

ManiGAN: Text-Guided Image Manipulation

Bowen Li¹ Xiaojuan Qi^{1,2} Thomas Lukasiewicz¹ Philip H. S. Torr¹
¹University of Oxford ²University of Hong Kong

{bowen.li, thomas.lukasiewicz}@cs.ox.ac.uk {xiaojuan.qi, philip.torr}@eng.ox.ac.uk

Abstract

The goal of our paper is to semantically edit parts of an image matching a given text that describes desired attributes (e.g., texture, colour, and background), while preserving other contents that are irrelevant to the text. To achieve this, we propose a novel generative adversarial network (ManiGAN), which contains two key components: text-image affine combination module (ACM) and detail correction module (DCM). The ACM selects image regions relevant to the given text and then correlates the regions with corresponding semantic words for effective manipulation. Meanwhile, it encodes original image features to help reconstruct text-irrelevant contents. The DCM rectifies mismatched attributes and completes missing contents of the synthetic image. Finally, we suggest a new metric for evaluating image manipulation results, in terms of both the generation of new attributes and the reconstruction of text-irrelevant contents. Extensive experiments on the CUB and COCO datasets demonstrate the superior performance of the proposed method. Code is available at <https://github.com/mrlibw/ManiGAN>.

1. Introduction

Image manipulation aims to modify some aspects of given images, from low-level colour or texture [9, 40] to high-level semantics [41], to meet a user’s preferences, which has numerous potential applications in video games, image editing, and computer-aided design. Recently, with the development of deep learning and deep generative models, automatic image manipulation has made remarkable progress, including image inpainting [12, 25], image colourisation [40], style transfer [9, 14], and domain or attribute translation [13, 16].

All the above works mainly focus on specific problems, and few studies [6, 23] concentrate on more general and user-friendly image manipulation by using natural language descriptions. More precisely, the task aims to semantically edit parts of an image according to the given text provided by a user, while preserving other contents that are not described in the text. However, current state-of-the-art text-

guided image manipulation methods are only able to produce low-quality images (see Fig. 1: first row), far from satisfactory, and even fail to effectively manipulate complex scenes (see Fig. 1: second row).

To achieve effective image manipulation guided by text descriptions, the key is to exploit both text and image cross-modality information, generating new attributes matching the given text and also preserving text-irrelevant contents of the original image. To fuse text and image information, existing methods [6, 23] typically choose to directly concatenate image and global sentence features along the channel direction. Albeit simple, the above heuristic may suffer from some potential issues. Firstly, the model cannot precisely correlate fine-grained words with corresponding visual attributes that need to be modified, leading to inaccurate and coarse modification. For instance, shown in the first row of Fig. 1, both models cannot generate detailed visual attributes like *black eye rings* and a *black bill*. Secondly, the model cannot effectively identify text-irrelevant contents and thus fails to reconstruct them, resulting in undesirable modification of text-irrelevant parts in the image. For example, in Fig. 1, besides modifying the required attributes, both models [6, 23] also change the texture of the bird (first row) and the structure of the scene (second row).

To address the above issues, we propose a novel generative adversarial network for text-guided image manipulation (ManiGAN), which can generate high-quality new attributes matching the given text, and at the same time effectively reconstruct text-irrelevant contents of the original image. The key is a text-image affine combination module (ACM) where text and image features collaborate to select text-relevant regions that need to be modified, and then correlate those regions with corresponding semantic words for generating new visual attributes semantically aligned with the given text description. Meanwhile, it also encodes original image representations for reconstructing text-irrelevant contents. Besides, to further enhance the results, we introduce a detail correction module (DCM) which can rectify mismatched attributes and complete missing contents. Our final model can produce high-quality manipulation results with fine-grained details (see Fig. 1: Ours).

