IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 April 2023, accepted 19 April 2023, date of publication 24 April 2023, date of current version 4 May 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3269847*

RESEARCH ARTICLE

# Text-Guided Image Manipulation via Generative Adversarial Network With Referring Image Segmentation-Based Guidance

**YUTO WATANABE[1], (Graduate Student Member, IEEE), REN TOGO[2], (Member, IEEE), KEISUKE MAEDA[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, Hokkaido University, Sapporo 060-0814, Japan

Corresponding author: Miki Haseyama (mhaseyama@lmd.ist.hokudai.ac.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP21H03456.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

**ABSTRACT** This study proposes a novel text-guided image manipulation method that introduces referring image segmentation into a generative adversarial network. The proposed text-guided image manipulation method aims to manipulate images containing multiple objects while preserving text-unrelated regions. The proposed method assigns the task of distinguishing between text-related and unrelated regions in an image to segmentation guidance based on referring image segmentation. With this architecture, the adversarial generative network can focus on generating new attributes according to the text description and reconstructing text-unrelated regions. For the challenging input images with multiple objects, the experimental results demonstrate that the proposed method outperforms conventional methods in terms of image manipulation precision.

**INDEX TERMS** Text-guided image manipulation, text-to-image synthesis, generative adversarial network, referring image segmentation.

## I. INTRODUCTION

The goal of image manipulation is to semantically alter the appearance of an image, e.g., color and texture [1], [2] or shape and type [3], to satisfy a user's requirements. It can facilitate various applications in computer-aided design, architecture, video games, and image editing. To reduce user burden during image manipulation, research on automation has attracted significant attention in the computer vision field.

With the emergence of generative models, e.g., the variational autoencoder (VAE) [4] and generative adversarial networks (GANs) [5], research on automated image manipulation is advancing. Automatic image manipulation can manipulate images without complex operations and has several approaches, including image inpainting [6], [7], image colorization [2], [8], style transfer [9], [10], and domain or attribute transformation [11], [12]. Image inpainting is the process of compositing alternative content into missing parts of an image and generating an image that is visually and semantically natural. Style transfer, image colorization, and domain or attribute transformation are classified as image-to-image translation tasks, and each model is designed to perform a particular conversion to an entire image. However, such methods are limited and have difficulties achieving high-precision image manipulation that effectively reflects a user's requirements.

To enable users to specify where and how to manipulate images, recent text-guided image manipulation approaches [13], [14], [15], [16] use natural language descriptions. These methods are designed to semantically manipulate

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

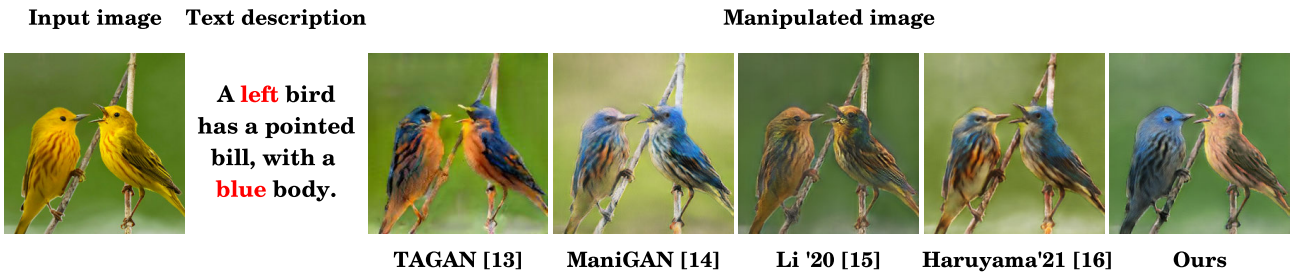


**FIGURE 1.** Examples of text-guided image manipulation. The red words in the text description express where and how to manipulate the image. Previous methods [13], [14], [15], [16] tend to fail to suppress the manipulation of the text-unrelated object (i.e., *a right bird*), and the colors of the left bird are a mixture of blue and yellow. The proposed method suppresses the manipulation of text-unrelated regions, and the color of the left bird is blue.

images according to text descriptions that indicate the image attributes to manipulate, such as texture, color, and backgrounds. These methods use GANs [5] as their base architecture and facilitate the generation of realistic manipulated images.

Previous text-guided image manipulation methods use mechanisms that can select the text-related region and attempt to semantically combine the text description with the region. However, as depicted in Fig. 1, such methods are limited in terms of reflecting the red words in the text description that express where and how to manipulate the image; thus, these methods achieve insufficient image manipulation. Existing methods manipulate undesired parts such as the *right bird* and backgrounds in an image, even though the text description does not expect these parts to be manipulated. In addition, due to the incorrect attention, an input image is partially reconstructed for the text-related region such as the *left bird*, and the colors of the region are an unnatural mixture of blue and yellow. Specifically, previous methods suffer from two problems. (1) Existing attention mechanisms are limited in that they explicitly correlate words in the text description, e.g., *left* between object information in an image. It is challenging to manipulate only the object specified by the text description among multiple objects of the same type in a single image. (2) GANs for image manipulation are expected to perform multiple actions. For a GAN, the simple reconstruction of an image is a relatively easy task; however, this leads to neglecting the matching of the text-related region in the image with the text description.

Thus, in this study, we propose a text-guided image manipulation method. The proposed method exploits the advantages of segmentation guidance based on referring image segmentation [17] and improves image manipulation precision. Referring image segmentation is designed to generate a segmentation mask that can extract the text-related region from an input image. Although the image can be reconstructed by applying this mask to it directly, there is no consistency between the reconstructed and manipulated regions. Thus, as segmentation guidance, we multiply the segmentation mask by the image features to obtain distinguished image features. We also leverage a recently proposed contrastive language-image pretraining (CLIP) model [18] as the loss function to reflect fine-grained words corresponding to visual attributes such as color and texture, in the manipulated image. The proposed method uses this segmentation guidance and explicitly identifies regions that should be manipulated or preserved in the GAN. Thus, this network can focus on generating new attributes expressed by the given text description and reconstructing input images in text-unrelated regions. Our primary contributions are summarized as follows.

- We enhance the visual and semantic consistency between the text and image in the mask regions by combining the feature map of the segmentation mask with the text description, according to the CLIP loss.
- Image manipulation is realized without the effect of the reconstruction of the input image in text-unrelated regions by distinguishing the features of the input image between the text-related and unrelated regions.

The proposed method can overcome the limitations of image manipulation in situations involving only one object that previous studies [13], [14], [15], [16] have tackled. The models in these previous studies could not handle image manipulation in complex situations involving multiple objects in the inference if they were trained on datasets consisting of images with a single object. Even if the proposed method is trained on such a dataset, it achieves high-precision image manipulation in the inference, regardless of the number of objects in the input image. This is accomplished by our network learning to associate a text with regions of an image extracted by the segmentation guidance.

The remainder of this paper is organized as follows. In Section II, we introduce related work on multimodal analyses of image and text representations. The proposed GAN based on the segmentation guidance for image manipulation is explained in Section III. In Section IV, we discuss our experimental results and verify the effectiveness of segmentation guidance. Finally, conclusions are provided in Section V.

Note that this paper is an extension of a previously published paper [19]. The main difference between our previous study and this one is the introduction of the CLIP loss. With this novelty, the proposed method is expected to improve the visual and semantic consistency between the

given text descriptions and the manipulated images. More extensive experiments than those in the previous study show that the proposed method can generate manipulated images that richly represent the given text description.

## II. RELATED WORK

### A. APPLICATION OF IMAGE GENERATION

With the emergence of GANs [5], image generation research has made significant advances, and various image generation tasks, including text-to-image synthesis [20], [21], [22], [23], [24] and image-to-image translation [2], [6], [7], [8], [9], [10], have been proposed.

