ISO 9001:2000
Reg. No. : RQ91/3688

MUSLIM ARTS COLLEGE

THIRUVITHANCODE-629174, KANYAKUMARI DISTRICT

TAMILNADU.

National Conference on

Inter disciplinary Research through New Age Information Technology (IRNAIT-2023)

2023, February 24, Friday

Certificate

This is to certify that ~~Prof.~~ / ~~Dr.~~ / ~~Dr.~~ / ~~Mrs.~~

FELSIA THOMPSON, Assist. Professor

Muslim Arts College, Thiruvithancode.

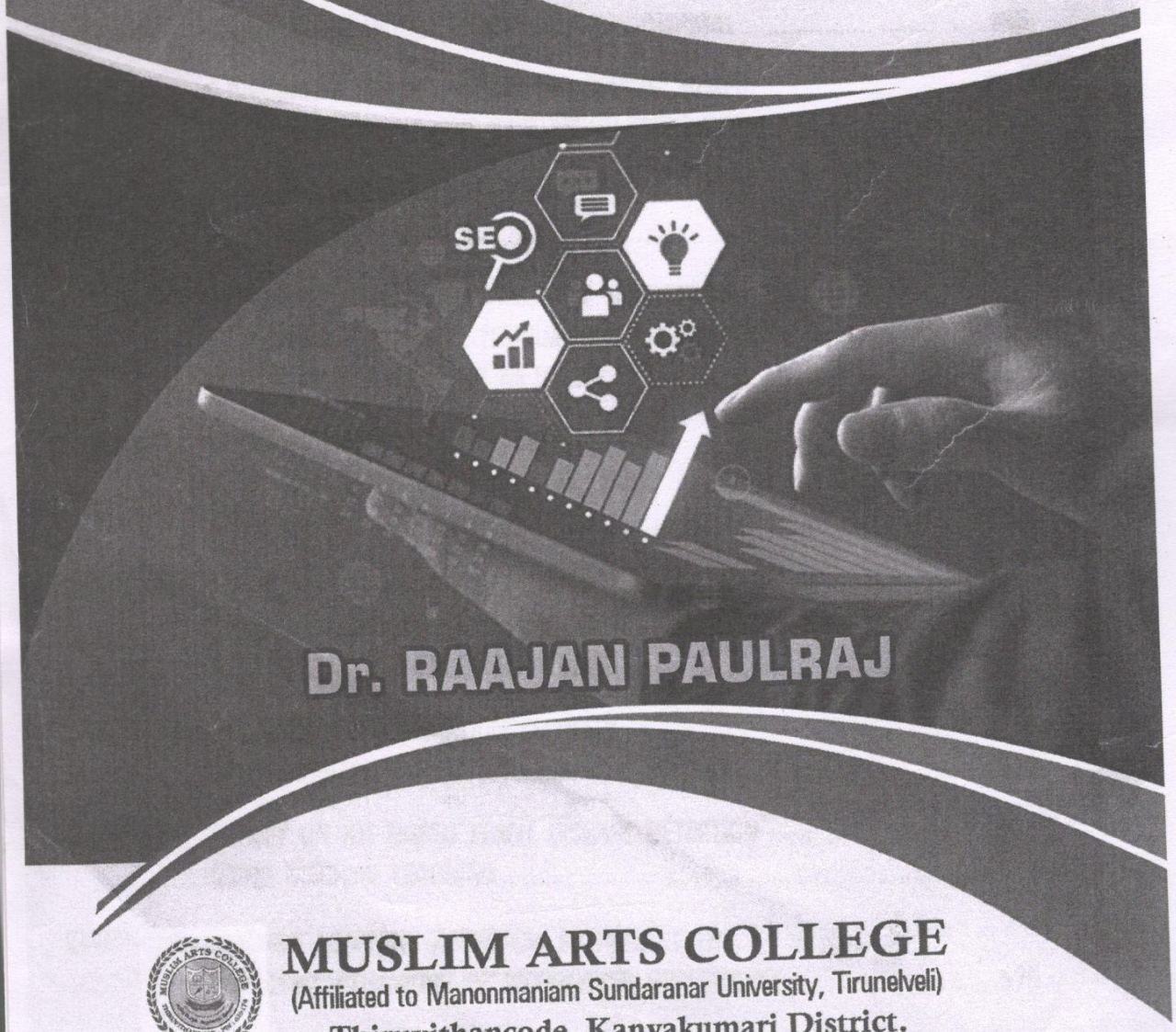
has participated / Best paper / presented a paper entitled

EXPLORING THE ADVANCEMENTS IN TEXT TO IMAGE MANIPULATION.

in the National Conference on "Inter disciplinary Research through New Age Information Technology" held on 24th February 2023, organized by the P.G and Research Department of Computer Science, Muslim Arts College, Thiruvithancode, Kanyakumari-629174, Tamil Nadu, India.

Lion Dr H. Mohamed Ali	Dr G. Edwin Sheela	Dr. P. Raajan
Chief Patron	Patron	Organizing Secretary

Research Trends in information Technology



Dr. RAAJAN PAULRAJ

MUSLIM ARTS COLLEGE
(Affiliated to Manonmaniam Sundaranar University, Tirunelveli)
Thiruvithancode, Kanyakumari District.

Published by

Tamilsuvadi

182, First Middle Street, Thiyagaraja Nagar,
Tirunelveli-627 011.
Cell : 95979 22250.
www.booksha.in

Disclaimer:

The findings/views/opinions expressed in the book are solely those of the authors and do not necessarily reflect the views of the publisher.

