoencoder [29], [30]. These methods have demonstrated higher compression ratios while maintaining better image quality than traditional image compression algorithms such as JPEG [10] and JPEG2000 [11]. However, these methods focus only on source coding to reduce the image data size. When a device transmits images through a network, a channel coding algorithm must be used to prevent errors caused by channel noise.

To address this issue, DeepJSCC-based methods were recently proposed to further improve image transmission efficiency [14], [15]. These methods encode an image in a batch using a CNN, which differs from the conventional procedures of employing separate source and channel coding steps. These methods exhibit better performance than conventional image compression algorithms, such as JPEG and JPEG2000, for a specific channel bandwidth and a low signal-to-noise ratio.

The goal of these methods is to transmit visually complete images between end-to-end devices. That is, these methods aim to transmit images while preserving all the semantic information visually recognizable by humans, including the composition, background, and target object type. However, in some situations, it is redundant to transmit all semantic information contained in images. Therefore, to significantly reduce the amount of data transmitted during image transmission, only some semantic information must be extracted from an image and transmitted to the receiver. Recently, owing to the development of neural networks, various machine learning algorithms, such as image captioning [31], [32] and semantic segmentation [33], [34], that can extract specific semantic information from an image have been proposed to convert images into descriptive or segmented information.

This approach, which focuses on the transmission of semantic information, is known as semantic communication [16]–[18]. Semantic communication aims to achieve end-to-end data transmission while maintaining the content at the semantic level rather than at the bit level. In other words, semantic communication is considered to be successful when original and received data are semantically the same, even if the received data have been changed from the original data at the bit level.

[23] proposed an image transmission method that transmits only the image caption to further improve communication efficiency. In this method, a transmitter generates a caption as descriptive information of the original image using an image captioning algorithm. Subsequently, the transmitter transmits only the generated caption, and the receiver reconstructs the image based on the received caption by using an image-generation model.

As this method extracts only the descriptive information from the original image, it may incur a loss of some semantic information contained in the original image that is difficult to represent through the descriptive information. For example, it is difficult to represent the exact position and composition of a target object in an image using only descriptive information. Therefore, to reconstruct an image more clearly at the receiver, the transmitter must extract and transmit multi-modal semantic information from the original image. To address this issue, we previously proposed a novel image transmission method that uses multi-modal semantic information [24]. In this method, a transmitter extracts descriptive, segmented, and color information from the original image as semantic features, and transmits only these features. After receiving these features, the receiver reconstructs the image using an image-generation model. By conveying diverse information, such as the composition and color of the original image, in addition to descriptive information, this method enables image transmission that can generate images that are more semantically similar to the original image compared with existing methods. However, [24] included only subjective classification of the generated images as a preliminary study. Therefore, quantitative evaluations of the similarities between original and generated images are lacking.

To evaluate the overall performance of image-compression algorithms, the peak signal-to-noise ratio (PSNR) is often employed as a metric to assess the visual similarity between two images. PSNR is defined as the ratio of the maximum power of a signal to the noise that degrades the image quality. Another metric, called the learned perceptual image patch similarity (LPIPS) [25], has been employed in existing semantic communication methods for image transmission [23]. LPIPS focuses on the distance in the latent space and compares the feature output of the convolution layers of a trained neural network model for image classification.

These metrics are useful for evaluating methods that aim to fully reconstruct the original image. However, the proposed image transmission method aims to reconstruct an image using specific semantic features required by the receiver and extracted from the original image. In other words, the proposed method does not involve transmitting visually complete images identical to the original image. Therefore, it is necessary to evaluate the semantic similarities between the original image and the images generated using the proposed method to compare them. However, evaluating semantic similarity using the conventional evaluation metrics is challenging.

3. Image Generation-based Transmission Method

This section describes the proposed image generation-based transmission method that uses multi-modal semantic information. A transmitter extracts multiple semantic features, such as composition, description, and color information, representing diverse semantic information from an image and transmits only these extracted features. The details of these extracted features are explained in Sec. 3.1. The receiver generates multiple images using the received features and selects an output image based on the semantic similarity. The details of these procedures at the receiver are explained in Sec. 3.2. The proposed method focuses on reducing both the amount of transmitted data and image reconstruction from the receiver's perspective while maintaining the extracted semantic features.

3.1 Feature Extraction at the Transmitter

Fig. 2 illustrates the process employed by the transmitter for extracting multi-modal semantic features from the original image (size $m \times n$ pixels) using existing machine learning algorithms.

First, the transmitter generates a caption as the descriptive information using an image captioning algorithm [31], [32], [35]. The descriptive information contains the target object type and broad background information of the original image.

Next, the transmitter performs semantic segmentation [33], [34] on the original image to obtain a segmentation array that represents the segment information of the image. Semantic segmentation algorithms realize image segmentation by labeling each pixel of an image. Therefore, the segmentation array is an $m \times n$ two-dimensional array composed of labels for all pixels of the original image. The segment information represents the position and composition of the target object in the original image.

Third, the transmitter generates a color palette by calculating the average RGB value for each label in the segmentation array. This is because the detailed color information of the image is lost when only the caption and segmentation array are generated, and it is difficult to reconstruct the color information of the image by the receiver. The color palette represents the color information of the segments contained in the original image. After extracting the three types of semantic features, the transmitter transmits this data, the size of which is significantly smaller than that of the original image, to the receiver.

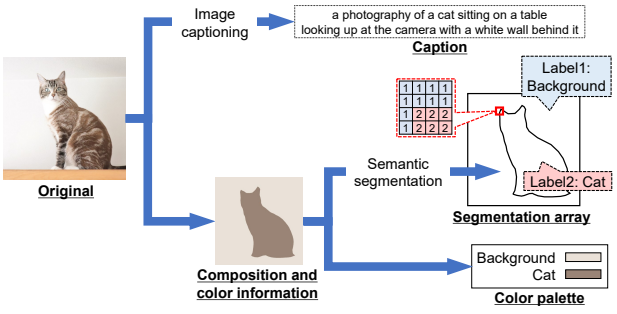


Fig. 2 Procedure for extracting semantic information from the original image using the transmitter. The transmitter generates a caption, segmentation array, and color palette from the original image, representing its descriptive, segment, and color information.