Text-to-image synthesis is a cross-modal image generation task conditioned by a natural language text description [20], [21], [22], [23] and scene graphs [24]. Reed et al. [20], for the first time, tackled the task of generating a 64 $\times$ 64 natural image from text features based on a deep convolutional GAN architecture [25]. Zhang et al. proposed StackGAN [21] to progressively increase the resolution of generated images by stacking multiple GANs, and Xu et al. [23] and Li et al. [22] proposed an attention-driven generator and word-level discriminator for fine-grained text-to-image synthesis at a word-level. In addition, to handle more complex cases, e.g., images with many objects and relationships, previous studies have proposed text-to-image synthesis methods conditioned by scene graphs [24]. Such methods generate a scene layout based on predicted segmentation masks and bounding boxes from input scene graphs and convert the layout into a realistic image using a GAN. By generating the scene layout according to the scene graphs, it is possible to explicitly infer multiple objects and their relationships in the images.

The recent studies [26], [27], [28] have reported that image-to-image translation has practical applications such as image quality enhancement, image retrieval, and identity anonymization. With this trend, in the research field of image-to-image translation, various approaches represented by image inpainting [6], [7], image colorization [2], [8], and style transfer [9], [10], have been proposed. These models have demonstrated effectiveness in specific tasks such as filling holes in an image, reflecting the style of well-known artworks in real images, and predicting the color version of black-and-white images. However, some models [6], [7] also require segmentation masks that specify the region of holes in an image to users, and others [2], [8], [9], [10] colorize and translate the entire image regardless of the user's requirements, e.g., manipulation of specific regions or colors.

### B. REFERRING IMAGE SEGMENTATION

Recently, several referring image segmentation approaches have been proposed to extract the object region in the input image corresponding to a text description as multimodal analyses of image and text representations. For example, Hu et al. [29] handled the referring segmentation task and generated a segmentation mask by directly concatenating multimodal features extracted using a convolutional neural network and a long short-term memory (LSTM) network [30]. To analyze word-to-image interactions, a multimodal LSTM [31] sequentially integrates visual and text features in multiple time steps. Li et al. [32] proposed a method for integrating multilevel visual features that can recurrently refine the local details of the segmentation mask. In addition, Huang et al. [17] proposed the state-of-the-art CMPC-Refseg referring image segmentation method, which has demonstrated excellent results. The CMPC-Refseg method progressively perceives correct objects using a cross-modal progressive comprehension (CMPC) and text-guided feature exchange (TGFE) modules. The CMPC module first recognizes all objects assumed from nouns in the text description and then emphasizes the correct object by multimodal graph reasoning. Based on the concept of integrating multilevel visual features, the TGFE module enables the refinement of the results from the CMPC module. In text-guided image manipulation tasks, segmentation masks generated using CMPC-Refseg method may provide promising benefits in terms of focusing on the text-related region.

### C. TEXT-GUIDED IMAGE MANIPULATION

By applying the concept of controlling image generation with natural language descriptions proposed in the text-to-image synthesis task, text-guided image manipulation [13], [14], [15], [16] achieves more user-friendly image manipulation. Such methods are designed to perform semantic manipulation according to text descriptions and preservation in text-unrelated regions. For example, Dong et al. [33] adopted a novel structure primarily based on a GAN and tackled the task of generating the manipulated image conditioned by the given image and text description. In addition, Nam et al. [13] proposed a text-adaptive discriminator that monitors the extent to which a text description is reflected in the image at the word level and obtains a generator that can generate fine-grained visual attributes. Li et al. [14] adopted a multi-stage architecture of multiple GANs and enabled the generation of 256 $\times$ 256 images to produce high-quality manipulated images reflecting text descriptions. Haruyama et al. In addition, [16] focused on differences in representation ability between images and text descriptions and achieved image manipulation that suppresses background manipulation.

In recent years, the integration of vision and language has been emphasized in the computer vision field, and this multimodal analysis can realize the development of user-friendly image manipulation techniques. CLIP [18] is an image classification model pretrained on 400 million image–text pairs. As a result, the model provides generalized text and image representation capabilities without overfitting to a specific dataset and has achieved excellent results in the zero-shot task. Benefiting from the capabilities of this model, text-guided image manipulation is expected to maximize the expressive capabilities of text descriptions and effectively

reflect the expression in the corresponding manipulated images.

While high-quality image manipulation can contribute to several fields, how to ensure the trust and credibility of data is an urgent problem to be solved. For this problem, a novel approach [34] for forgery detection based on GANs analyzed the traces that the forgery method may leave on the tampered data and constructed a multi-scale forgery trace generation system. Since data manipulation can have dangerous aspects depending on how it is used, the research on text-guided image manipulation needs to be cooperatively conducted with the research mentioned above.

## III. PROPOSED IMAGE MANIPULATION METHOD

We design a model that can manipulate input image $I$ according to text description $T$ while preserving text-unrelated regions to generate manipulated image $I'$. As depicted in Fig. 2, the proposed GAN includes a guided text-image affine combination module (GACM) and a guided detail correction module (GDCM). We extend the original ACM and DCM [14] by introducing segmentation guidance to distinguish the image between text-related and unrelated regions such that each module can focus on image manipulation rather than region selection. To generate manipulated image $I'$, the proposed method refines multimodal features by fusing images and text descriptions by passing through two units: the main unit with the multi-stage architecture and the GDCM unit. Here, three generators $G_i$ $(i = 1, 2, 3)$ in the main unit take hidden features $\boldsymbol{X}_i$ as inputs and generate images gradually in small-to-large scales, i.e., 64 × 64, 128 × 128, and 256 × 256 pixels. Note that $i$ denotes the order of each stage in the main unit.

### A. GAN ARCHITECTURE WITH REFERRING IMAGE SEGMENTATION

Fig. 2 depicts the structure of the proposed GAN with the GACM and GDCM. To make the GACM and GDCM function with segmentation guidance, we use a segmentation mask $S$ obtained by referring image segmentation. To provide segmentation guidance for the hidden and image features in the GACM and GDCM, the segmentation mask $S$ is transformed into feature maps for compatibility with these features. We use a feature transformation module (FTM) and generate feature maps by resizing the width and height of $S$ and replicating the same along the channel dimension. We expect these feature maps to provide segmentation guidance for the two modules and instruct the network about the target regions to manipulate or reconstruct. Specifically, in the GACM, segmentation guidance provides the module with information that can identify the region, where the text description $T$ is to be embedded, and the module outputs features useful for manipulating only the text-related region. Each output feature from the GACM is input into the corresponding generators $G_i$, which produce temporarily generated images containing text information in only the text-related regions. In addition, in the GDCM, image features of an input image are distinguished into text-related and unrelated regions according to the segmentation guidance. These features are used to prompt the network to reconstruct the contents of the input image $I$ in text-unrelated regions while retaining new attributes acquired in the main unit. Finally, we acquire the manipulated image $I'$ from the generator $G_{\mathrm{GDCM}}$. In the following, we describe the segmentation guided modules and their objective functions.