Book Name	:	RESEARCH TRENDS IN INFORMATION TECHNOLOGY
Author Name	:	Dr.P Raajan
Toatal Pages	:	700
Rate	:	Rs. 1550/-
First Edition	:	2023
ISBN No	:	ISBN 978-81-962277-1-5

தமிழ்ச்சுவடி

Tamilsuvadi

182, First Middle Street, Thiyagaraja Nagar,
Tirunelveli-627 011.
Cell : 95979 22250. www.booksha.in

IRNAIT-053.
EXPLORING THE ADVANCEMENTS IN TEXT TO IMAGE MANIPULATION

M. Sapna,
Reg No: 20213092506221,
II M. Sc Computer Science,
Department of Computer Science,
Muslim Arts College, Thiruvithancode, Azhagiamandapam - 629174
(Affiliated to Manonmaniam Sundaranar University, Tirunelveli - 627 012)
E-Mail: sapnanowshad5@gmail.com

Dr. Felsia Thompson,
Assistant professor, Department of Computer Science,
Muslim Arts College, Thiruvithancode, Azhagiamandapam - 629174
(Affiliated to Manonmaniam Sundaranar University, Tirunelveli - 627 012)
E-Mail: felsiashanu@gmail.com

ABSTRACT:

Text to image manipulation refers to the process of generating images from textual descriptions. In recent years, this field has gained significant attention due to its numerous applications, such as in the gaming and entertainment industry, digital media, and even in scientific research. The purpose of this research report is to explore the current state of the field, including the latest advancements and emerging trends, as well as the challenges faced by researchers and practitioners.This research report focuses on the advancements and challenges in the field of text to image manipulation, where images are generated from textual descriptions. The report provides an overview of the techniques used in this field, including generative adversarial networks (GANs), variational autoencoders (VAEs), and attention-based models. It also highlights the latest advancements and emerging trends, such as the integration of virtual and augmented reality and the use of text to image manipulation in scientific research. The report concludes by addressing the challenges faced by researchers and practitioners, including the lack of robust evaluation metrics and

comprehensive datasets, and the commitment to exploring and solving these challenges to advance the field

KEYWORD: Text to image manipulation, Generative adversarial networks, Variational autoencoders, Attention-based models, Virtual reality, Scientific research.

I.INTRODUCTION

Image manipulation refers to the modification of specific aspects of images, from low-level elements like color and texture to high-level semantics. This technique has numerous applications in video games, image editing, and computer-aided design. With the advancement of deep learning and deep generative models, automatic image manipulation has seen significant progress in various fields including image inpainting, colorization, style transfer, and domain or attribute translation. However, most existing methods concentrate on specific problems, and few studies focus on general and user-friendly image manipulation using natural language descriptions.This report focuses on the task of semantically editing parts of an image according to a user's text description while preserving unmentioned contents. Although the current state-of-the-art methods are able to generate images guided by text descriptions, the resulting images are often low-quality and fail to effectively manipulate complex scenes. To address these limitations, this report proposes a novel generative adversarial network, ManiGAN, for text-guided image manipulation.The key to ManiGAN is the text-image affine combination module (ACM) that collaborates between text and image features to select text-relevant regions for modification, correlate them with semantic words, and generate new attributes aligned with the given text description. The model also encodes original image representations for reconstructing text-irrelevant contents. Additionally, a detail correction module (DCM) is introduced to rectify mismatched attributes and complete missing information. These modifications allow ManiGAN to produce high-quality, fine-grained manipulation results. Image manipulation is a rapidly growing field with numerous potential applications in a variety of industries. The traditional approach to image manipulation requires manual effort and is time-consuming, making it difficult for non-experts to use. With the development of deep learning and deep generative models, automatic image manipulation has made remarkable progress, enabling the creation of high-quality, visually appealing images.However, the current state-of-the-art methods for text-guided image manipulation are limited in their ability to produce high-quality images, particularly when dealing with complex scenes. These limitations stem from the inability of existing models to precisely correlate fine-grained words with corresponding visual attributes, as well as their inability to effectively identify and preserve text-irrelevant contents in the original image.To overcome these limitations, this report proposes ManiGAN, a novel generative adversarial network for text-guided image manipulation. ManiGAN employs a text-image affine combination module (ACM) to collaborate between text and image features to select text-relevant regions for modification, correlate them with semantic

ords, and generate new attributes aligned with the given text description. Additionally, ManiGAN includes a detail correction module (DCM) to rectify mismatched attributes and complete missing information, further enhancing the quality of the final results.Our final model can produce high-quality manipulation results with fine-grained details (see Fig. 1: Ours).