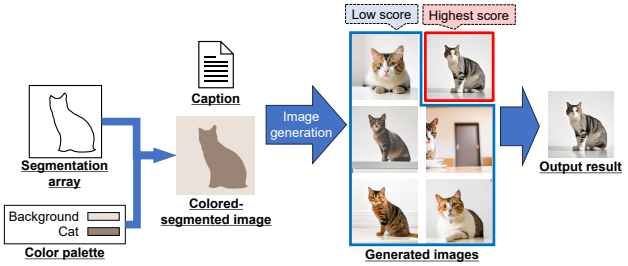


Fig. 3 Procedure for reconstructing the image at the receiver using the image-generation model. The receiver generates multiple images by inputting the received caption and generated segmented-colored image into an image-generation model. After the generation, the receiver selects an output image based on the semantic similarity between the received features and those of the generated images.

3.2 Image Reconstruction by the Receiver

Fig. 3 illustrates the image reconstruction procedure employed by the receiver. After receiving the semantic features, the receiver reconstructs the image using the received information and an image-generation model. First, the receiver generates a colored-segmented image from the segmentation array and color palette, which is then input into the image-generation model. The colored-segmented image is created by changing the value of each label in the segmentation array to its corresponding RGB value in the color palette. Subsequently, the receiver inputs the received caption and generated colored-segmented image into the image-generation model to generate multiple images. Here, the receiver must evaluate the semantic similarities between the original and generated images to output the best result among the generated images. However, it is not possible to directly compare the similarities between the images because the receiver does not have the original image. To address this issue, the receiver assigns similarity scores based on the received semantic features and generated images. The detailed scoring procedures for semantic similarity are discussed below. After scoring, the receiver selects the image with the highest score as the output. Thus, the proposed method realizes image transmission based on the semantic information contained in an image using these procedures.

3.3 Background Recoloring for Enhancing Object Contours

As shown in Fig. 4, the image-generation model cannot accurately identify the segment information if similar RGB values are assigned to the "background" and target object labels. Therefore, the receiver generates images with completely different compositions than that of the original image when using this segment information. To avoid this undesirable image generation, this study proposes an extensive approach that changes the color of pixels labeled "background" to white. When employing this approach, the segment information of the original image is more likely to be recognized by the image-generation model because the area of the target object in the colored-segmented image stands out. However, the background color information of the original image may be lost because the background color of the colored-segmented image is converted to white.

4. Evaluation Metrics for the Proposed Method

This section describes the similarity scoring procedures for selecting an output from the generated images. As described in Sec. 3.2, to score the generated images, the receiver requires similarity metrics. However, the receiver cannot directly compare the similarity of the images because it cannot access the original image. To address this issue, this study proposes two scoring procedures to compare the semantic similarity between the received features of the original image and those of the images generated by the receiver.

Scoring procedure for descriptive information similarity: This procedure compares the descriptive information in the original images and those generated by the receiver. The receiver generates captions for the generated images and calculates the text similarity between them and that of the original image. This procedure uses BERTScore [36], which is a metric used to assess the semantic similarity between

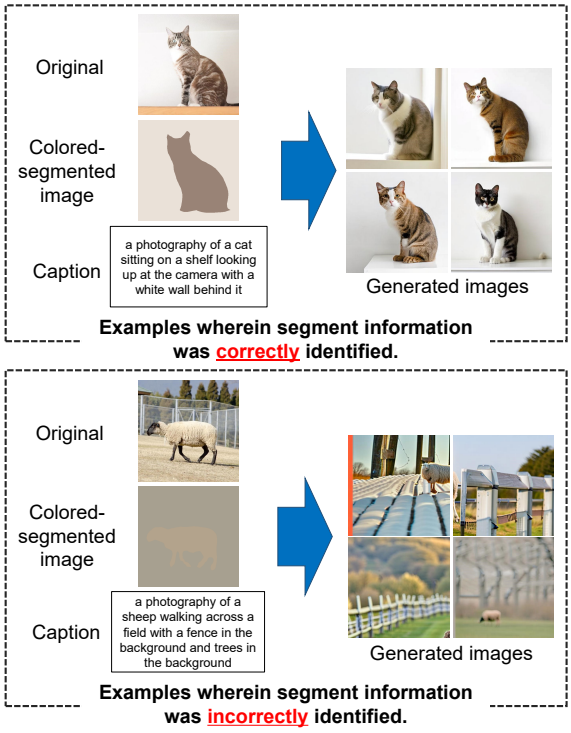


Fig. 4 Examples wherein color information affects the segment information recognition of the image-generation model. In the upper example, the image-generation model correctly recognizes the composition of the target object because the boundary between the object and background in the colored-segmented image is clear. In contrast, in the lower example, the positions of the target object in the generated images differ significantly from that in the original image, which may be caused by the similar RGB values of the target object and background in the colored-segmented image.

sentences to evaluate text similarity. BERTScore calculates text similarity using bidirectional encoder representations from transformers (BERT) [37], which is a natural language processing model. Text similarity is calculated based on the vectors of input sentences generated by the trained BERT.

Scoring procedure for segment similarity: This procedure compares the segment information contained in the original and generated images. The receiver first converts the generated images into segmentation arrays using a semantic segmentation algorithm, similar to the transmitter. As described in Sec. 3, each element of the segmentation array contains a label for the corresponding pixel. Note that the segmentation arrays of the original image S_{org} and the generated image S_{seg} are defined as follows:

$$\begin{aligned} S_{\text{org}} &:= (p_1, p_2, \cdots p_N) \\ S_{\text{seg}} &:= (q_1, q_2, \cdots q_N), \end{aligned} \tag{1}$$

where p and q denote the labels stored in each element of the segmentation arrays, that is the corresponding pixels. Additionally, $N := mn$ denotes the size of segmentation arrays, i.e. the number of pixels of the original and generated images, which contain, the same number of pixels. Subsequently, the receiver calculates the segmentation matching rate SMR between S_{org} and S_{seg} as follows:

$$\text{SMR} := \frac{1}{N} \sum_{i=1}^{N} \delta(p_i, q_i). \tag{2}$$

where $\delta(p_i, q_i)$ represents the Kronecker delta that indicates whether the two arguments are the same, and is defined as follows:

$$\delta(x, y) := \begin{cases} 1 & (x = y) \\ 0 & (x \neq y). \end{cases} \tag{3}$$

The receiver calculates SMR by determining the percentage of pixels with matching labels among all pixels. SMR can compare the position and composition of the target object between the images.

