GENERATIVE ADVERSARIAL NETWORK INCLUDING REFERRING IMAGE SEGMENTATION FOR TEXT-GUIDED IMAGE MANIPULATION

Yuto Watanabe, Ren Togo, Keisuke Maeda, Takahiro Ogawa, Miki Haseyama Hokkaido University, Japan

【Paper ID : 3843】

ACKNOWLEDGEMENT : This work was partly supported by JSPS KAKENHI Grant Number JP20K19857 and JP20K19856.

INTRODUCTION

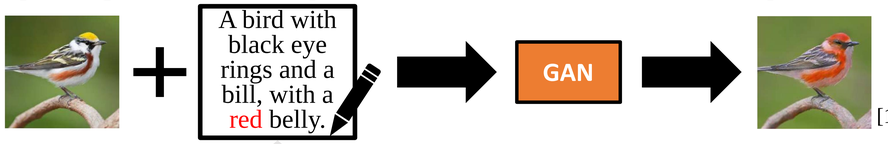
Image manipulation is expected to have many fields such as image inpainting [1], image colorization [4], style transfer [7] and domain transformation [11].

There are several methods focusing on more user-friendly image manipulation.

Text-guided image manipulation methods [14-17]

Generative adversarial networks (GANs) that manipulate the image by using natural language descriptions.

Input image + Text description (A bird with black eye rings and a bill, with a red belly.) → GAN → Manipulated image [15]



The text description contains the user's demands.

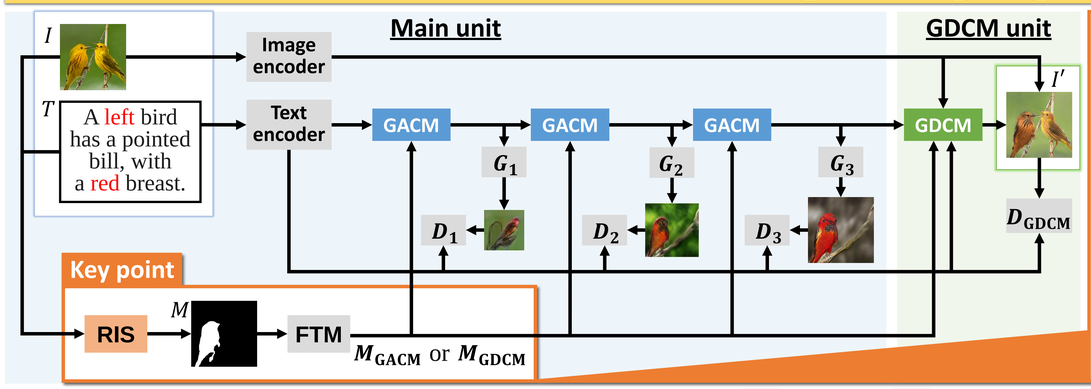
There are the following two problems.

Previous methods [14-17]

- **manipulate incorrect object that are not specified in the text description.**
- **still have attributes of an input image in the manipulated image.**

PROPOSED METHOD

Generative Adversarial Network Introducing Referring Image Segmentation (RIS)



The key point is the introduction of RIS.

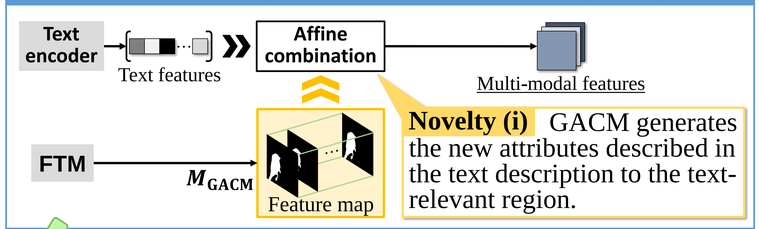
RIS aims to generate the segmentation mask to extract the region relevant to the text description.

Novelty (i) GACM concatenates the segmentation mask M with the text description T.

Novelty (ii) GDCM distinguishes image features of the input image I into the text-relevant and irrelevant regions.

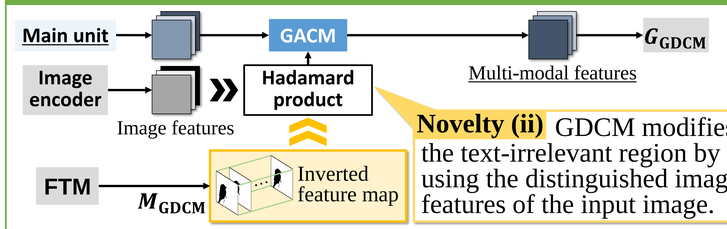
I: Input image, T: Text description, M: Segmentation mask, I': Manipulated image, G : Generator, D : Discriminator, **FTM** : Feature transformation module

Guided Text-Image Affine Combination Module (GACM)



Novelty (i) GACM generates the new attributes described in the text description to the text-relevant region.