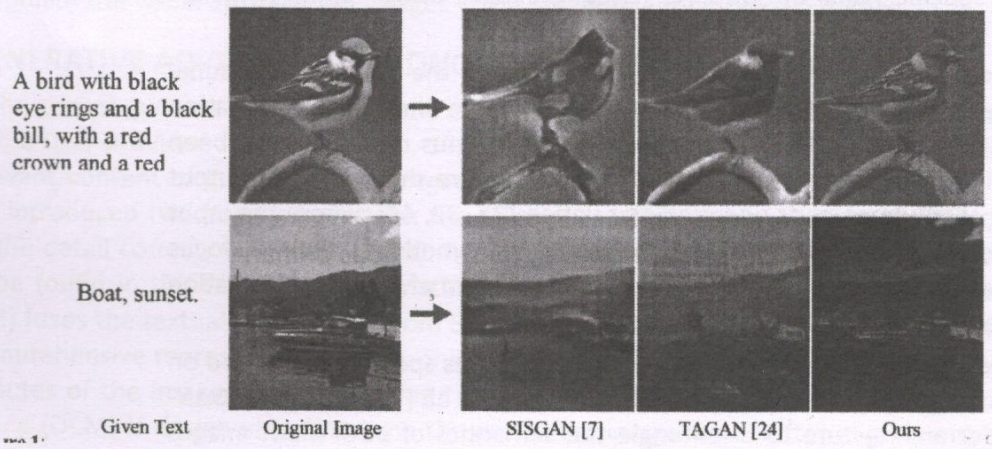


Figure 1: Given an original image that needs to be edited and a text provided by a user describing desired attributes, the goal is to edit parts of the image according to the given text while preserving text-irrelevant contents. Current state-of-the-art methods only generate low-quality images, and fail to do manipulation on COCO. In contrast, our method allows the original image to be manipulated accurately to match the given description, and also reconstructs text-irrelevant contents.

In conclusion, ManiGAN represents a significant advancement in the field of text-guided image manipulation, enabling the creation of high-quality, fine-grained images that accurately reflect the user's specifications, and has the potential to revolutionize the field with its numerous practical applications, from video games and image editing to computer-aided design. Additionally, a new metric is suggested to assess image manipulation results, which appropriately reflects the performance in terms of generating new visual attributes and reconstructing text-irrelevant contents of the original image. The superiority of ManiGAN is demonstrated through extensive experiments on the CUB and COCO datasets, outperforming existing state-of-the-art methods both qualitatively and quantitatively.

RELATED WORK

Text-to-image Generation: Text-to-image generation has been a widely studied topic due to the breakthroughs in Generative Adversarial Networks (GANs) in generating highly realistic images [11]. One approach to text-to-image generation is using conditional GANs, where given text descriptions are used as conditions to generate plausible images [28].

Another approach involves stacking multiple GANs to generate high-resolution images from coarse to fine-scale, where each GAN is responsible for generating a different level of details [40, 41]. Additionally, attention mechanisms have been used to better capture fine-grained information at the word-level [39, 18]. However, despite the impressive results, these methods primarily focus on generating new photo-realistic images from texts and not on manipulating specific visual attributes of existing images using natural language descriptions.

Conditional Image Synthesis: Our work is closely related to the field of conditional image synthesis, where additional information is used to guide the image generation process [1, 2, 4, 9, 20, 23, 25, 33, 34, 43]. In recent years, various methods have been proposed for paired image-to-image translation [3, 14, 37], where the input and output images belong to the same domain, or unpaired translation [21, 32, 44], where the input and output images belong to different domains. However, these methods primarily focus on same-domain image translation and do not address the problem of image manipulation using cross-domain text descriptions.

Text-guided Image Manipulation: There have been few studies specifically focused on image manipulation using natural language descriptions. Dong et al. [7] proposed a GAN-based encoder-decoder architecture to disentangle the semantics of both input images and text descriptions, where the generator learns to generate images that match the given text descriptions. Nam et al. [24] implemented a similar architecture but with a text-adaptive discriminator that provides specific word-level training feedback to the generator. However, both methods are limited in their performance due to an ineffective text-image concatenation method and a coarse sentence condition.

Affine Transformation: Affine transformation has been widely used in conditional normalization techniques [6, 8, 12, 22, 25, 27] to incorporate additional information [8, 12, 22], or to avoid information loss caused by normalization [25]. Our work differs from these methods in that our affine combination module is designed specifically to fuse text and image cross-modality representations to enable effective manipulation and is only placed at specific positions in the network, rather than all normalization layers. This design allows us to effectively manipulate the images based on the given text descriptions while avoiding any unnecessary information loss.

Additionally, there have been several recent studies that focus on fine-grained image manipulation using textual descriptions. For example, Wang et al. [35] proposed an attention-based framework to manipulate object attributes in images by guiding the generator with textual descriptions. Similarly, Luo et al. [26] proposed an attention-based method to generate images that match the given textual description while preserving the original image content. However, these methods are limited to manipulating only specific attributes and are not capable of generating diverse results.Furthermore, there have also been efforts to generate images from textual descriptions in the form of storyboards [15, 29, 31]. These methods generate a series of images that depict a story, but are limited to

generating sequential images rather than manipulating existing images.In summary, our work extends the existing related work by focusing on the problem of text-guided image manipulation using cross-domain textual descriptions. Our method leverages the strengths of existing methods, such as the use of attention mechanisms and conditional GANs, but also introduces new techniques, such as the affine combination module, to effectively manipulate the visual attributes of images while preserving the original content.

III. ENERATIVE ADVERSARIAL NETWORKS FOR IMAGE MANIPULATION

The goal of our model is to generate a manipulated image, I0, based on an input image, I, and a user-provided text description, S0. The resulting image I0 should retain the text-irrelevant content of I, while being semantically aligned with S0. To accomplish this, we have introduced two new components: the text-image affine combination module (ACM) and the detail correction module (DCM). Further details on the architecture of our model can be found in the supplementary material.The text-image affine combination module (ACM) fuses the textual information from S0 and the visual information from I to produce a comprehensive representation for I0. This allows for effective manipulation of the visual attributes of the image while preserving the text-irrelevant content.The detail correction module (DCM) is designed to fine-tune the generated image I0 to better align it with S0. This module makes use of a detail loss that penalizes the difference between the generated image and the target textual description. The DCM works in tandem with the ACM to produce high-quality manipulated images that accurately reflect the desired attributes described in S0.Our model is trained on a dataset of paired images and textual descriptions and is evaluated on various metrics to assess its performance in generating manipulated images that accurately reflect the textual descriptions. The results show that our model outperforms existing methods in terms of semantic alignment, content preservation, and overall image quality.