5. Experiment Setup

We conducted an experiment to evaluate the performances of the proposed methods and scoring procedures for determining the similarity between the semantic features of the original and generated images. This experiment included 105 images; five each containing 21 different objects (airplane, bicycle, bird, boat, bottle, bus, car, cat, chair, table and chairs, cow, dog, horse, motorbike, person, potted plant, sheep, sofa, table, train, and TV). All images were in the JPEG format with a quality of 80 and size of 512×512 pixels, and were obtained from photoAC [38], which contains copyright-free images.

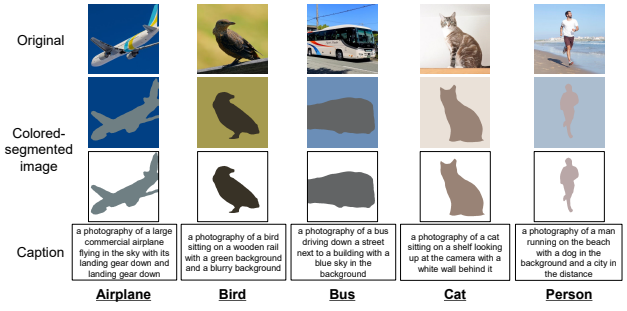


Fig. 5 Examples of original images, segmented colored images, and captions used in the experiment.

First, all the images were converted into captions and segmentation arrays using image captioning and semantic segmentation algorithms. This experiment used bootstrapping language-image pre-training for unified vision-language understanding and generation (BLIP) [31] for image captioning and DeepLabV3 [33] for semantic segmentation. In addition, color palettes for each image were simultaneously created based on the segmentation arrays. Note that BLIP was set to generate captions containing 20–30 words beginning with "a photography of" because the experiment employed only photos. After conversion, the colored-segmented images were created in the PNG format using the segmentation arrays and corresponding color palettes. Fig. 5 shows examples of the colored-segmented images and captions used in this experiment. Subsequently, 50 images were generated for each original image by inputting the colored-segmented image and the caption into Stable Diffusion [19], which is a widely used image-generation model. Note that in this experiment, Stable Diffusion was instructed to avoid the following factors to prevent generating images with textures different from the original images.

"low quality, worst quality, out of focus, ugly, error, jpeg artifacts, lowers, blurry, broken, illustration, animation, painting, 2D, oil painting, sketch, watercolor, ink, flat color"

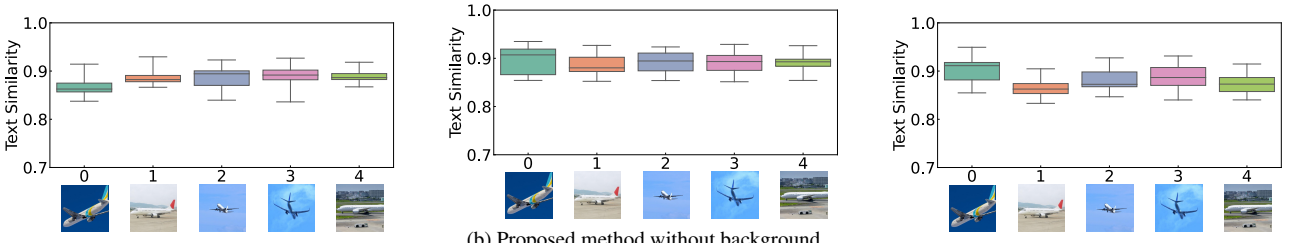
We compared the semantic features of the original images and those generated by each method using the scoring procedures described in Sec. 4. Note that in the text similarity evaluation, we compared the performance with and without removing stop words from the captions. Stop words are prepositions and articles that are removed during preprocessing for natural language processing as their presence can result in excessively high similarity scores. We also evaluated and compared the performance of the simplest method, which generated images using only the caption generated from the original images, with that of the two methods proposed in Sec. 3.

6. Experiment Results and Discussion

This section presents and discusses the results of the experiment described in Sec. 5, wherein the following aspects. First, the reduction in transmission data size using the proposed method was evaluated in Sec. 6.1. Second, the image-generation performances of each method were compared using the proposed scoring procedures in Sec. 6.2 and Sec. 6.3. Third, the impacts of the content and combinations of seman-

Table 1 Data sizes employed by each method. The average data size of the captions is approximately $1/4000^{\text{th}}$ of that of the images in the JPEG format. Additionally, even when multiple semantic features are combined, the data size is approximately $1/20^{\text{th}}$ of that of the JPEG format images.

Uncompressed	JPEG	Caption	Caption + color palette + segmentation array
786 [KB]	41.8 [KB]	0.0998 [KB]	2.09 [KB]



(a) Using only the caption. (b) Proposed method without background recoloring. (c) Proposed method with background recoloring.

Fig. 6 Text similarity scores of the "airplane" images generated by each method. Note that the stop words are not removed in these results. The original images are shown at the bottom of the graphs. The results of the proposed method with background recoloring exhibit lower similarity scores than those obtained using the method without background recoloring. This is because the background color information in the original image is lost via background recoloring.

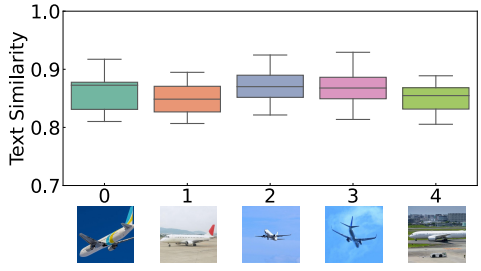


Fig. 7 Text similarity scores of the "airplane" images generated by the proposed method without background recoloring. Note that the stop words are removed in the results. Compared with the results of not removing the stop words, these results exhibit lower similarity scores and larger variances. This is because removing the stop words prevents the text similarity scores from becoming excessively high.

tic features on the generated images are discussed in Sec. 6.4 and Sec. 6.5. Finally, the advantages and effectiveness of the proposed method and scoring procedures are discussed in Sec. 6.6 and Sec. 6.7.