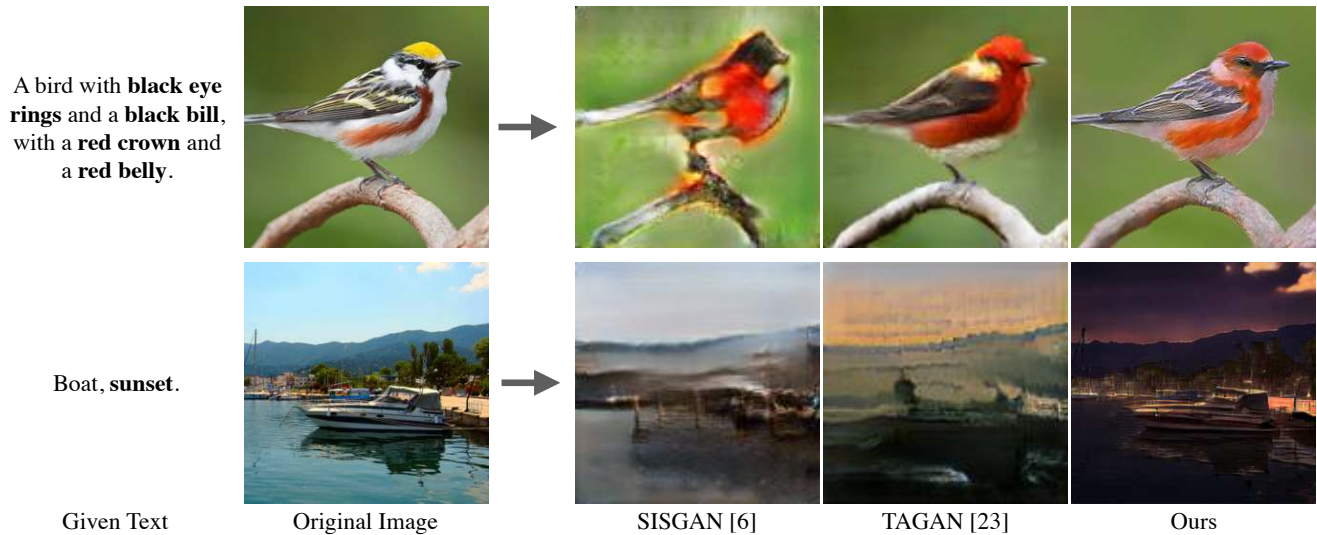


Figure 1: Given an original image that needs to be edited and a text provided by a user describing desired attributes, the goal is to edit parts of the image according to the given text while preserving text-irrelevant contents. Current state-of-the-art methods only generate low-quality images, and fail 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.

Finally, we suggest a new metric to assess image manipulation results. The metric can appropriately reflect the performance of image manipulation, in terms of both the generation of new visual attributes corresponding to the given text, and the reconstruction of text-irrelevant contents of the original image. Extensive experiments on the CUB [34] and COCO [18] datasets demonstrate the superiority of our model, where our model outperforms existing state-of-the-art methods both qualitatively and quantitatively.

2. Related Work

Text-to-image generation has drawn much attention due to the success of GANs [10] in generating realistic images. Reed et al. [27] proposed to use conditional GANs to generate plausible images from given text descriptions. Zhang et al. [38, 39] stacked multiple GANs to generate high-resolution images from coarse- to fine-scale. Xu et al. [37] and Li et al. [17] implemented attention mechanisms to explore fine-grained information at the word-level. However, all aforementioned methods mainly focus on generating new photo-realistic images from texts, and not on manipulating specific visual attributes of given images using natural language descriptions.

Conditional image synthesis. Our work is related to conditional image synthesis [1, 19, 22, 24, 32, 33, 41]. Recently, various methods have been proposed to achieve paired image-to-image translation [3, 13, 35], or unpaired translation [20, 31, 42]. However, all these methods mainly focus on same-domain image translation instead of image

manipulation using cross-domain text descriptions.