Guided Detail Correction Module (GDCM)



Novelty (ii) GDCM modifies the text-irrelevant region by using the distinguished image features of the input image.

The proposed method can manipulate only the text-relevant region and preserve other regions by GACM and GDCM.

EXPERIMENTAL RESULTS

Conditions

Training data is 8,855 images and text descriptions of Caltech-UCSD Birds (CUB) [20], such as (A).

Evaluation Methods

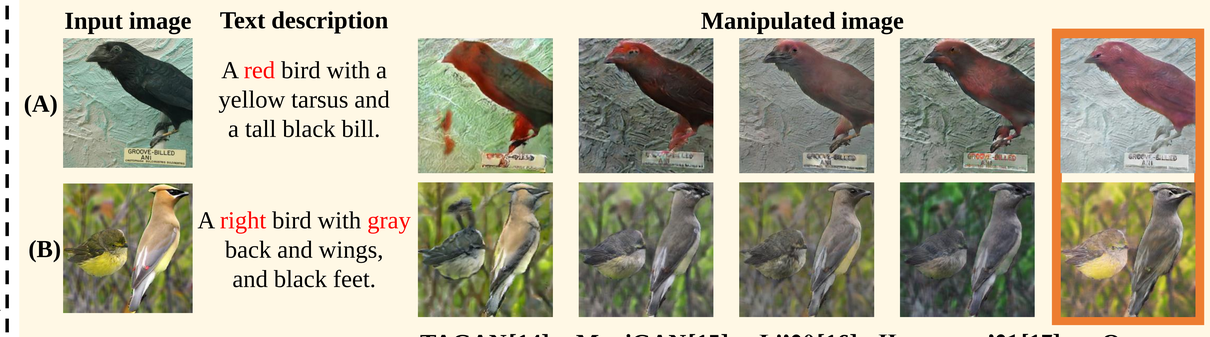
1) Evaluation for the quality of generated images
- Inception score (IS) [27] and Fréchet inception distance (FID) [28] for manipulated images (CUB).

2) Evaluation for the accuracy of image manipulation
- Subjective experiments for manipulated images based on Realism and Accuracy, according to [16] (A newly created dataset, such as (B)).
- A total of 18 subjects participated.

Comparative Methods

Existing text-guided image manipulation methods: TAGAN[14], ManiGAN[15], Li'20[16], Haruyama'21[17]

Qualitative Results



- In (A), the manipulated image by our method has accurate attributes described in the text description.
- In (B), our method successfully suppresses the manipulation of the text-irrelevant object.

Quantitative Results

	IS(↑)	FID(↓)	Realism(↑)	Accuracy(↑)
TAGAN[14]	3.64	57.20	2.82	2.43
ManiGAN[15]	4.58	11.30	3.01	2.64
Li'20[16]	4.64	**9.10**	3.67	2.68
Haruyama'21[17]	4.54	9.47	3.01	2.64
Ours	**5.86**	9.33	**3.83**	**4.16**

IS, FID and Realism The naturalness of the manipulated images by our method is almost equal to or better than those by the state-of-the-art methods [14-17].

Accuracy Manipulated images by our method align with the text description and preserve the text-irrelevant regions, successfully.

Our method qualitatively and quantitatively outperforms the performance of the four state-of-the-art methods.

[1] D. Pathak *et al.*, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544. [4] Z. Cheng *et al.*, in *Proc. IEEE/CVF International Conference on Computer Vision*, 2015, pp. 415–423. [7] L. A. Gatys *et al.*, *arXiv preprint arXiv:1508.06576*, 2015. [11] P. Isola *et al.*, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.
[14] S. Nam *et al.*, *arXiv preprint arXiv:1810.11919*, 2018. [15] B. Li *et al.*, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7880–7889. [16] B. Li *et al.*, *Advances in Neural Information Processing Systems*, pp. 22020–22031, 2020. [17] T. Haruyama *et al.*, in *Proc. IEEE International Conference on Image Processing*, 2021, pp. 2433–2437. [20] C. Wah *et al.*, *California Institute of Technology*, 2011.
[27] T. Salimans *et al.*, *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016. [28] M. Heusel *et al.*, *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.