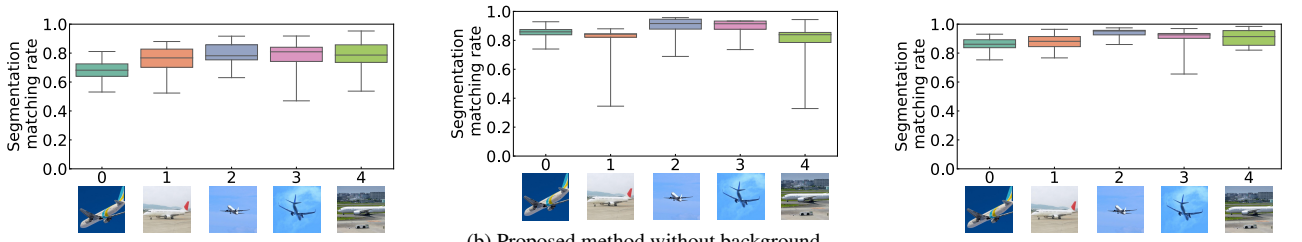
6.1 Data Sizes of Semantic Features

We compared the data size transmitted by the proposed method with that of the raw bit maps, JPEG formatted data, and generated captions of the original images to validate its communication efficiency. Table 1 presents the data sizes for each compression approach. "Uncompressed" denotes the size of the uncompressed bit map of the original images, "JPEG" denotes the average size of the original images in the JPEG format, and "Caption" denotes average size of the captions for each original image, which is used by the simple method employing only the descriptive information of the images. In addition, "Caption + color palette + segmentation array" denotes the average total size of the caption, color palette, and segmentation array for each image, which is the size of semantic features used in the proposed method. The sizes of all elements of the segmentation arrays and captions characters were calculated as 1 byte each. The data size of each color palette was calculated as the number of labels in an image $\times$ 4 bytes: 3 bytes for the RGB values and 1 byte for the label index. Each segmented array was compressed using run-length encoding for further data reduction. Additionally, "Caption + color palette + segmentation array" includes the RGB value of the "background" label.

The results indicate that the proposed methods achieved higher data reduction than the images in JPEG format, thereby underscoring the utility of the proposed method in situations with limited network resources. Additionally, data size comparisons of the semantic features showed that only the caption extracted from the images had a smaller data size than the total size of the multiple semantic features, indicating that the communication efficiency was better when only captions were used. However, the proposed method was superior in terms of color and composition reproduction, as described in the following sections.

6.2 Comparison of Text Similarity Scores of Each Method

Fig. 6 shows the text similarity scores for the captions generated from the original image and 50 images generated using each method. The results indicate the text similarity scores of the five original images of "airplane," and similar trends were observed for the results of other images. Note that stop words were not removed from these results and the original images are shown at the bottom of the figures. The results demonstrate that text similarity was higher when multiple semantic features were used for image generation than when only captions were used for some original images. This is because the proposed method generates images with more similar semantic information, such as the composition of the object and color information, using multiple semantic features. In addition, the method that considers segment



(a) Using only captions. (b) Proposed method without background recoloring. (c) Proposed method with background recoloring.

Fig. 8 Segmentation matching rates of the "airplane" images generated using each method. The original images are shown at the bottom of the graphs. The results of the proposed methods exhibit higher matching rates than those obtained using only captions. Additionally, the proposed method with background recoloring exhibits a higher matching rate than that without background recoloring. This is because background recoloring increases the segment information recognition performance of the image-generation model.

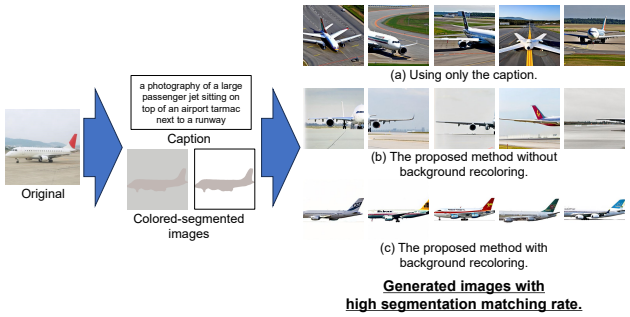


Fig. 9 Top five generated images with the highest segmentation matching rates with the original "airplane" image. Compared to the images generated using the method with background recoloring, the compositions of the images generated using the proposed method without background recoloring are less similar to that of the original image. The image-generation model cannot recognize the segment information using the proposed method without background recoloring owing to the similar RGB values of the target object and background. In such situations, background recoloring can generate images with more similar compositions to the target object.

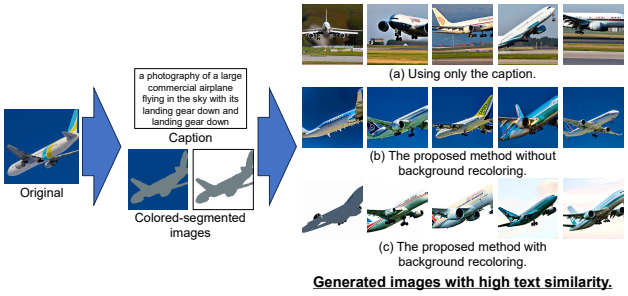


Fig. 10 Top five images with the highest text similarity scores for each method for the original "airplane" image. The proposed method without background recoloring can reconstruct the entire color information of original image, including the background. In contrast, the proposed method with background recoloring results in the loss of background color information.

information (i.e., when the background color of the colored-segmented image is white), exhibited lower text similarity for some images compared to the method that does not consider segment information. This is because, in the method that considers segment information, the color information of the background in the original image is lost, which may affect its text similarity performance.

Fig. 7 shows the text similarity scores for the captions generated from the original image and 50 images generated using the proposed method without background recoloring. Note that the stop words were removed from all the captions in this result. Therefore, this result can be compared to that shown in Fig. 6 (b) based on the presence or absence of stop words. Fig. 6 (b) shows high text similarity scores for all original images, which may be caused by the presence of stop words that make similarity excessively high. In contrast, in Fig. 7, the similarity scores for all images are lower than that in Fig. 6; however, their variances are higher because of the removal of stop words. Therefore, these results suggest that removing stop words is an effective strategy to improve the evaluation accuracy of the descriptive information in images.