### B. MANIPULATION OF TEXT-RELATED REGIONS BASED ON GACM

To embed text information in the target region to be manipulated, the GACM performs the process according to the segmentation guidance. As depicted in Fig. 2 (a), the GACM takes two inputs, i.e., feature maps $\boldsymbol{S}_{\mathrm{GACM}} \in \mathbb{R}^{256\times17\times17}$ using the FTM and hidden features $\boldsymbol{X}_i \in \mathbb{R}^{32\times H_i\times W_i}$ calculated from the previous stage of the GACM or $\boldsymbol{X}_{\mathrm{init}} \in \mathbb{R}^{32\times64\times64}$. To obtain $\boldsymbol{X}_{\mathrm{init}}$, we encode the text description $T$ into global text features using a pretrained bi-directional LSTM [23] and reshape them with neural networks introduced in [14]. Note that $H_1$ and $W_1$ are 64 pixels, $H_2$ and $W_2$ are 128 pixels, and $H_3$ and $W_3$ are 256 pixels. To make the size of the feature maps $\boldsymbol{S}_{\mathrm{GACM}}$ equal to $\boldsymbol{X}_i$, we process the map using two convolutional layers and acquire $\boldsymbol{W}_{i+1}(\boldsymbol{S}_{\mathrm{GACM}})$ encoding the text-related content and $\boldsymbol{B}_{i+1}(\boldsymbol{S}_{\mathrm{GACM}})$ encoding text-unrelated contents. We calculate the hidden features $\boldsymbol{X}_{i+1} \in \mathbb{R}^{32\times H_{i+1}\times W_{i+1}}$, embedding the text features in the region to be manipulated according to the segmentation guidance as follows:

$$\boldsymbol{X}_{i+1} = \boldsymbol{X}_i \odot \boldsymbol{W}_{i+1}(\boldsymbol{S}_{\mathrm{GACM}}) + \boldsymbol{B}_{i+1}(\boldsymbol{S}_{\mathrm{GACM}}), \quad (1)$$

where $\odot$ denotes the Hadamard product. The main unit of the proposed method uses a multi-stage architecture comprising GACMs to gradually expand the size of the hidden features $\boldsymbol{X}_{i+1}$ by upsampling blocks. In the GACM, we let segmentation guidance handle the attention to text-related regions and allow the network to focus on gradually acquiring fine-grained visual attributes in each stage. As a result, the proposed method realizes a model reflecting the text description $T$ in only the text-related region.

### C. RECONSTRUCTION OF TEXT-UNRELATED REGIONS BASED ON GDCM

The temporarily manipulated image produced by $G_3$ in the main unit acquires rich visual attributes according to the text description in the text-related region. To further highlight the detailed content and recover text-unrelated regions lost in the main unit, the GDCM performs the process according to the segmentation guidance. Here, as depicted in Fig. 2 (b), we first apply the pretrained VGG16 network [35] and acquire the visual features $\boldsymbol{Q} \in \mathbb{R}^{32\times256\times256}$ of the input image $I$ from the ReLU layer in the second convolutional block (also known as *relu2_2*). However, the visual features $\boldsymbol{Q}$ contain too many content details (e.g., color, texture, and edge information), which makes the generator simply reconstruct the input image and potentially lose rich visual attributes in

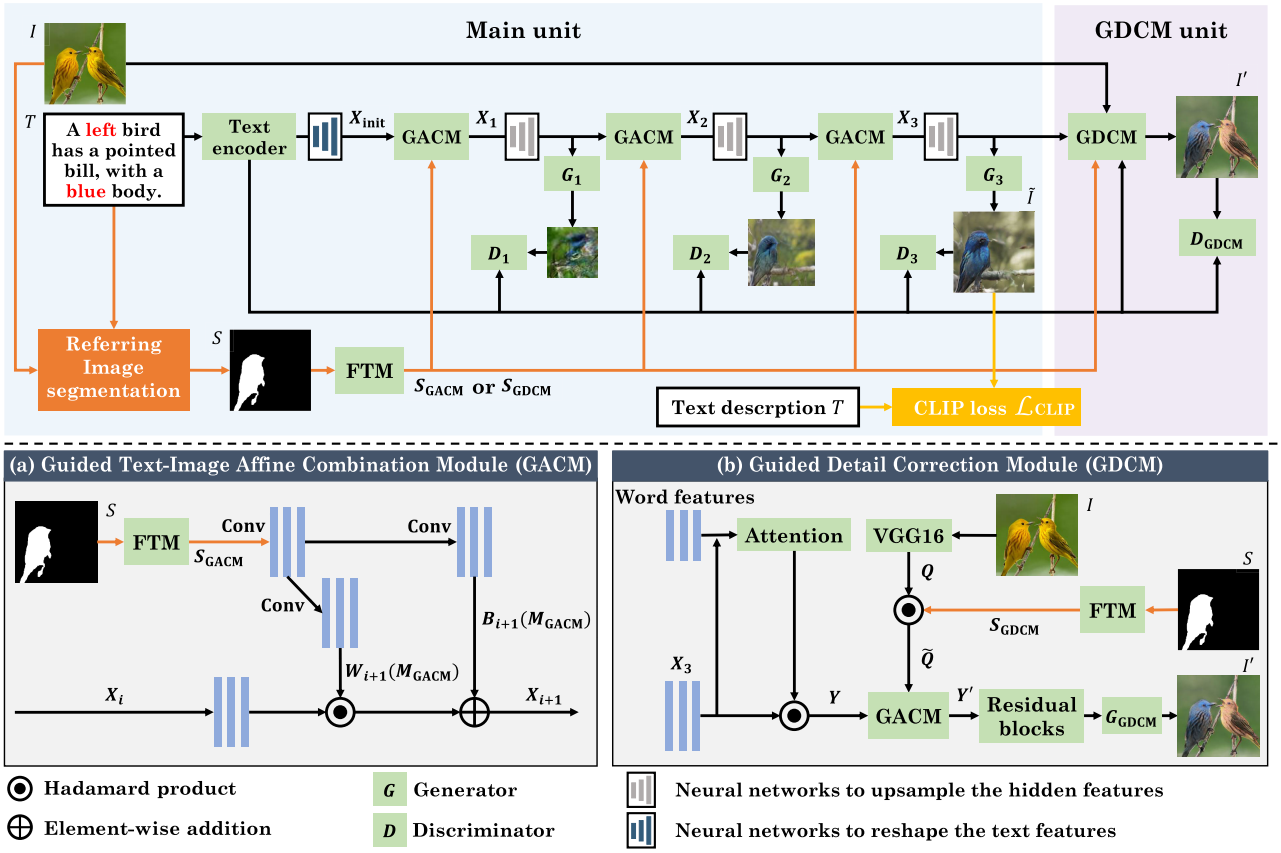


**FIGURE 2.** Top: Structure of the proposed GAN for text-guided image manipulation. We introduce feature maps $S_{\mathrm{GACM}}$ and $S_{\mathrm{GDCM}}$ output from the feature transformation module (FTM) into the network as segmentation guidance. Bottom: Details of segmentation-guided modules: (a) guided text-image affine combination module (GACM) and (b) guided detail correction module (GDCM).

the text-related region acquired by the main unit. To address this problem, we utilize inverted feature maps $\boldsymbol{S}_{\mathrm{GDCM}} \in \mathbb{R}^{32\times256\times256}$ by FTM and acquire distinguished features $\tilde{\boldsymbol{Q}} \in \mathbb{R}^{32\times256\times256}$ by concatenating visual features $\boldsymbol{Q}$ with feature maps $\boldsymbol{S}_{\mathrm{GDCM}}$. By letting segmentation guidance handle the selection of regions that preserve the attributes acquired in the main unit, our model can focus exclusively on reconstructing the input image $I$ in text-unrelated regions.