Text-guided image manipulation. There are few studies focusing on image manipulation using natural language descriptions [2, 4, 6, 8, 23]. Dong et al. [6] proposed a GAN-based encoder-decoder architecture to disentangle the semantics of both input images and text descriptions. Nam et al. [23] introduced a text-adaptive discriminator that can provide specific word-level training feedback to the generator. However, both methods are limited in performance due to a less effective text-image concatenation method and a coarse sentence condition.

Affine transformation has been widely implemented in conditional normalisation techniques [5, 7, 11, 21, 24, 26] to incorporate additional information [7, 11, 21], or to avoid information loss caused by normalisation [24]. Differently from these methods, our affine combination module is designed to fuse text and image cross-modality representations to enable effective manipulation, and is only placed at specific positions instead of all normalisation layers.

3. Generative Adversarial Networks for Image Manipulation

Given an input image I , and a text description S' provided by a user, the model aims to generate a manipulated image I' that is semantically aligned with S' while preserving text-irrelevant contents existing in I . To achieve this, we propose two novel components: (1) a text-image affine combination module (ACM), and (2) a detail correction module (DCM). We elaborate our model as follows.

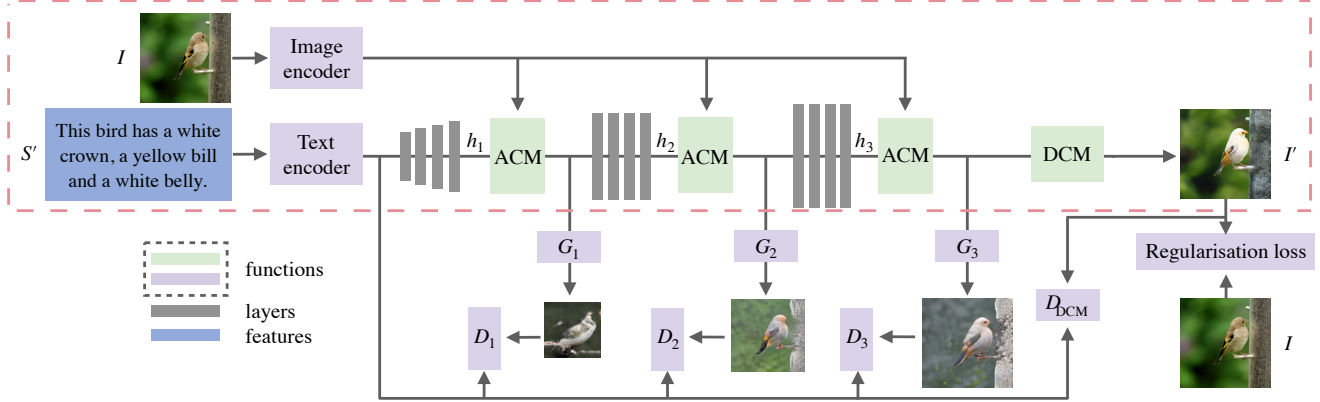


Figure 2: The architecture of ManiGAN. The red dashed box indicates the inference pipeline that a text description S' is given by a user, while in training, the text S' is replaced by S that correctly describes I . ACM denotes the text-image affine combination module. DCM denotes the detail correction module. The attention is omitted for simplicity. Please see the supplementary material for the full architecture.

3.1. Architecture

As shown in Fig. 2, we adopt the multi-stage ControlGAN [17] architecture as the basic framework, as it achieves high-quality and controllable image generation results based on the given text descriptions. We add an image encoder, which is a pretrained Inception-v3 network [30], to extract regional image representations v . Our proposed text-image affine combination module (ACM) is utilised to fuse text representations, encoded from a pretrained RNN [23], and regional image representations before each upsampling block at the end of each stage. For each stage, the text features are refined with several convolutional layers to produce hidden features h . The proposed ACM further combines h with the original image features v in order to effectively select image regions corresponding to the given text, and then correlate those regions with text information for accurate manipulation. Meanwhile, it also encodes the original image representations for stable reconstruction. The output features from the ACM module are fed into the corresponding generator to produce an edited image, and are also upsampled serving as input to the next stage for image manipulation at a higher resolution. The whole framework gradually generates new visual attributes matching the given text description at a higher resolution with higher quality, and also reconstructs text-irrelevant contents existing in the input image at a finer scale. Finally, the proposed detail correction module (DCM) is used to rectify inappropriate attributes, and to complete missing details.