6.3 Comparison of Segmentation Matching Rate of Each Method

Fig. 8 shows the segmentation matching rates for the segmentation arrays generated from the original 5 "airplane" images and the 50 images generated using each method. Approximately, the same trends can be observed in the results of the other images, wherein the proposed methods exhibit higher matching rates than using only captions for the image generation. This is because the proposed methods can generate images with similar compositions and positions of the target object as those in the original image using the colored-segmented images.

Among the proposed methods, that without background recoloring exhibits a large variance in the matching rate for some images, whereas that with background recoloring maintains a high matching rate for all images. This is because, as shown in Fig. 9 (b), when the colors of the background and target objects are similar, the image-generation model cannot correctly recognize the object position and composition. By contrast, when the background color was set to white, the generated images had a composition that was more similar to the original image than that generated using the method without background recoloring as shown in Fig. 9 (c).

6.4 Characteristics of Images Generated Using Each Method

Fig. 10 shows the top five images with the highest text similarity scores generated using each method for one "airplane" image. Note that the stop words were not removed from the captions in this result. The images generated using only captions exhibit large composition and color information differences compared with the original image. In contrast, the compositions of the images generated using the proposed methods are similar to those of the original image. These results demonstrate that images with a composition similar to that of the original image can be generated by inputting the multi-modal semantic features contained in the original image into the image-generation model. However, in Fig. 10 (c), the background color information of the original image is lost in the generated images because background recoloring sets a color other than that of the target object to white. This result suggests that in situations where background information is important, inputting all the color information, including that of the background, is advantageous.

However, as shown in Fig. 9 (b) and Fig. 11 (b), when the colors of the background and target object are similar, the composition and position of the object in the generated images deviate from that in the original image. This is because, if the RGB values of the object and background colors are close, the object boundary becomes ambiguous, and the image-generation model cannot recognize the position and composition of the object in the colored-segmented image. In such situations, background recoloring that sets the background color to white is effective for making the object in the colored-segmented image stand out.

6.5 Impacts of Captions on Contents of Generated Images

Fig. 11 shows the top five images with the highest text similarity scores generated using each method for one "person" image. In this figure, the caption generated from the original image contains an unrelated word "dog." This phenomenon is caused due to misidentification by the image captioning algorithm. Although there were no dogs in the original image, some of the generated images contained dog-like objects, and these images exhibited high text similarity scores. To prevent this phenomenon, further improvements in the caption generation algorithm are required.

Fig. 12 shows the semantic features generated from the two "bird" images and examples of images generated using these features. These results indicate that if the original

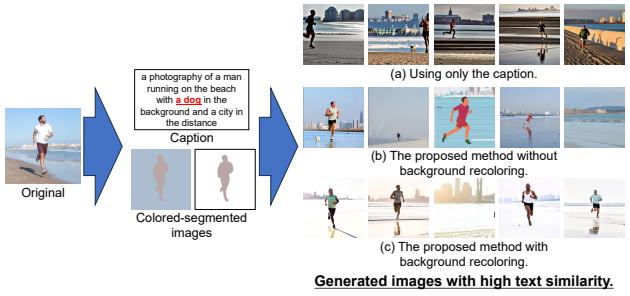


Fig. 11 Top five images with the highest text similarity scores generated by each method for the original "person" image. The generated images with irrelevant objects exhibit high text similarity scores because the caption contains a word about the object that is not included in the original image.

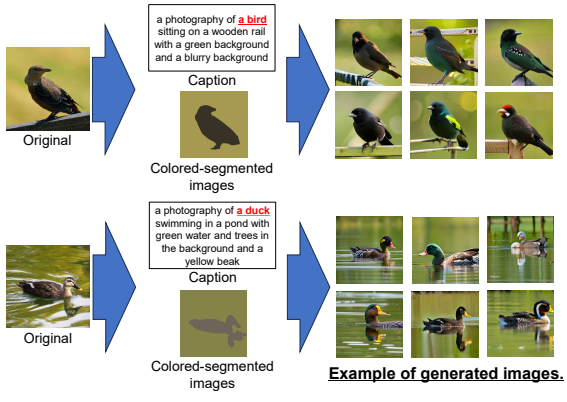


Fig. 12 Semantic features and examples of the images generated using the "bird" images. If the caption contains specific words describing the target object, such as the type of living thing, this content is reflected in the generated images.

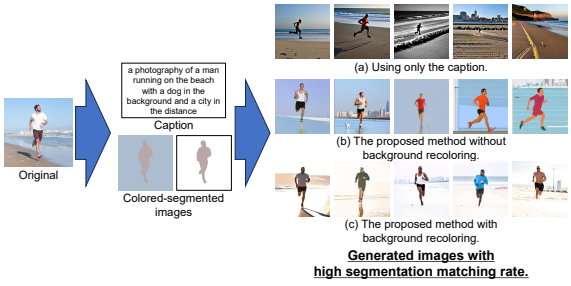


Fig. 13 Top five images with the highest segmentation matching rates generated using each method for the original "person" image. The generated images with similar composition and position of the target object exhibit higher rates compared to the text similarity scores.