To modify the hidden features $\boldsymbol{X}_3 \in \mathbb{R}^{32\times256\times256}$ output from the main unit, we use features $\tilde{\boldsymbol{Q}}$ and the word features of the text description $T$. Here, using spatial and channel-wise attention modules [22], [23] (i.e., "Attention" in Fig. 2 (b)) based on words, we generate intermediate features $\boldsymbol{Y} \in \mathbb{R}^{32\times256\times256}$ by multiplying the hidden features $\boldsymbol{X}_3$ with the spatial and channel-wise attention features of the same size as $\boldsymbol{X}_3$, following [14]. By reusing the capabilities of the GACM, we associate $\boldsymbol{Y}$ with $\tilde{\boldsymbol{Q}}$ in the same manner as Eq. (1), where $\boldsymbol{Y}$ and $\tilde{\boldsymbol{Q}}$ correspond to $\boldsymbol{X}_i$ and $\boldsymbol{S}_{\mathrm{GACM}}$, respectively, and generate features $\boldsymbol{Y}' \in \mathbb{R}^{32\times256\times256}$. To refine the features, we pass $\boldsymbol{Y}'$ through a residual block [14] containing multiple convolutional layers. The residual block is beneficial for stabilizing the learning of networks with deep layers. Finally, the generator $G_{\mathrm{GDCM}}$ in the GDCM generates the final manipulated image $I'$ from the refined features.

### D. OBJECTIVE FUNCTIONS

To perform adversarial training of generators $G_{i,\mathrm{GDCM}}$ and discriminators $D_{i,\mathrm{GDCM}}$, we define the objective functions of the generators $\mathcal{L}_G$ and discriminators $\mathcal{L}_D$, minimizing the losses alternately. The details of these losses are explained below.

#### 1) WORD-CONTEXT CORRELATION LOSS

To determine whether the visual attributes specified in the text description exist in the image, we use a word-context correlation loss [22]. Here, the GoogleNet-based [36] image encoder outputs the feature maps $\boldsymbol{C} \in \mathbb{R}^{C_c\times(H_cW_c)}$ of the input or manipulated images, and the pretrained bi-directional LSTM [23] outputs the word features $\boldsymbol{D} \in \mathbb{R}^{D\times L}$ of the text description. The feature maps $\boldsymbol{C}$ of the input or manipulated images are processed in the same manner to calculate this loss; thus, we assume the feature maps of the input image. A perception layer generates a feature $\tilde{\boldsymbol{C}} \in \mathbb{R}^{D\times(H_cW_c)}$ from $\boldsymbol{C}$, matching the channel dimension to $\boldsymbol{D}$. We use the features $\tilde{\boldsymbol{C}}$ and $\boldsymbol{D}$ and calculate the word-context correlation matrix $\boldsymbol{Q} = \mathrm{softmax}(\boldsymbol{D}^\top\tilde{\boldsymbol{C}}) \in \mathbb{R}^{L\times H_cW_c}$, which describes the correlation between each region in the image and the word. Then, we obtain weighted word features $\tilde{\boldsymbol{D}} = \tilde{\boldsymbol{C}}\boldsymbol{Q}^\top \in \mathbb{R}^{D\times L}$ by $\boldsymbol{Q}$. Finally, we calculate the similarity between the $j$-th word and

the entire image and define the word-context correlation loss by aggregating all similarities as follows:

$$r_j = \sigma\left(\frac{\tilde{\boldsymbol{D}}_j^{\top}\boldsymbol{D}_j}{||\tilde{\boldsymbol{D}}_j||_2||\boldsymbol{D}_j||_2}\right), \tag{2}$$

$$\mathcal{L}_{\text{cor}}(I,T) = \sum_{j=1}^{L} r_j, \tag{3}$$

where $\sigma(\cdot)$ denotes the sigmoid function. $r_j$ denotes the similarity between the $j$-th word and the image, and $\tilde{\boldsymbol{D}}_j$ and $\boldsymbol{D}_j$ describe the $j$-th column of $\tilde{\boldsymbol{D}}$ and $\boldsymbol{D}$, respectively. Here, we minimize $\mathcal{L}_{\text{cor}}(I,T)$ and $-\mathcal{L}_{\text{cor}}(I',T)$ for each correlation between the input or manipulated image and the text description.

#### 2) PERCEPTUAL LOSS

Without constraints on object information (e.g., shape or element location) not indicated in the text description, the generated images can be highly random and lose semantic alignment with the input images. To moderate this randomness, we use the perceptual loss [22] based on the pretrained VGG16 network [35]. The network extracts semantic features from the manipulated image $I'$ and input image $I$. The perceptual loss is defined as follows:

$$\mathcal{L}_{\text{per}}(I,I') = \frac{1}{C_p H_p W_p}||\text{VGG}(I') - \text{VGG}(I)||_F^2, \tag{4}$$

where $\text{VGG}(\cdot)$ denotes the VGG16 network, which generates the feature maps of the image from the *relu2_2* layer. $C_p$, $H_p$, and $W_p$ denote the number of color channels, the height, and the width of the feature maps, respectively. $||\cdot||_F^2$ denotes the Frobenius norm.

#### 3) CLIP LOSS

In the proposed method, to perceive fine-grained words in the text description, we implement the CLIP loss. Specifically, the CLIP model $\text{CLIP}(\cdot)$ comprises an image encoder based on ResNet50 [37] and a text encoder based on Transformer [38]. This model outputs 512-dimensional image and text features in the CLIP space from the text description $T$ and the image $\tilde{I} = G_3(\boldsymbol{X}_3)$ generated from the last stage in the main unit. We calculate the similarity between these features in the final stage of the main unit and define the CLIP loss as follows:

$$\mathcal{L}_{\text{CLIP}}(\tilde{I},T) = -\frac{\text{CLIP}(\tilde{I})\cdot\text{CLIP}(T)}{||\text{CLIP}(\tilde{I})||_2||\text{CLIP}(T)||_2}. \tag{5}$$

By minimizing the CLIP loss, the proposed method can effectively improve the visual and semantic consistency between the given text descriptions and the high-quality manipulated images acquired from the final stage of the multi-stage architecture.

#### 4) REGULARIZATION TERM

To ensure the diversity of generated images, we apply a regularization term [14] while training the GDCM unit. This regularization term is defined as follows:

$$\mathcal{L}_{\text{reg}}(I,I') = -\frac{1}{C_r H_r W_r}||I' - I||_F^2, \tag{6}$$

where $C_r$, $H_r$, and $W_r$ denote the number of color channels, the height, and the width of the input image $I$, respectively. This regularization term produces a significant penalty if the manipulated image $I'$ resembles the input image $I$, which helps prevent the network from learning identity mapping.

#### 5) OBJECTIVE FUNCTIONS OF GENERATOR

Based on [14], [22], the objective function of the generator $\mathcal{L}_G$ comprises the adversarial loss $\mathcal{L}_{\text{Gadv}}$, the perceptual loss $\mathcal{L}_{\text{per}}$, the word-level correlation loss $\mathcal{L}_{\text{cor}}$, the CLIP loss $\mathcal{L}_{\text{CLIP}}$, and the regularization term $\mathcal{L}_{\text{reg}}$. To ensure the diversity of the manipulated images, we use only the regularization term when training the generator $G_{\text{GDCM}}$. We calculate the objective function of the generator $\mathcal{L}_G$ as follows:

$$\begin{aligned}\mathcal{L}_G = {} & \mathcal{L}_{\text{Gadv}} + \mathcal{L}_{\text{CLIP}}(I',T) + \{1 - \mathcal{L}_{\text{cor}}(I',T)\} \\ & + \mathcal{L}_{\text{per}}(I,I') + \mathcal{L}_{\text{reg}}(I,I').\end{aligned} \tag{7}$$

The adversarial loss $\mathcal{L}_{\text{Gadv}}$ for the generators $G_{i,\text{GDCM}}$ is defined as follows:

$$\mathcal{L}_{\text{Gadv}} = \underbrace{-\mathbb{E}_{I'\sim p_G}[\log(D(I'))]}_{\text{unconditional adversarial loss}} \underbrace{-\mathbb{E}_{I'\sim p_G}[\log(D(I',T))]}_{\text{conditional adversarial loss}}. \tag{8}$$

Here, the unconditional adversarial loss makes the generated image $I'$ indistinguishable from the real image $I$, and the conditional adversarial loss aligns the generated image $I'$ with the given text description $T$.