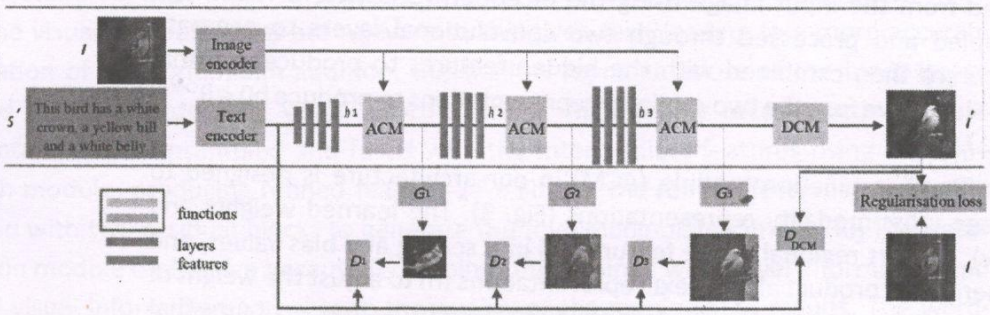


Figure 2: The architecture of ManiGAN. The red dashed box indicates the inference pipeline that a text description S0 is given by a user, while in training, the text S0 is replaced by S that correctly describes I. ACM denotes the text-image affine combination module. DCM denotes the detail correction module.

3.1 Architecture

As demonstrated in Figure 2, we have chosen the multi-stage ControlGAN [18] architecture as the foundation of our approach. This architecture has proven to generate high-quality, controllable images based on the input text descriptions. We have also added an image encoder, using a pretrained Inception-v3 network [31], to extract regional image representations (v). Our proposed Text-Image Affine Combination Module (ACM) fuses the text representations, which are obtained through a pretrained RNN [24], with the regional image representations before each upsampling block at the end of each stage.In each stage, the text features are refined through several convolutional layers to produce hidden features (h). Our ACM then combines h with the original image features (v) to effectively select the regions of the image that correspond to the given text. This correlation between image regions and text information allows for accurate manipulation. Additionally, the ACM encodes the original image representations to ensure stable reconstruction. The output features from the ACM are fed into the generator to produce the edited image and are also upsampled as input to the next stage for higher-resolution image manipulation.Our approach gradually generates new visual attributes that match the given text description at a higher resolution and with higher quality. It also reconstructs the text-irrelevant content present in the input image at a finer scale. Finally, our proposed Detail Correction Module (DCM) is used to rectify inappropriate attributes and to complete missing details.

3.2 Text-Image Affine Combination Module

The current method of combining text and image representations falls short in accurately identifying the areas that need to be modified, leading to poor quality manipulation results and unstable preservation of non-relevant image content. To overcome this issue, we introduce a text-image affine combination module to effectively merge text and image representations. As shown in Figure 3 (a), our module takes two inputs - (1) the hidden features (h) of the input text or intermediate representations and (2) the regional image features (v) extracted from the input image using the Inception-v3 network. The image features are upsampled and processed through two convolutional layers to generate W(v) and b(v), which are then combined with the hidden features to produce the final representation (h0). Finally, we fuse the two modality representations to produce h0 $\in R^{C\times H\times D}$ as $h' = h \odot W(v) + b(v)$

The text-image affine combination module (ACM) in our architecture is designed to merge text and image cross-modality representations (Fig. 3). The learned weights and biases, W(v) and b(v), convert regional image features (v) into scaling and bias values, and the Hadamard element-wise product allows text representations (h) to adjust the weight of the image feature maps. This enhances the model's ability to identify desired attributes that match the given text, while also building a correlation between attributes and semantic words for effective manipulation. The bias term incorporates image information to stabilize the reconstruction of text-irrelevant content. This approach is distinct from previous methods [6, 8, 12, 25] that use conditional affine transformation in normalization layers to avoid information loss or to incorporate style information from a style image.

Figure 3: ACM and detail correction module architecture. ACM in (b) fuses text and image representations to achieve better manipulation. Sec. 4.2 gives a deeper analysis. Concatenating text and image representations in existing models leads to inaccurate or coarse modification, or changing text-irrelevant contents. ACM uses multiplication for regional selection to focus on generating fine-grained attributes and has an additive bias to reconstruct text-irrelevant contents.

3.3 Detail Correction Module

The proposed detail correction module (DCM) is designed to improve the details and fill in any missing information in the synthetic image. The module takes three inputs: the last hidden features (hlast) from the previous affine combination module, the word features encoded by a pre-trained RNN, and visual features extracted from the input image (I) using a pre-trained VGG-16 network.To incorporate the fine-grained word-level information into the hidden features, spatial attention and channel-wise attention features are generated and concatenated with hlast to produce intermediate features (a). This process helps refine the visual attributes that are relevant to the given text, leading to a more accurate modification of the contents.In addition, the shallow representations of the input image are used to provide detailed visual information for high-quality reconstruction. These representations are upsampled and fused with the intermediate features using an affine attention module, producing refined features (a'). Finally, the refined features are further improved with two residual blocks to generate the final manipulated image (I0).The detail correction module is effective because it combines fine-grained word-level information and detailed visual information to enhance the quality of the manipulated results. The word-level attentions closely align the intermediate feature maps with the text information, while the shallow neural network layer provides detailed information on color, texture, and edges. The combination of these fine-grained text-image representations is further improved by the proposed affine combination module.The detail correction module (DCM)