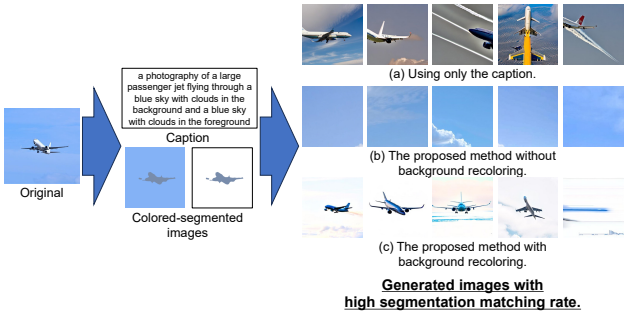
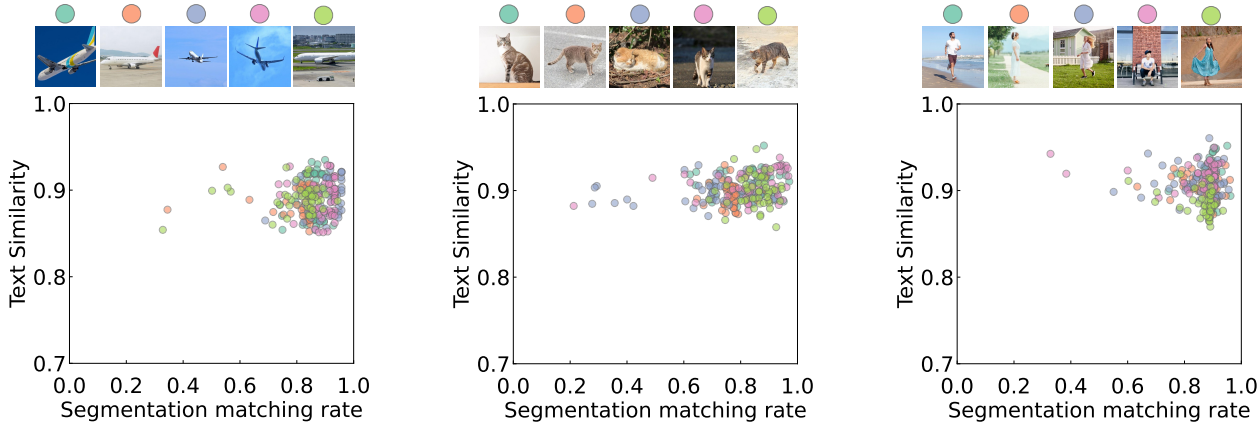


Fig. 14 Top five images with the highest segmentation matching rates generated using each method for the original "airplane" image. The images with no objects generated using the proposed method without background recoloring exhibit high matching rates. Therefore, it is necessary to reconsider the calculation method for the segmentation matching rate in a future study.



(a) Airplane. (b) Cat. (c) Person.

Fig. 15 Scatter plots of text similarity scores and segmentation matching rates of the images generated using proposed method without background recoloring. To prevent outputting images with low scores, it is necessary to select the output result from multiple generated images based on their semantic similarity.

image contains a living thing, the generated caption may contain a description of a specific type of this living thing. For example, in the caption generated from the image shown at the top of Fig. 12, the target object is simply described as a "a bird." By contrast, the caption generated from the image at the bottom of Fig. 12 describes the object using the specific species name "a duck." Consequently, the images generated using the bottom caption contain duck-like birds. This result suggests that the content of the generated images is affected by the granularity of the descriptive information. Therefore, it can be possible to adjust the granularity of the descriptive information included in the caption based on the situation and application.

6.6 Advantages and Disadvantages of Evaluating Text Similarity and Segmentation Matching Rate

Fig. 11 and Fig. 13 show the top five generated images with the highest text similarity scores and segmentation matching rates generated by each method. Note that the images in these figures are generated based on the same original "person" image. As shown in Fig. 11, because the caption contains the word "dog," the generated images that contain dog-like objects exhibit high text similarity scores. Additionally, because the captions include the words "beach" and "city," the generated images containing these elements in the background also exhibit high text similarity scores. These results indicate that the scoring procedure for text similarity is effective for evaluating the type of target object and broad background information.

As shown in Fig. 13, the generated images with high segmentation matching rates are more similar to the original image in terms of the composition and position of the target object than those with high text similarity scores. Therefore, these results indicate that the segmentation matching rate is effective for evaluating the composition and position of the target object in the generated images.

Fig. 14 shows the top five generated images with the highest segmentation matching rates for the "airplane" image by each method. As shown in Fig. 14 (b), if the target object in the original image is too small, the images generated without the target object exhibit higher scores than the other images. This is because the number of pixels labeled "airplane" is considerably lower than the number of pixels; therefore, the matching rate of the images generated without any object where all pixels are labeled "background" is higher. To address this issue, future studies should develop a novel calculation method for the segmentation matching rate.

6.7 Effectiveness of the Proposed Methods and Scoring Procedures

Fig. 15 shows the scatter plots of the text similarity scores and segmentation matching rates for the images generated by the proposed method without background recoloring. Note that the stop words were not removed in the text similarity evaluations, and the color of each point indicates the type of semantic feature in the original image used for image generation. Some of the generated images exhibit significantly lower scores across both metrics than the other images. This implies that if only a single image is reconstructed as an output in an image-generation-based transmission, the image can be semantically different from the original image. Therefore, as employed in the proposed method, it is essential to select an image that is similar to the original image from multiple generated images based on the proposed scoring procedure.

Based on these results, the proposed method realizes both data reduction and image generation using semantic information that is more similar to the original image than using only a single semantic feature of the image. In addition, these results indicate that the text similarity scores and segmentation matching rates can quantify the semantic similarity of the generated images in terms of descriptive and segment information, respectively. The proposed method

without background recoloring can realize image transmission while maintaining the background color information around the target object. However, if the colors of the target object and the background are similar, the image-generation model cannot correctly recognize the segment information of the colored-segmented image. This issue can be addressed by applying background recoloring that changes the colors of pixels labeled "background" to white.

7. Conclusion

This paper proposed an image-generation-based transmission method. In the proposed method, a transmitter extracts multi-modal semantic features from an image and transmits them to a receiver, thereby significantly reducing the amount of transmitted data compared to traditional methods. After receiving the data, the receiver generates multiple images using an image-generation model and selects an output image based on the semantic similarity between the original and generated images. Therefore, the proposed method can realize both a significant reduction in the transmitted data size and reconstruction of the image with the semantic information required by the receiver. The evaluation results validated the ability of the proposed method to significantly reduce the data size compared with typical image compression algorithms.

In the proposed method, the receiver must compare and evaluate the similarity between the original and generated images using only the received data when selecting the output. Therefore, this study focused on descriptive and segment features of images and proposed two scoring methods for comparing them between the original and generated images.