#### 6) OBJECTIVE FUNCTIONS OF DISCRIMINATOR

Based on [22], the objective function of the discriminator $\mathcal{L}_D$ comprises the adversarial loss $\mathcal{L}_{\text{Dadv}}$ and the word-level correlation loss $\mathcal{L}_{\text{cor}}$. The objective function of the discriminator $\mathcal{L}_D$ is calculated as follows:

$$\mathcal{L}_D = \mathcal{L}_{\text{Dadv}} + \mathcal{L}_{\text{cor}}(I,T') + \{1 - \mathcal{L}_{\text{cor}}(I,T)\}, \tag{9}$$

where $T'$ represents an unpaired text description randomly selected from the training dataset. The adversarial loss $\mathcal{L}_{\text{Dadv}}$ for discriminators $D_{i,\text{GDCM}}$ is defined as follows:

$$\begin{aligned}\mathcal{L}_{\text{Dadv}} = {} & \underbrace{-\mathbb{E}_{I\sim p_{\text{data}}}[\log(D(I))] - \mathbb{E}_{I'\sim p_G}[\log(1 - D(I'))]}_{\text{unconditional adversarial loss}} \\ & \underbrace{-\mathbb{E}_{I\sim p_{\text{data}}}[\log(D(I,T))] - \mathbb{E}_{I'\sim p_G}[\log(1 - D(I',T))]}_{\text{conditional adversarial loss}}.\end{aligned} \tag{10}$$

The unconditional adversarial loss determines whether the given image is real, and the conditional adversarial loss

reflects the semantic similarity between the images and text descriptions.

This GAN is trained by alternately minimizing the objective functions of the generator $\mathcal{L}_G$ and discriminator $\mathcal{L}_D$ calculated above. As a result, the GAN can generate manipulated images that richly represent the given text description.

## IV. EXPERIMENTS

Here, we describe experimental results and verify the effectiveness of the segmentation guidance process within our proposed GAN.

### A. EXPERIMENTAL SETTINGS

To provide the segmentation guidance in the proposed method, we adopted the referring image segmentation model [17] pretrained on RefCOCO [39]. The model applies multimodal graph reasoning to associate images with texts and achieves high accuracy in generating the segmentation mask that extracts text-related regions. In the implementation, before training our GAN, we obtained the segmentation mask from all the image and text description sets in each dataset using the referring image segmentation model. During the training of our GAN, we did not change the parameters of the referring image segmentation model because we used the obtained mask without running the model each time. In this network, we transformed the segmentation mask into feature maps using the FTM.

We evaluated image manipulation performance on the Oxford-102 [40] dataset, the Caltech-UCSD-Birds (CUB) [41] dataset, and a new more complicated dataset. Each image in each dataset represents the details of a flower or bird, respectively. Table 1 shows the detailed statistics for each dataset. Our segmentation guidance process is expected to be powerful in situations involving multiple objects in an image; thus, it cannot be validated sufficiently on the CUB dataset, consisting of a single bird image. Therefore, to demonstrate the effectiveness of our image manipulation process, we constructed a CUB-based unique dataset. Here, we applied a semantic segmentation model [42] and automatically separated the regions of birds and backgrounds from all images in the CUB dataset. However, unlike the segmentation model [17] used in the proposed method, the model [42] takes only an image as input and generates the segmentation mask of objects in the image. We then manually selected 20 images to be used as backgrounds by cropping out areas of images in the CUB dataset that did not contain birds. Then, we selected one of the backgrounds created above and randomly aligned the two extracted birds in vertical or horizontal alignment. Finally, depending on the alignment relationship, we added the words *right*, *left*, *top*, or *bottom* to the text description originally attached to the image of either bird.

We compared the proposed method to state-of-the-art text-guided image manipulation methods, namely, the TAGAN [13], ManiGAN [14], Li'20 [15], and Haruyama'21 [16] methods. These methods are designed for image manipulation in situations with a single object in an image and have demonstrated promising results on the CUB dataset. To the best of our knowledge, there are no extensions to these methods that can manipulate only specific objects among multiple objects in an image based on the given text description. By comparing our method with these methods, we demonstrate that the proposed method enables robust image manipulation even for complex images.

**TABLE 1. Detailed statistics for each dataset.**

| | Oxford-102 [40] | CUB [41] | CUB-based unique dataset |
|---|---|---|---|
| sample | 8,189 | 11,788 | 11,788 |
| train : test | 6,149 : 2,040 | 8,855 : 2,933 | 8,855 : 2,933 |
| text / image | 10 | 10 | 10 |
| category | 102 | 200 | 200 |

During training, we followed the literature [14] and trained the two units independently to stabilize the output of the GAN. Specifically, via adversarial training, we first optimized the parameters in the main unit and then trained the GDCM unit with the fixed parameters from the main unit. For a fair comparison with the previous studies that have the same multi-stage architecture [14], [16] as the proposed method, we followed the experimental setup of those methods. Specifically, the main and GDCM units were trained over 600 and 100 epochs, respectively, using the Adam optimizer [43] with a learning rate of 0.0002. Note that we applied the proposed and comparison methods trained on the original CUB dataset and performed image manipulation for images in the CUB-based unique dataset.

To demonstrate the effectiveness of segmentation guidance, we applied the inception score (IS) [44], Fréchet inception distance (FID) [45], and kernel inception distance (KID) [46] metrics and verified that the quality of the manipulated images obtained by the proposed method is comparable to that of the comparison methods. Based on previous studies, we calculated the values of these metrics for the test images on each dataset. By calculating IS, FID, and KID on each dataset as well as in previous studies on text-guided image manipulation, we can demonstrate that the quality of images manipulated using the proposed method is not degraded by the introduction of segmentation guidance.

We conducted a subjective experiment on the CUB-based unique dataset to evaluate image manipulation precision in complex situations. In this experiment, we randomly selected 30 sets of images and text descriptions from the CUB-based unique dataset. We then randomly changed the words related to the attributes of the selected text description to different words (i.e., *yellow* and *white*) and created a new text description representing the user's desired image manipulation. Here, to create a sample evaluated by subjects, we used the proposed and comparison methods to obtain a manipulated image with the selected image and new text description as inputs. In this experiment, the manipulated images generated using the proposed and comparison methods were presented

to participants, who were asked to evaluate each image in terms of accuracy and realism based on [15]. These metrics are described as follows. (1) Accuracy: indicates whether the text-related region is manipulated according to the given text description, and whether the text-unrelated regions are preserved. (2) Realism: indicates whether the manipulated image appears realistic. We obtained informed consent from 24 participants and then asked them to assign scores of 1–5 (1: inaccurate or unreal to 5: accurate or real) to each image according to these two metrics.

### B. QUANTITATIVE RESULTS

The IS, FID, and KID results for the generated images are listed in Table 2. The IS, FID, and KID values of the proposed method are the first or second-best results, which demonstrates the images generated using the proposed method have similar quality compared with those generated using the comparison state-of-the-art methods. From these results, we confirm that our segmentation guidance process does not negatively affect the quality of the generated images.