3.2. Text-Image Affine Combination Module

The existing concatenation scheme for combining text-image cross-modality representations cannot effectively locate desired regions that need to be modified, and thus fails

to achieve fine-grained image manipulation, regarding both the generation quality of new attributes corresponding to the given text, and the reconstruction stability of text-irrelevant image contents. To address the above issue, we propose a simple text-image affine combination module to fuse text-image cross-modality representations as discussed below.

As shown in Fig. 3 (a), our affine combination module takes two inputs: (1) the hidden features $h \in \mathbb{R}^{C \times H \times D}$ from the input text or intermediate hidden representations between two stages, where C is the number of channels, H and D are the height and width of the feature map, respectively, and (2) the regional image features $v \in \mathbb{R}^{256 \times 17 \times 17}$ from the input image I encoded by the Inception-v3 network [30]. Then, v is upsampled and further processed with two convolutional layers to produce $W(v)$ and $b(v)$ that have the same size as h . Finally, we fuse the two modality representations to produce $h' \in \mathbb{R}^{C \times H \times D}$ as

$$h' = h \odot W(v) + b(v), \quad (1)$$

where $W(v)$ and $b(v)$ are the learned weights and biases based on the regional image features v , and \odot denotes Hadamard element-wise product. We use W and b to represent the functions that convert the regional features v to scaling and bias values.

Our affine combination module (ACM) is designed to fuse text and image cross-modality representations. $W(v)$ and $b(v)$ encode the input image into semantically meaningful features as shown in Fig. 4. The multiplication operation enables text representations h to re-weight image feature maps, which serves as a regional selection purpose to help the model precisely identify desired attributes matching the given text, and in the meantime the correlation between attributes and semantic words is built for effective manipulation. The bias term encodes image information to help

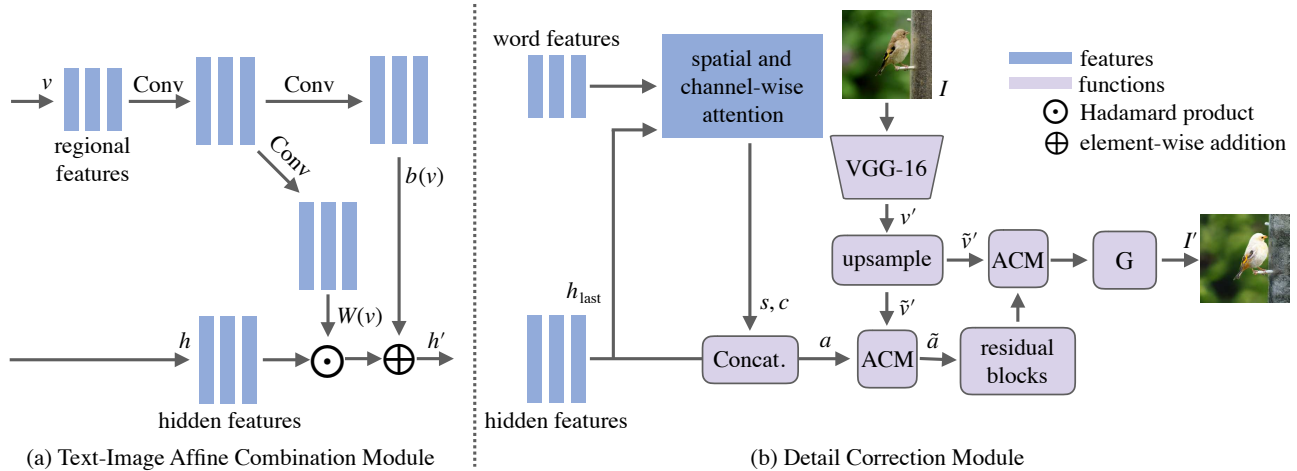


Figure 3: The architecture of the text-image affine combination module and the detail correction module. In (b), ACM denotes the text-image affine combination module.