An experiment was conducted to verify the effectiveness of the proposed image transmission method and scoring procedures. The results indicated that the proposed method using multi-modal semantic information can generate images that are more similar to the original images than those using only a single piece of information. In particular, the proposed method exhibits good reconstruction performance for the composition and position of the objects in the original image because a colored-segmented image is input into the image-generation model as segment and color information. In addition, the proposed scoring procedures can quantify the semantic similarity between images. The text similarity scores can be used to compare the types of target objects and rough background information of images. In contrast, the segmentation matching rates can be used to compare the composition and position of the target object in the images. However, we found that if the target object in the image was too small, the generated images with no objects exhibited high segmentation matching rates.

In a future work, we will improve the scoring procedures used for evaluating the semantic similarity between images at the receiver end. In the proposed scoring procedure for the segmentation matching rate, images generated with no objects had a high matching rate when the target object in the original image was too small. This is because the proposed procedure calculates the matching rate based on the number of pixels including those labeled as "background." Therefore, it is necessary to develop a calculation method that considers only the pixels labeled as "target object." Moreover, we plan to develop an image reconstruction method based on the evaluation results of semantic similarity between the images. In the proposed method, the receiver selects the output result from the generated images based on semantic similarity. However, if all generated images have low similarity scores, the receiver cannot select the result using the semantic information required by its application. To address this issue, it is necessary to develop a method to regenerate images or, if necessary, request the sender to transmit the semantic features again.

Acknowledgments

This work was partly supported by NICT (JPJ012368C01101).

References

[1] M.F. Ali, D.N.K. Jayakody, and Y. Li, "Recent trends in underwater visible light communication (UVLC) systems," IEEE Access, vol.10, pp.22169–22225, Feb. 2022, DOI:10.1109/ACCESS.2022.3150093.

[2] D.N. Sandeep and V. Kumar, "Review on clustering, coverage and connectivity in underwater wireless sensor networks: A communication techniques perspective," IEEE Access, vol.5, pp.11176–11199, June 2017, DOI:10.1109/ACCESS.2017.2713640.

[3] O. Kodheli, E. Lagunas, N. Maturo, S.K. Sharma, B. Shankar, J.F.M. Montoya, J.C.M. Duncan, D. Spano, S. Chatzinotas, S. Kisseleff, J. Querol, L. Lei, T.X. Vu, and G. Goussetis, "Satellite communications in the new space era: A survey and future challenges," IEEE Commun. Surveys & Tutorials, vol.23, no.1, pp.70–109, Firstquarter 2021, DOI:10.1109/COMST.2020.3028247.

[4] M.A. Ullah, K. Mikhaylov, and H. Alves, "Enabling mMTC in remote areas: LoRaWAN and LEO satellite integration for offshore wind farm monitoring," IEEE Trans. on Industrial Informatics, vol.18, no.6, pp.3744–3753, June 2022, DOI:10.1109/TII.2021.3112386.

[5] Y. Zheng, Y. Zhao, W. Liu, S. Liu, and R. Yao, "An intelligent wireless system for field ecology monitoring and forest fire warning," Sensors, vol.18, no.12, pp.3744–3753, Dec. 2018, DOI:10.3390/s18124457.

[6] "Cisco Annual Internet Report (2018–2023) White Paper." https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.pdf. (Accessed on 01/22/2024).

[7] "Information and Communications in Japan 2022." https://www.soumu.go.jp/johotsusintokei/whitepaper/eng/WP2022/2022-index.html. (Accessed on 01/22/2024).

[8] "Ericsson Mobility Report June 2023." https://www.ericsson.com/49dd9d/assets/local/reports-papers/mobility-report/documents/2023/ericsson-mobility-report-june-2023.pdf. (Accessed on 01/22/2024).

[9] A. Hussain, A. Al-Fayadh, and N. Radi, "Image compression techniques: A survey in lossless and lossy algorithms," Neurocomputing, vol.300, pp.44–69, July 2018, DOI:10.1016/j.neucom.2018.02.094.

[10] G.K. Wallace, "The JPEG still picture compression standard," Communications of the ACM, vol.34, no.4, pp.30–44, April 1991, DOI:10.1145/103085.103089.

[11] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG 2000 still image compression standard," IEEE Signal Processing Magazine, vol.18, no.5, pp.36–58, Sept. 2001, DOI:10.1109/79.952804.