The accuracy and realism results obtained in the subjective experiment are listed in Table 3. Note that accuracy and realism values shown are the mean values for each text-guided image manipulation method, which were calculated on the basis of the scores given by the participants. The accuracy results demonstrate the effectiveness of the proposed method in complex situations. The accuracy of the proposed method is the highest among all comparison methods. The accuracy metric is used to evaluate the precision of the manipulated image generated from an image with two birds and a text description representing the manipulation of only a single bird. These results prove the advantages of our image manipulation method for complex situations. We also conducted a Welch's t-test on the proposed method and ManiGAN [14], and we verified statistically significant differences of 1% ($p$-values $<0.01$) in terms of accuracy. From the realism results, the quality of manipulated images generated using the proposed method is favorable in terms of the two automatic evaluation metrics (i.e., IS, FID, and KID) and the subjective user evaluations.

### C. QUALITATIVE RESULTS

To evaluate the visual quality of the manipulated images, we compare the results obtained using the proposed method and comparison methods in Fig. 3. Here, samples (A1), (A2), (B1), and (B2) on the datasets of a single object verify that the proposed method achieves improved reflection of the text description in the text-related region with nearly the same image quality as the comparison methods. Note that TAGAN [13] does not adopt a multi-stage architecture; thus, the image quality of the manipulated images by TAGAN is inferior to that of the other methods. Specifically, in sample (A1), only the proposed method succeeds in generating the *blue* attribute while preserving the white region in the background. For sample (A2), the manipulated images generated using the comparison methods lose the texture of the stripes or still retain the *purple* attribute. Moreover, the manipulated image generated using the proposed method successfully retains the texture of the stripes and modifies its colors. For sample (B1), only the proposed method generated a manipulated image with the *red* attribute while preserving the background and the details of the bird, e.g., the texture. For sample (B2), we compared the precision of image manipulation when the text description contained two attribute words (i.e., *blue* and *white*). The proposed method successfully reflects all elements of the text description, e.g., *white breast* and *blue tail*, in the bird shown in the manipulated image. This is achieved because the proposed method lets the segmentation guidance process handle the selection of the text-related region, which allows the GAN to focus on generating new attributes that correspond to the text description based on the CLIP loss.

For samples (C1) and (C2) with complex situations from the CUB-based unique dataset, the proposed method can focus on only text-related objects and maintains sufficient image manipulation precision. The proposed and comparison methods were trained on the CUB dataset, which does not include phrases such as *a left bird* in the text description. Thus, the comparison methods are primarily influenced by the word *bird* and manipulate the two birds in the image. In contrast, the proposed method suppresses the manipulation of text-unrelated regions by introducing segmentation guidance based on the referring image segmentation model, which can select the regions in the image identified by the text description. Specifically, images manipulated by the proposed maintain the original colors of text-unrelated birds, e.g., *a right bird* in sample (C1) and *a bottom bird* in sample (C2). In addition, some of the comparison methods achieve partial success in suppressing background manipulation; however, the proposed method appears to most appropriately suppress background manipulation in the manipulated images. These experimental results demonstrate the advantages of the proposed method over the comparison state-of-the-art methods [13], [14], [15], [16], especially in terms of image manipulation precision.

### D. DISCUSSION

#### 1) EFFECTIVENESS OF SEGMENTATION GUIDANCE

Fig. 4 depicts the segmentation masks that are transformed into feature maps $\boldsymbol{S}_{\mathrm{GACM}}$ or $\boldsymbol{S}_{\mathrm{GDCM}}$ and used as the segmentation guidance in the proposed method's GAN. The segmentation mask successfully extracts only text-related regions and can indicate to the network which regions should be manipulated. With this segmentation guidance, the generators in the proposed method can focus on generating images based on the text description and reconstructing input images without selecting regions. Specifically, as depicted in Fig. 4, the temporarily generated images produced by $G_3$ reflect the text descriptions in the regions extracted by the segmentation masks; thus, the GACM works effectively. The image produced by the last generator $G_3$ in the main unit has a rich representation of the visual attributes indicated by the text description, whereas the text-unrelated regions are not

**TABLE 2.** Quantitative results of proposed and comparison methods. IS, FID, and KID are automatic metrics to evaluate the quality of generated images. Note that each model trained on the CUB dataset was used to perform image manipulation on test images in the CUB-based unique dataset. (↑ and↓ imply that higher or lower is better. Bold and underlined values are the best and second-best results for each metric, respectively.)

| | Oxford-102 [40] | | | CUB [41] | | | CUB-based unique dataset | | |
|---|---|---|---|---|---|---|---|---|---|
| | IS(↑) | FID(↓) | KID(↓) | IS(↑) | FID(↓) | KID(↓) | IS(↑) | FID(↓) | KID(↓) |
| TAGAN [13] | 2.86 | 66.48 | 0.384 | 3.64 | 57.20 | 0.205 | 3.29 | 58.75 | 0.334 |
| ManiGAN [14] | 2.90 | 52.53 | 0.223 | 4.58 | 11.30 | 0.036 | 3.60 | <u>19.45</u> | <u>0.069</u> |
| Li'20 [15] | 3.81 | 36.78 | 0.171 | <u>4.64</u> | **9.10** | <u>0.018</u> | 3.31 | 34.96 | 0.181 |
| Haruyama'21 [16] | <u>3.83</u> | <u>27.52</u> | <u>0.125</u> | 4.54 | 9.47 | 0.020 | <u>3.66</u> | 20.22 | 0.074 |
| Ours | **3.95** | **19.67** | **0.052** | **6.91** | <u>9.24</u> | **0.016** | **3.85** | **16.97** | **0.067** |