the model stably reconstruct text-irrelevant contents. The above is in contrast with previous approaches [5, 7, 11, 24] which apply conditional affine transformation in normalisation layers to compensate potential information loss due to normalisation [24] or to incorporate style information from a style image [7, 11]. To better understand what has been actually learned by different components of our affine combination module, we give a deeper analysis in Sec. 4.2.

Why does the affine combination module work better than concatenation? By simply concatenating the text and image representations along the channel direction, existing models cannot explicitly distinguish regions that are required to be modified or to be reconstructed, which makes it hard to achieve a good balance between the generation of new attributes and the reconstruction of original contents. As a result, this imbalance leads to either inaccurate/coarse modification or changing text-irrelevant contents. In contrast, our affine combination module uses multiplication on text and image representations to achieve a regional selection effect, aiding the model to focus on generating required fine-grained visual attributes. Besides, the additive bias part encodes text-irrelevant image information to help reconstruct contents that are not required to be edited.

3.3. Detail Correction Module

To further enhance the details and complete missing contents in the synthetic image, we propose a detail correction module (DCM), exploiting word-level text information and fine-grained image features.

As shown in Fig. 3 (b), our detail correction module takes three inputs: (1) the last hidden features $h_{\text{last}} \in \mathbb{R}^{C' \times H' \times D'}$ from the last affine combination module, (2) the word features encoded by a pretrained RNN following [37], where each word is associated with a feature vec-

tor, and (3) visual features $v' \in \mathbb{R}^{128 \times 128 \times 128}$ that are extracted from the input image I , which are the *relu2_2* layer representations from a pretrained VGG-16 [29] network.

Firstly, to further incorporate fine-grained word-level representations into hidden features h_{last} , we adopt the spatial attention and channel-wise attention introduced in [17] to generate spatial and channel-wise attention features $s \in \mathbb{R}^{C' \times H' \times D'}$ and $c \in \mathbb{R}^{C' \times H' \times D'}$, respectively, which are further concatenated with h_{last} to produce intermediate features a . The features a can further aid the model to refine visual attributes that are relevant to the given text, contributing to a more accurate and effective modification of the contents corresponding to the given description. Secondly, to introduce detailed visual features from the input image for high-quality reconstruction, the shallow representations v' of layer *relu2_2* from the pretrained VGG network are utilised, which are further upsampled to be the same size as a , denoted as \tilde{v}' . Then, our proposed affine attention module is utilised to fuse visual representations \tilde{v}' and hidden representations a , producing features \tilde{a} . Finally, we refine \tilde{a} with two residual blocks (details in the supplementary material) to generate the final manipulated image I' .

Why does the detail correction module work? This module aims to refine the manipulated results by enhancing details and completing missing contents. On the one hand, the word-level spatial and channel-wise attentions closely correlate fine-grained word-level information with the intermediate feature maps, enhancing the detailed attribute modification. On the other hand, the shallow neural network layer is adopted to derive visual representations, which contain more detailed colour, texture, and edge information, contributing to missing detail construction. Finally, further benefiting from our ACM, the above fine-grained text-image representations collaborate to enhance the quality.

3.4. Training

To train the network, we follow [17] and adopt adversarial training, where our network and the discriminators (D_1, D_2, D_3, D_{DCM}) are alternatively optimised. Please see the supplementary material for more details about training objectives. We only highlight some training differences compared with [17].

Generator objective. We follow the ControlGAN [17] to construct the objective function for training the generator. Besides, we add a regularisation term as

$$\mathcal{L}_{reg} = 1 - \frac{1}{CHW} \|I' - I\|, \quad (2)$$

where I is the real image sampled from the true image distribution, and I' is the corresponding modified result produced by our model. The regularisation term is used to ensure diversity and to prevent the network learning an identity mapping, since this term can produce a large penalty when the generated image I' is the same as the input image.