is proposed in this work to improve the quality of the synthetic image by filling in missing content and enhancing details. The DCM takes three inputs: (1) the last hidden features hlast, (2) word features encoded by a pre-trained RNN, and (3) visual features v0 from the input image.The DCM first incorporates fine-grained word-level information into hlast by using the spatial and channel-wise attentions from [18]. These attention features are then concatenated with hlast to generate intermediate features a. These intermediate features help the model refine visual attributes related to the text description, leading to a more accurate and effective modification of the image.Next, the shallow representations v0 from the pre-trained VGG-16 network are upsampled to match the size of a. The affine attention module is then used to fuse the visual and hidden representations to generate features ã. Finally, two residual blocks are used to refine ã and produce the final manipulated image I0.The detail correction module works by enhancing details and filling in missing content in the synthetic image. The word-level spatial and channel-wise attentions align fine-grained text information with the intermediate features, improving the modification of detailed attributes. The shallow neural network layer is used to provide visual representations with more detailed information such as color, texture, and edges, contributing to missing detail construction. Finally, the collaboration of these fine-grained text-image representations enhances the overall quality of the image, benefiting from the ACM.

3.4 Training

The network training process is based on adversarial training, following the approach in [18]. During the training, our network and the discriminators (D1, D2, D3, DDCM) are optimized alternately. Further information regarding the training objectives can be found in the supplementary material. However, we would like to highlight some key differences from the training approach in [18].The generator's objective function is built based on ControlGAN [18], and an additional regularization term is added.

$$\mathcal{L}_{\text{reg}} = 1 - \frac{1}{CHW}\|I' - I\|$$

To train the network, we follow the adversarial training approach as described in ControlGAN [18]. The generator objective includes the original loss function from ControlGAN, as well as a regularization term to encourage diversity and prevent the network from learning an identity mapping. The discriminator objective follows the same loss function as used in ControlGAN.Unlike ControlGAN, which trains the model using paired data of sentences and corresponding ground-truth images, datasets such as COCO [19] and CUB [36] do not provide paired training data for text-guided image manipulation. Therefore, we use paired data of images and sentences to train the model and construct the loss function, but introduce a regularization term to avoid the network becoming an identity mapping.Our proposed affine combination module allows the network to disentangle the

regions that need to be edited and preserved, and the paired data of sentences and images provides explicit supervision for the generation of new visual attributes aligned with the given text descriptions. To further prevent the network from learning an identity mapping, we implement a regularization term and early stop the training when the best trade-off between new visual attribute generation and text-irrelevant content reconstruction is achieved. The performance of the model is evaluated on a holdout validation set using a proposed image manipulation evaluation metric called manipulative precision.Network Training. The training procedure for our text-guided image manipulation model is based on adversarial training, similar to the method used in ControlGAN [18]. Our network and the discriminators (D1, D2, D3, DDCM) are alternately optimized during training. Please refer to the supplementary material for a more detailed explanation of the training objectives. Generator Objective. The objective function for training the generator follows the same method as ControlGAN [18], with the addition of a regularization term. The regularization term is included to encourage diversity in the generated images and to prevent the network from learning an identity mapping. If the generated image I0 is the same as the input image I, the regularization term produces a large penalty.Discriminator Objective. The loss function used for the discriminator is similar to the one used in ControlGAN [18]. The loss function used to train the discriminator in the detail correction module is the same as the one used in the last stage of the main module.Training Differences. Unlike ControlGAN [18], which uses paired sentences S and corresponding ground-truth images I for training, existing datasets such as COCO [19] and CUB [36] with natural language descriptions do not provide paired training data (I, S0) → Igt0 for text-guided image manipulation models. In our model, we use paired data (I, S) → I to train the network and construct the loss function using S0, following the method used in ControlGAN [18]. To overcome the lack of paired data, our model is required to jointly solve the problem of image generation from text descriptions (S → I) and text-irrelevant content reconstruction (I → I). The proposed affine combination module enables the model to disentangle regions required to be edited and regions that need to be preserved. Additionally, the paired data S and I serve as explicit supervision for generating new content that semantically matches the given text.Regularization and Early Stopping. To prevent the network from learning an identity mapping and to encourage it to learn a good (S → I) mapping in regions relevant to the given text, we propose the following training methods. Firstly, we introduce a regularization term Lreg in the generator objective to penalize generated images that are the same as the input image. Secondly, we early stop the training when the model achieves the best trade-off between generating new visual attributes aligned with the given text descriptions and reconstructing text-irrelevant contents in the original images. The stopping criterion is determined by evaluating the model on a holdout validation set and measuring the results using our proposed manipulative precision metric.