[12] S. Jamil, M.J. Piran, M. Rahman, and O.J. Kwon, "Learning-driven lossy image compression: A comprehensive survey," Engineering App. of Artificial Intel., vol.123, pp.1–17, Aug. 2023, DOI:10.1016/j.engappai.2023.106361.
[13] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wang, "Image and video compression with neural networks: A review," IEEE Trans. on Circuits and Systems for Video Tech., vol.30, no.6, pp.1683–1698, June 2020, DOI:10.1109/TCSVT.2019.2910119.
[14] E. Bourtsoulatze, D.B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," 2019 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2019), Brighton, UK, pp.4774–4778, May 2019.
[15] E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," IEEE Trans. on Cognitive Commun. and Netw., vol.5, no.3, pp.567–579, Sept. 2019, DOI:10.1109/TCCN.2019.2919300.
[16] X. Luo, H.H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," IEEE Wireless Commun., vol.29, no.1, pp.210–219, Feb. 2022, DOI:10.1109/MWC.101.2100269.
[17] D. Gündüz, Z. Qin, I.E. Aguerri, H.S. Dhillon, Z. Yang, A. Yener, K.K. Wong, and C.B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," IEEE Journal on Selected Areas in Commun., vol.41, no.1, pp.5–41, Jan. 2023, DOI:10.1109/JSAC.2022.3223408.
[18] Z. Qin, X. Tao, J. Lu, W. Tong, and G.Y. Li, "Semantic communications: Principles and challenges," arXiv preprint arXiv:2201.01389, pp.1–32, June 2021.
[19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR 2022), New Orleans, United States, pp.10684–10695, June 2022.
[20] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," 38th Int. Conf. on Machine Learning (ICML 2021), pp.8821–8831, July 2021.
[21] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," arXiv preprint arXiv:2204.06125, pp.1–27, April 2022.
[22] A. Kodaira, C. Xu, T. Hazama, T. Yoshimoto, K. Ohno, S. Mitsuhori, S. Sugano, H. Cho, Z. Liu, and K. Keutzer, "StreamDiffusion: A pipeline-level solution for real-time interactive generation," arXiv preprint arXiv:2312.12491, pp.1–13, Dec. 2023.
[23] H. Nam, J. Park, J. Choi, M. Bennis, and S.L. Kim, "Language-oriented communication with semantic coding and knowledge distillation for text-to-image generation," arXiv preprint arXiv:2309.11127, pp.1–5, Sept. 2023.
[24] E. Hosonuma, T. Yamazaki, T. Miyoshi, A. Taya, Y. Nishiyama, and K. Sezaki, "Exploiting spatial and descriptive information for generative compression," IEEE Consumer Commun. and Netw. Conf. (CCNC 2024), Jan. 2024.
[25] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, United States, pp.586–595, June 2018.
[26] D. Tellez, G. Litjens, J. van der Laak, and F. Ciompi, "Neural image compression for gigapixel histopathology image analysis," IEEE Trans. on Pattern Analysis and Machine Intel., vol.43, no.2, pp.567–578, Feb. 2021, DOI:10.1109/TPAMI.2019.2936841.
[27] L. Cavigelli, P. Hager, and L. Benini, "CAS-CNN: A deep convolutional neural network for image compression artifact suppression," 2017 Int. Joint Conf. on Neural Networks (IJCNN 2017), Anchorage, United Staes, pp.752–759, May 2017.
[28] M. Li, W. Zuo, S. Gu, D. Zhao, and D. Zhang, "Learning convolutional networks for content-weighted image compression," IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2018), Salt Lake City, United States, pp.3214–3223, June 2018.
[29] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Deep convolutional autoencoder-based lossy image compression," 2018 Picture Coding Symposium (PCS 2018), San Francisco, United Stetes, pp.253–257, June 2018.
[30] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," arXiv preprint arXiv:1703.00395, pp.1–19, March 2017.
[31] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," 39th Int. Conf. on Machine Learning (ICML 2022), Baltimore, United States, pp.12888–12900, July 2022.
[32] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 38th Int. Conf. on Machine Learning (ICML 2021), pp.8748–8763, July 2021.
[33] L.C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, pp.1–14, Dec. 2017.
[34] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE Trans. on Pattern Analysis and Machine Intel., vol.39, no.12, pp.2481–2495, Dec. 2017, DOI:10.1109/TPAMI.2016.2644615.
[35] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "CoCa: Contrastive captioners are image-text foundation models," arXiv preprint, arXiv:2205.01917, pp.1–19, June 2022.
[36] T. Zhang, V. Kishore, F. Wu, K.Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with bert," 2020 Int. Conf. on Learning Representations (ICLR 2020), pp.1–43, April 2020.
[37] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, pp.1–16, May 2019.
[38] "PhotoAC." `https://www.photo-ac.com/`. (Accessed on 01/22/2024).

Eri Hosonuma received the B.E. and M.S. degrees in electronic information systems from Shibaura Institute of Technology, Tokyo, Japan, in 2020 and 2022, respectively. She is presently a doctoral course student at the Department of Socio-Cultural Environmental Studies, The University of Tokyo, Tokyo, Japan.

Taku Yamazaki received the B.E. and M.S. degrees in electronic information systems from Shibaura Institute of Technology, Tokyo, Japan, in 2012 and 2014, respectively. He received the D.E. degree in computer science and communications engineering from Waseda University, Tokyo, Japan, in 2017. He is presently an associate professor at Department of Electronic Information Systems, College of Systems Engineering and Science, Shibaura Institute of Technology, Saitama, Japan. His research interests include wireless networks, internet of things, and network security.

Takumi Miyoshi received his B.Eng., M.Eng., and Ph.D. degrees in electronic engineering from the University of Tokyo, Japan, in 1994, 1996, and 1999, respectively. He started his career as a research associate in Waseda University from 1999 to 2001, and is presently a professor at Department of Electronic Information Systems, College of Systems Engineering and Science, Shibaura Institute of Technology, Saitama, Japan. He is also a research fellow in Institute of Industrial Science, the University of Tokyo, Tokyo, Japan. He was a visiting scholar in Laboratoire d'Informatique de Paris 6 (LIP6), Sorbonne Université, Paris, France, from 2010 to 2011. His research interests include overlay networks, location-based services, and mobile ad hoc and sensor networks.

Akihito Taya received the B.E. degree in electrical and electronic engineering from Kyoto University, Kyoto, Japan in 2011, and the master and Ph.D. degree in Informatics from Kyoto University in 2013 and 2019, respectively. From 2013 to 2017, he joined Hitachi, Ltd., where he participated in the development of computer clusters. From 2019 to 2022, he was an assistant professor of Aoyama Gakuin University. He has been an assistant professor of The University of Tokyo, since 2022. He received the IEEE VTS Japan Young Researcher's Encouragement Award and the IEICE Young Researcher's Award in 2012 and 2018, respectively. His current research interests include distributed machine learning and human activity and emotion recognition using sensor networks. He is a member of the IEEE, ACM, IEICE and IPSJ.

Yuuki Nishiyama is an Assistant Professor at the Center for Spatial Information Science in the University of Tokyo. He obtained M.S.(2014) in Media and Governance from Keio University, and Ph.D. in Media and Governance (2017) from Keio University, respectively. He had worked at Keio University in Japan and the University of Oulu in Finland, as a post-doctoral researcher respectively. He started work at the Institute of Industrial Science in the University of Tokyo as a Research Associate in 2019, and has held his current position since 2022. His current research interests include ubiquitous computing, contex-aware systems, human behavior change, and human ability augmentation. He is a member of ACM, IEEE, and Information Processing Society of Japan (IPSJ).

Kaoru Sezaki is the director of Center for Spatial Information Science at the University of Tokyo and co-appointed as professor of Institute of Industrial Science, the University of Tokyo. He is a steering member of e-Health Technical Committee COMSOC. He has been general chair and TPC Chair of many IEEE international conferences. He also served as Treasurer of IEEE Tokyo Section as well as that of Japan Council from 2003 to 2004.He received B.Eng., M.Eng., and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 1984, 1986, and 1989, respectively, all in Electrical Engineering. Since 1989, he has been with the University of Tokyo. He was a Visiting Researcher at University of California at San Diego in 1996. His research interests include e-Health, sensor networks, IoT, and urban computing.