Discriminator objective. The loss function for the discriminator follows those used in ControlGAN [17], and the 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. Differently from [17], which has paired sentence S and corresponding ground-truth image I for training text-guided image generation models to learn the mapping $S \rightarrow I$, existing datasets such as COCO [18] and CUB [34] with natural language descriptions do not provide paired training data $(I, S') \rightarrow I'_{gt}$ for training text-guided image manipulation models, where S' is a text describing new attributes, and I'_{gt} is the corresponding ground-truth modified image.

To simulate the training data, we use paired data $(I, S) \rightarrow I$ to train the model, and adopt S' to construct the loss function following [17]. A natural question may arise: how does the model learn to modify the image I if the input image I and ground-truth image are the same, and the modified sentence S' does not exist in the input? In theory, the optimal solution is that the network becomes an identity mapping from the input image to the output. The text-guided image manipulation model is required to jointly solve image generation from text descriptions ($S \rightarrow I$), similarly to [17], and text-irrelevant contents reconstruction ($I \rightarrow I$). Thanks to our proposed affine combination module, our model gains the capacity to disentangle regions required to be edited and regions needed to be preserved. Also, to generate new contents semantically matching the given text, the paired data S and I can serve as explicit supervision.

Moreover, to prevent the model from learning an identity mapping and to promote the model to learn a good ($S \rightarrow I$) mapping in the regions relevant to the given text, we propose the following training schemes. Firstly, we introduce a

regularisation term \mathcal{L}_{reg} as Eq. (2) in the generator objective to produce a penalty if the generated image becomes the same as the input image. Secondly, we choose to early stop the training when the model achieves the best trade-off between the generation of new visual attributes aligned with the given text descriptions and the reconstruction of text-irrelevant contents existing in the original images. The stop criterion is determined by evaluating the model on a hold-out validation and measuring the results by our proposed image manipulation evaluation metric, called manipulative precision (see Fig. 5), which is discussed in Sec. 4.

4. Experiments

Method	CUB				COCO			
	IS	sim	diff	MP	IS	sim	diff	MP
SISGAN [6]	2.24	.045	.508	.022	3.44	.077	.442	.042
TAGAN [23]	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	.068

Table 1: Quantitative comparison: inception score (IS), text-image similarity (sim), L_1 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 the 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 [34] and more complicated COCO [18] datasets, comparing with two state-of-the-art approaches SISGAN [6] and TAGAN [23] on image manipulation using natural language descriptions.

Datasets. CUB bird [34]: there are 8,855 training images and 2,933 test images, and each image has 10 corresponding text descriptions. COCO [18]: there are 82,783 training images and 40,504 validation images, and each image has 5 corresponding text descriptions. We preprocess these two datasets according to the method in [37].

Implementation. In our setting, we train the detail correction module (DCM) separately from the main module. Once the main module has converged, we train the DCM subsequently and set the main module to the evaluation mode. There are three stages in the main module, and each stage contains a generator and a discriminator. We train the three stages at the same time, and three different-scale images $64 \times 64, 128 \times 128, 256 \times 256$ 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)