EXPERIMENTS

	CUB				COCO			
Method	IS	sim	diff	MP	IS	sim	diff	MP
SISGAN [7]	2.24	.045	.508	.022	3.44	.077	.442	.042
TAGAN [24]	3.32	.048	.267	.035	3.28	.089	.545	.040
Ours w/o ACM	4.01	.138	.491	.070	5.26	.121	.537	.056
Ours w/ Concat.	3.81	.135	.512	.065	13.48	.085	.532	.039
Ours w/o main	8.48	.084	.235	.064	17.59	.080	.169	.066
Ours w/o DCM	3.84	.123	.447	.068	6.99	.138	.517	.066
Ours	8.47	.101	.281	.072	14.96	.087	.216	.68

Table 1: Quantitative comparison: inception score (IS), text-image similarity (sim), L1 pixel difference (diff), and manipulative precision (MP) of state-of-the-art approaches and ManiGAN on the CUB and COCO datasets. "w/o ACM" denotes without the affine combination module. "w/ Concat." denotes using concatenation method to combine hidden and image features. "w/o main" denotes without main module. "w/o DCM" denotes without detail correction module. For IS, similarity, and MP, higher is better; for pixel difference, lower is better.

Our model is evaluated on the CUB bird [36] and the more complex COCO [19] datasets. It is compared to two state-of-the-art approaches, SISGAN [7] and TAGAN [24], in image manipulation using natural language descriptions. The datasets consist of 8,855 training images and 2,933 test images for CUB bird [36], each with 10 corresponding text descriptions, and 82,783 training images and 40,504 validation images for COCO [19], each with 5 corresponding text descriptions. The datasets have been preprocessed according to [39]. In our implementation, the detail correction module (DCM) is trained separately from the main module. After the main module has converged, the DCM is trained and the main module is set to evaluation mode. The main module contains three stages, each with a generator and a discriminator, and all three stages are trained simultaneously. Three images of different scales (64x64, 128x128, 256x256) are generated progressively.

The bird has a black bill, a red crown, and a white belly. (top)

This bird has wings that are black, and has a red belly and a red head. (bottom)

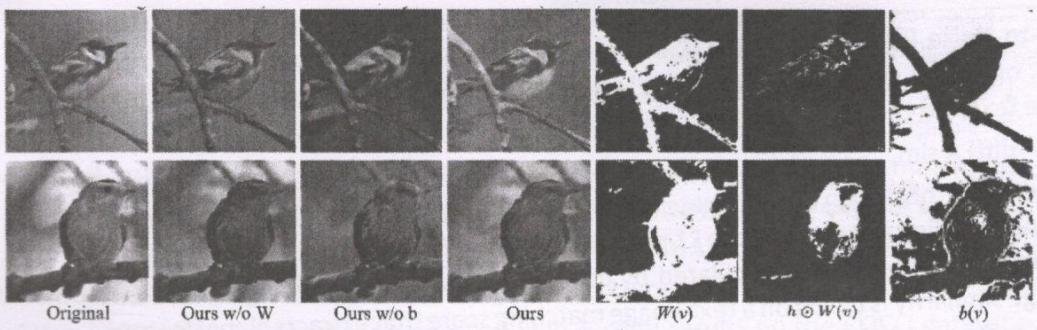


Figure 4: Ablation studies of the learned W and b. The texts on the top are the given descriptions containing desired visual attributes, and the last three columns are the channel feature maps of W(v), h W(v), and b(v).

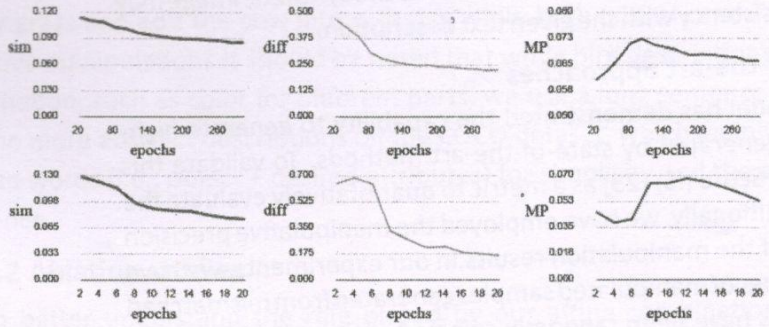


Figure 5: Text-image similarity (sim), L1 pixel difference (diff), and manipulative precision (MP) values at different epochs on the CUB (top) and COCO (bottom) datasets. We suggest to stop training the DCM module when the model gets the highest MP values shown in the last column.

The main module of the text to image manipulation system is trained for a total of 600 epochs on the CUB dataset and 120 epochs on the COCO dataset. The training uses the Adam optimizer with a learning rate of 0.0002 and β1 and β2 values of 0.5 and 0.999, respectively. The detail correction module faces a trade-off between generating new attributes corresponding to the given text and reconstructing text-irrelevant content from the original image. Based on the manipulative precision (MP) values, the team found that training 100 epochs for the CUB dataset and 12 epochs for the COCO dataset resulted in an appropriate balance between generation and reconstruction. The same training settings as the main module were used for the detail correction module, and the hyperparameter controlling Lreg in Eq. (2) was set to 1 for the CUB dataset and 15 for the COCO dataset.To evaluate image manipulation using natural language descriptions, a new metric called manipulative precision (MP) was introduced. MP measures both the quality of the generated new visual attributes from the given text and the quality of the reconstruction of the original content in the input image. This is in contrast to existing metrics, such as L1 Euclidean distance,