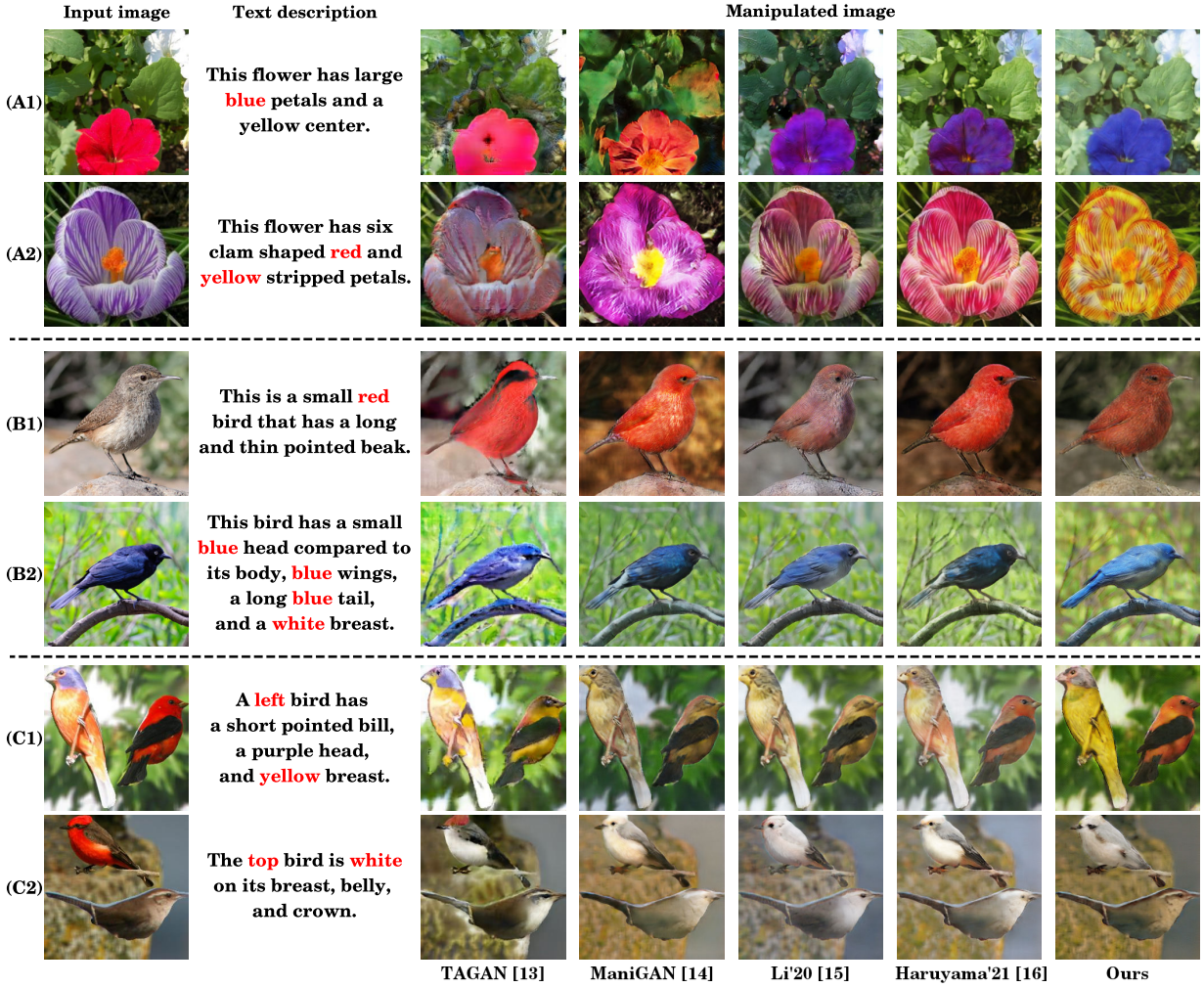


**FIGURE 3.** Qualitative results of proposed method and four comparison methods [13], [14], [15], [16]. Samples (B1) and (B2) are the comparison on the Oxford-102 dataset. Samples (C1) and (C2) are the comparison on the CUB dataset. Samples (C1) and (C2) are the comparison on the CUB-based unique dataset representing complex cases used in the subjective experiment.

properly reconstructed. By utilizing segmentation guidance based on feature maps $S_{\mathrm{GDCM}}$, the GDCM reconstructs the contents of the input image in text-unrelated regions. The difference between the image generated by $G_3$ and the final manipulated image is due to the contribution of the GDCM.

### 2) REPRESENTATION CAPABILITIES OF GENERATORS

The generated images corresponding to $G_i$ are shown in the three columns on the right side of Fig. 4. Here, we show $64 \times 64$, $128 \times 128$, and $256 \times 256$ images generated using our multi-stage architecture. As shown in the figure, the visual representation becomes increasingly sophisticated as the resolution increases. By benefiting from segmentation guidance based on feature maps $\boldsymbol{S}_{\mathrm{GACM}}$, these images have the visual attributes indicated by the text description in each text-related region. The generators achieve high representational capability by introducing the CLIP loss, and they can reflect fine-grained words in the generated images. In addition, by benefiting from segmentation guidance based on feature maps $\boldsymbol{S}_{\mathrm{GDCM}}$, the generator $G_{\mathrm{GDCM}}$ in the GDCM

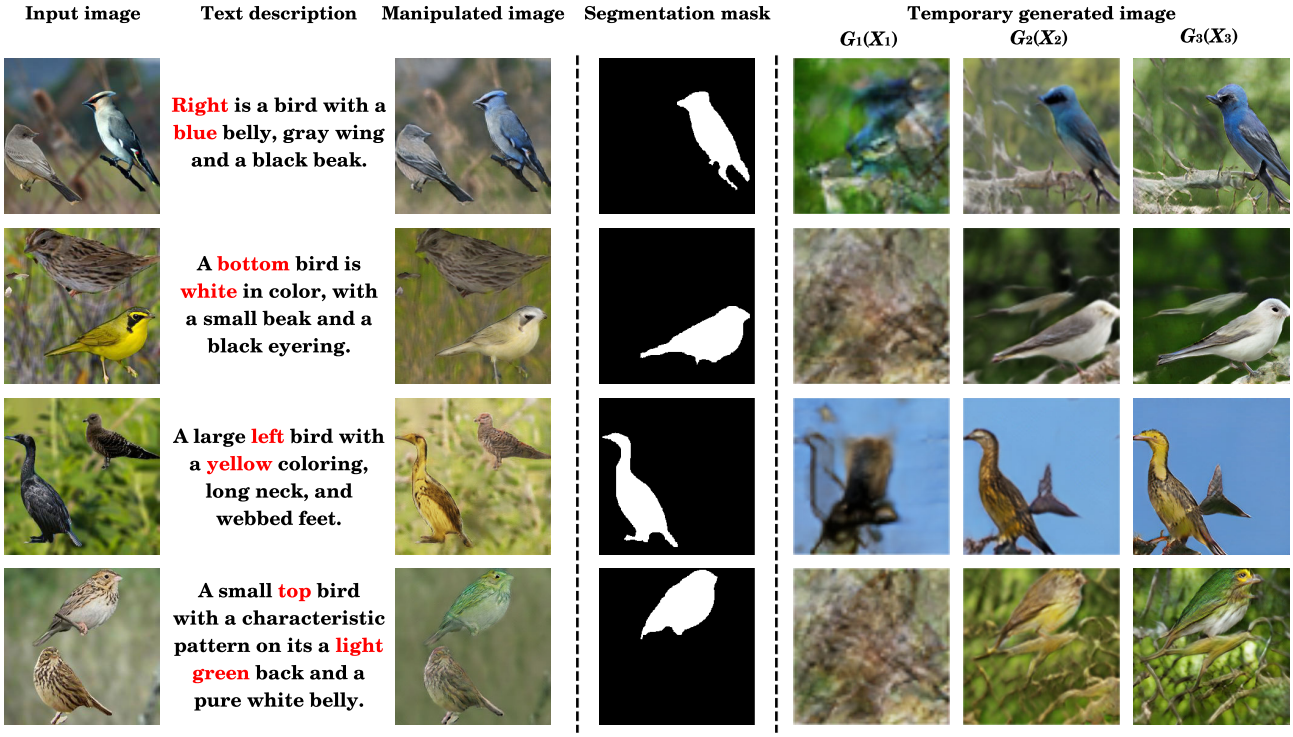


**FIGURE 4.** Detailed visual analyses of the proposed method. All samples are the image manipulation results obtained on the CUB-based unique dataset representing complex situations. In addition to the image manipulation results, the segmentation mask used as segmentation guidance and the 64 × 64, 128 × 128, and 256 × 256 images generated by inputting $X_i$ into $G_i$ in the main unit are listed from left to right.

**TABLE 3.** The results of accuracy and realism obtained in the subjective experiment using the CUB-based unique dataset that describes more complex situations than the CUB dataset. These values are the mean values of 30 manipulated images for each method. Note that each model trained on the CUB dataset was used to perform image manipulation on test images in the CUB-based unique dataset. (↑ and ↓ imply that higher or lower is better. Bold and underlined values are the best and second-best results for each metric, respectively.)

| | Accuracy(↑) | Realism(↑) |
|---|---|---|
| TAGAN [13] | 2.74 | 3.20 |
| ManiGAN [14] | <u>2.84</u> | 3.21 |
| Li'20 [15] | 2.83 | **3.68** |
| Haruyama'21 [16] | 2.83 | 3.25 |
| Ours | **4.12** | <u>3.47</u> |


**FIGURE 5.** Examples of failed image manipulation on the CUB-based unique dataset. The proposed method performed image manipulation using the text description: "A right bird with a red color has a long beak."

unit can properly reconstruct the content of the input image without canceling the attributes generated by $G_3$.