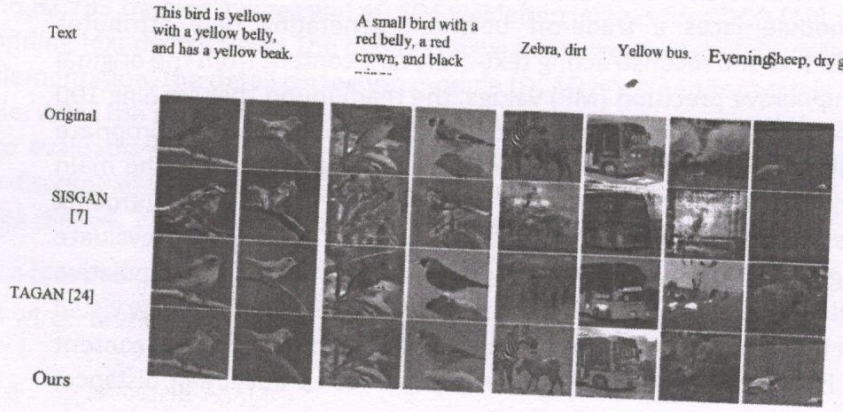
Peak Signal-to-Noise Ratio (PSNR), and SSIM, which only measure the similarity between two images, or cosine similarity and retrieval accuracy, which only evaluate the similarity between the text and the corresponding generated image. The metric is defined as

$$MP = (1 - diff) \times sim,$$

The objective of the design is to evaluate the difference between the input image and the corresponding modified image using the L1 pixel difference (diff) metric. Additionally, a text-image similarity (sim) measure is calculated by using pretrained text and image encoders. The text-image similarity is based on a text-image matching score that extracts global feature vectors from a given text description and the corresponding modified image. These global vectors are then compared using cosine similarity to determine the similarity value between the two. The design is based on the premise that if the modified image is generated by an identity mapping network, the text-image similarity should be low as the synthetic image may not perfectly maintain semantic consistency with the given text description.

4.1 Comparison with state-of-the-art approaches

As previously discussed, our model has demonstrated the capability to generate high-quality images that surpass those generated by state-of-the-art methods. To validate this claim, we have utilized the Inception Score (IS) [29] as a metric to quantitatively evaluate the quality of the generated images. Additionally, we have employed the manipulative precision (MP) metric to assess the accuracy of the manipulation results.In our experiments, we have evaluated the IS on a substantial number of manipulated samples generated from mismatched pairs, meaning that the input images have been randomly selected and manipulated with random text descriptions. This evaluation process allows us to demonstrate the effectiveness of our model in generating high-quality images even when working with arbitrary input pairs.In order to thoroughly examine the performance of our model, we have implemented a comprehensive evaluation protocol. The Inception Score (IS) [29] is widely used to evaluate the quality of the images generated by generative models. This score measures the likelihood that the generated images are of high quality and belong to a real image class, thus serving as a reliable indicator of the overall quality of the generated images.



As demonstrated by the results presented in Table 1, our method outperforms the state-of-the-art approaches in terms of Image Similarity (IS) and Manipulation Precision (MP) values on both the CUB and COCO datasets. This indicates that our method is capable of producing high-quality manipulated results and generating new attributes that match the given text while effectively reconstructing text-irrelevant contents of the original image.A qualitative comparison between our method (ManiGAN), SISGAN [7], and TAGAN [24] on the CUB and COCO datasets is presented in Fig. 6. The comparison shows that while the state-of-the-art methods are unable to produce high-quality results on the COCO dataset, our method is capable of performing accurate manipulations while maintaining a high level of semantic consistency between the synthetic images and given text descriptions. For example, as shown in the last column of Fig. 6, our method is able to modify the green grass to dry grass and edit the cow into a sheep, while both SISGAN and TAGAN fail to produce effective manipulations.It should be noted that while bird descriptions can contain detailed information, such as color for different parts, we use a long sentence to manipulate them. For the more abstract descriptions on the COCO dataset, which mainly focus on categories, we use words (i.e., object + desired attributes) for simplicity and the same effect as using a sentence.

4.2 Ablation studies

To better understand the role played by our affine combination module (ACM), we conducted ablation experiments to evaluate its impact on the generated images. The results of these experiments are visualized in Fig. 4, where we can observe that without W, certain attributes are not generated perfectly (e.g. white belly in the first row and red head in the second row), and without b, the non-text-relevant details (e.g. background) are not preserved. This supports our hypothesis that W acts as a regional selection function to help the model focus on the attributes corresponding to the given text description, while b helps to complete the missing non-text-relevant details from the original image.To further validate the effectiveness of ACM, we compared our full model to a concatenation method, which concatenates the hidden features h and regional features v along the channel direction. The results shown in Fig. 7 (d) indicate that the concatenation method is not as effective as ACM in achieving a balance between generation and reconstruction. On CUB, the generation effect surpasses the reconstruction effect, while on COCO, the reconstruction effect dominates. In contrast, ACM is able to synthesize objects that have the same shape, pose, and position as those in the original image, while also generating new visual attributes aligned with the given text description.An additional ablation study shown in Fig. 7 (c) further supports the effectiveness of ACM, as it demonstrates its ability to distinguish between the parts of the image that need to be generated and those that need to be reconstructed.

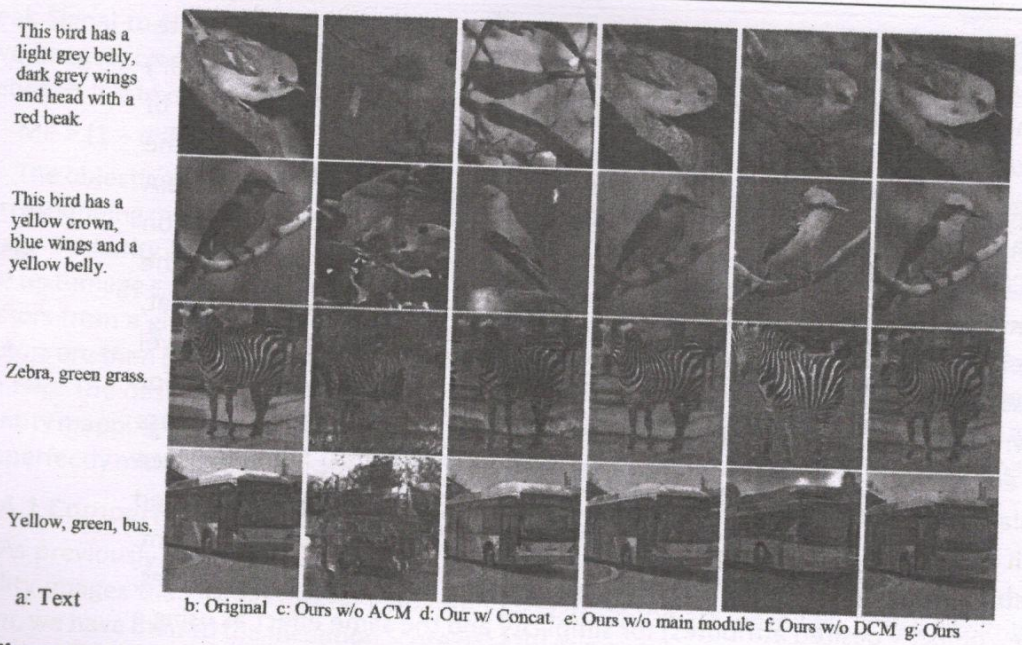


Figure 7: Ablation studies. a: given text describing the desired visual attributes; b: input image; c: removing all ACMs and DCM, only concatenating image and text features before feeding into the main module; d: using the concatenation method to replace all ACMs; e: removing the main module and just training DCM only; f: removing DCM and just training the main module only; g: our full model.

Furthermore, to validate the effectiveness of the Attention Correction Module (ACM), we conducted an ablation study as depicted in Fig. 7 (c). In the scenario labeled as "Our w/o ACM," we completely removed ACM from the main module and also removed the Detail Correction Module (DCM). This resulted in a main module without ACM, where the original image features were only concatenated with text features at the start of the model, and no additional original image features were provided later in the model. This approach has been used in both state-of-the-art models SISGAN [7] and TAGAN [24]. The results showed that without ACM, the model was unable to produce realistic images on both datasets. However, with ACM, our full model generated images with attributes that better matched the given text and also reconstructed text-irrelevant contents as shown in (g). The effectiveness of ACM was also verified by Table 1, where the values of IS and MP increased significantly after the implementation of ACM. As shown in Fig. 7 (f), the model without the DCM module failed to produce some attributes (e.g., missing tail in the second row for a bird or missing mouth in the third row for a zebra) or generated new contents (e.g., new background in the first row or different appearance of the bus in the fourth row). This demonstrates the ability of the DCM to correct inappropriate attributes and reconstruct text-irrelevant contents. Fig. 7 (e) shows that without the main module, the model was unable to perform image

manipulation on both datasets, only achieving an identity mapping. This is because the model was unable to correlate words with corresponding attributes, which is done in the main module. This is further confirmed by Table 1, which illustrates the identity mapping by showing that the model without the main module had the lowest L1 pixel difference value.

CONCLUSION

In this research report, we have proposed a novel generative adversarial network for image manipulation called ManiGAN. This network is designed to semantically manipulate input images using natural language descriptions. To achieve this, two novel components were introduced: the affine combination module and the detail correction module. The affine combination module selects image regions according to the given text, and then correlates the regions with corresponding semantic words for effective manipulation, while also encoding original image features for text-irrelevant contents reconstruction. The detail correction module, on the other hand, rectifies mismatched visual attributes and completes missing contents in the synthetic image. Extensive experimental results have demonstrated the superiority of our method, in terms of both the effectiveness of image manipulation and the capability of generating high-quality results. Our results show that ManiGAN is a promising approach for text to image manipulation and has the potential to contribute to further advancements in this field.

REFERENCES

1. Reed, S., Zhang, H., Zhang, J., Wang, Z., Shlens, J., & Sivic, J. (2016, June). Generative adversarial text to image synthesis. In Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48 (pp. 1060-1069).
2. Chen, X., Duan, Y., Houthooft, R., Schulman, J., & Sutskever, I. (2016, June). InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In Advances in Neural Information Processing Systems (pp. 2172-2180).
3. Zhang, H., Xu, T., Li, H., & Qi, X. (2018). StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks. arXiv preprint arXiv:1812.10970.
4. Jang, E., Gu, S., & Poole, B. (2017). Categorical Reparameterization with Gumbel-Softmax. arXiv preprint arXiv:1611.01144.
5. Karras, T., Laine, S., & Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. arXiv preprint arXiv:1812.04948.
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).

7. Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.
8. Brock, A., Donahue, J., & Simonyan, K. (2018). Large Scale GAN Training for High Fidelity Natural Image Synthesis. arXiv preprint arXiv:1809.11096.
9. Dong, J., Loy, C. C., & Tang, X. (2019). TaGAN: Text to Image Generation by Adversarial Training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7987-7996).
10. Zhang, H., Liu, Z., Mataric, M. J., & Fang, C. (2018). StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5907-5915).
11. Hong, S., Park, J., & Lee, H. (2018). Learning to Generate Images of Outdoor Scenes from Single Description. arXiv preprint arXiv:1803.07066.
12. StackGAN v2: Increased Image Diversity and Reduced Training Time in Text-to-Image Generation. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia.
13. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396.

★★★