### 3) EXAMPLES OF FAILED IMAGE MANIPULATION

Although the proposed method significantly improves image manipulation precision by introducing segmentation guidance, there are some cases where images could not be manipulated inadequately. Our segmentation guidance plays a major role in improving image manipulation precision, and the output results of the pretrained referring image segmentation model have a significant impact on the proposed method's performance. In Fig. 5, we show an example in which the proposed method fails to realize effective image manipulation on the CUB-based unique dataset. Here, the segmentation mask fails to extract *a right bird*, and the region in the manipulated image does not reflect the text description. In addition, the incorrect region extracted by the segmentation mask does not completely cover the object, and the region of *a left bird* is interrupted unnaturally. Thus, in the future, we plan to investigate the matching between images and texts to enable the extraction of small objects and improve the performance of the segmentation guidance process.

Although a segmentation mask can be directly applied to an image to distinguish the regions to be reconstructed and manipulated, we attempt to reduce disconnectedness by applying a segmentation mask to image features. However, few artifacts certainly remain in the manipulated image generated using the proposed method. In future studies, we will consider introducing a module to mitigate artifacts at the end of the network as well as using a segmentation mask of the state of the score map prior to the binarization.

## V. CONCLUSION

This study has proposed a text-guided image manipulation method that can handle complex image manipulation situations. During the image manipulation process, by introducing a segmentation guidance process into a GAN, the proposed method can focus on the generation of new attributes and reconstruction of text-unrelated regions without having to consider the selection of regions. Our experimental results demonstrate that the novel GACM and GDCM provide effective segmentation guidance to our GAN architecture, and we enable robust image manipulation even for complex images. Automatic and subject evaluations across extensive experiments on multiple datasets substantiate this fact. However, there is still room for improvement in the selection of regions in the proposed method, and we will consider the optimal use of a segmentation mask to mitigate artifacts in GANs.

## REFERENCES

[1] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[2] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 649–666.

[3] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 597–613.

[4] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[5] I. 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. (NeurIPS)*, vol. 27, 2014, pp. 2672–2680.

[6] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[7] V. Lempitsky, A. Vedaldi, and D. Ulyanov, "Deep image prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 9446–9454.

[8] Z. Cheng, Q. Yang, and B. Sheng, "Deep colorization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 415–423.

[9] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," 2015, *arXiv:1508.06576*.

[10] A. Madaan, A. Setlur, T. Parekh, B. Poczos, G. Neubig, Y. Yang, R. Salakhutdinov, A. W. Black, and S. Prabhumoye, "Politeness transfer: A tag and generate approach," 2020, *arXiv:2004.14257*.

[11] 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. (CVPR)*, Jul. 2017, pp. 1125–1134.

[12] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, "CrDoCo: Pixel-level domain transfer with cross-domain consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1791–1800.

[13] S. Nam, Y. Kim, and S. J. Kim, "Text-adaptive generative adversarial networks: Manipulating images with natural language," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018, pp. 42–51.

[14] B. Li, X. Qi, T. Lukasiewicz, and P. H. S. Torr, "ManiGAN: Text-guided image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7880–7889.

[15] B. Li, X. Qi, P. H. Torr, and T. Lukasiewicz, "Lightweight generative adversarial networks for text-guided image manipulation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 22020–22031.

[16] T. Haruyama, R. Togo, K. Maeda, T. Ogawa, and M. Haseyama, "Segmentation-aware text-guided image manipulation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 2433–2437.

[17] S. Huang, T. Hui, S. Liu, G. Li, Y. Wei, J. Han, L. Liu, and B. Li, "Referring image segmentation via cross-modal progressive comprehension," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10488–10497.

[18] 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," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[19] Y. Watanabe, R. Togo, K. Maeda, T. Ogawa, and M. Haseyama, "Generative adversarial network including referring image segmentation for text-guided image manipulation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 4818–4822.

[20] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1060–1069.

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

[22] B. Li, X. Qi, T. Lukasiewicz, and P. Torr, "Controllable text-to-image generation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019, pp. 2065–2075.

[23] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1316–1324.

[24] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1219–1228.

[25] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[26] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6306–6314.

[27] Z. Pang, J. Guo, Z. Ma, W. Sun, and Y. Xiao, "Median stable clustering and global distance classification for cross-domain person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3164–3177, May 2022.

[28] H. Kim, Z. Pang, L. Zhao, X. Su, and J. S. Lee, "Semantic-aware deidentification generative adversarial networks for identity anonymization," *Multimedia Tools Appl.*, vol. 82, no. 10, pp. 15535–15551, 2023.

[29] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 108–124.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, and A. Yuille, "Recurrent multimodal interaction for referring image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1271–1280.

[32] R. Li, K. Li, Y.-C. Kuo, M. Shu, X. Qi, X. Shen, and J. Jia, "Referring image segmentation via recurrent refinement networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5745–5753.

[33] H. Dong, S. Yu, C. Wu, and Y. Guo, "Semantic image synthesis via adversarial learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5706–5714.

[34] S. Xiao, J. Yang, and Z. Lv, "Protecting the trust and credibility of data by tracking forgery trace based on GANs," *Digit. Commun. Netw.*, vol. 8, no. 6, pp. 877–884, Dec. 2022.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.

[39] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 69–85.

[40] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process. (ICVGIP)*, Dec. 2008, pp. 722–729.
[41] 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.
[42] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.
[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[44] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 29, 2016, pp. 2234–2242.
[45] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 6626–6637.
[46] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–36.

**YUTO WATANABE** (Graduate Student Member, IEEE) received the B.S. degree in electronics and information engineering from Hokkaido University, Japan, in 2022, where he is currently pursuing the M.S. degree with the Graduate School of Information Science and Technology. His research interests include computer vision, image manipulation, and image quality assessment.

**REN TOGO** (Member, IEEE) 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 currently a specially appointed Assistant Professor with the Faculty of Information Science and Technology, Hokkaido University. He is also a Radiological Technologist. His research interest includes machine learning and its applications. He is a member of ACM and IEICE.

**KEISUKE MAEDA** (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronics and information engineering from Hokkaido University, Japan, in 2015, 2017, and 2019, respectively. He is currently a specially appointed Assistant Professor with the Faculty of Information Science and Technology, Hokkaido University. His research interests include multimodal signal processing and machine learning and its applications. He is a member of IEICE.

**TAKAHIRO OGAWA** (Senior Member, IEEE) 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. He joined the Graduate School of Information Science and Technology, Hokkaido University, in 2008. He is currently a Professor with the Faculty of Information Science and Technology, Hokkaido University. His research interests include artificial intelligence, the Internet of Things, and big data analysis for multimedia signal processing and its applications. He is a member of ACM, IEICE, and ITE. He was the Special Session Chair of IEEE ISCE 2009, the Doctoral Symposium Chair of ACM ICMR 2018, an Organized Session Chair of IEEE GCCE 2017–2019, the TPC Vice Chair of IEEE GCCE 2018, and the Conference Chair of IEEE GCCE 2019. He was an Associate Editor of *ITE Transactions on Media Technology and Applications*.

**MIKI HASEYAMA** (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronics from Hokkaido University, Japan, in 1986, 1988, and 1993, respectively. She joined the Graduate School of Information Science and Technology, Hokkaido University, 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, Hokkaido University. Her research interests include image and video processing and its development into semantic analysis. She is a fellow of ITE and a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and ASJ. She was the Vice-President of the Institute of Image Information and Television Engineers (ITE), Japan, the Editor-in-Chief of *ITE Transactions on Media Technology and Applications*, and the Director of the International Coordination and Publicity at IEICE.