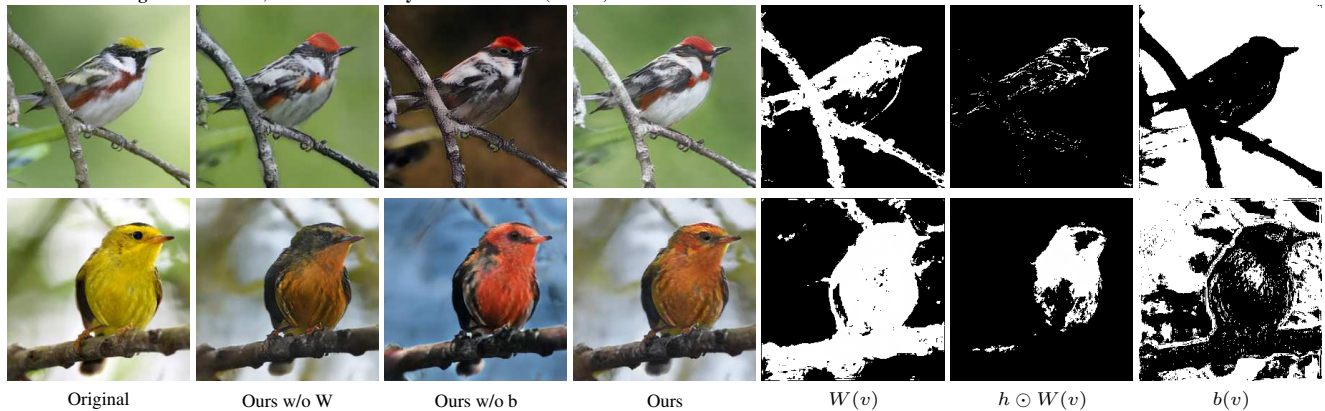


Figure 4: Ablation studies of the learned W and b . The texts on the top are the given descriptions containing desired visual attributes, and the last three columns are the channel feature maps of $W(v)$, $h \odot W(v)$, and $b(v)$.

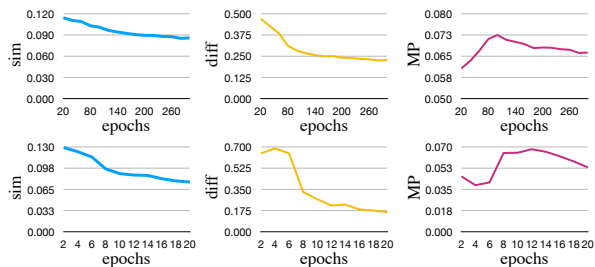


Figure 5: Text-image similarity (sim), L_1 pixel difference (diff), and manipulative precision (MP) values at different epochs on the CUB (top) and COCO (bottom) datasets. We suggest to stop training the DCM module when the model gets the highest MP values shown in the last column.

The main module is trained for 600 epochs on CUB and 120 epochs on COCO using the Adam optimiser [15] with the learning rate 0.0002, and $\beta_1 = 0.5$, $\beta_2 = 0.999$. As for the detail correction module, there is a trade-off between the generation of new attributes corresponding to the given text and the reconstruction of text-irrelevant contents of the original image. Based on the manipulative precision (MP) values (see Fig. 5), we find that training 100 epochs for CUB, and 12 epochs for COCO to achieve an appropriate balance between generation and reconstruction. The other training setting is the same as in the main module. The hyperparameter controlling \mathcal{L}_{reg} in Eq. (2) is set to 1 for CUB and 15 for COCO.

Manipulative precision metric. Image manipulation using natural language descriptions should be evaluated in terms of both the generation of new visual attributes from the given text, and the reconstruction of original contents existing in the input image. However, existing metrics only focus on one aspect of this problem. For example, the L_1 Eu-

clidean distance, Peak Signal-to-Noise Ratio (PSNR), and SSIM [36] only measure the similarity between two images, while the cosine similarity and the retrieval accuracy [17, 23, 37] only evaluate the similarity between the text and the corresponding generated image. Based on this, we contribute a new metric, called manipulative precision (MP), for this area to simultaneously measure the quality of generation and reconstruction. The metric is defined as

$$\text{MP} = (1 - \text{diff}) \times \text{sim}, \quad (3)$$

where diff is the L_1 pixel difference between the input image and the corresponding modified image, sim is the text-image similarity, which is calculated by using pretrained text and image encoders [37] based on a text-image matching score to extract global feature vectors of a given text description and the corresponding modified image, and then the similarity value is computed by applying the cosine similarity between these two global vectors. Specifically, the design is based on the intuition that if the manipulated image is generated from an identity mapping network, then the text-image similarity should be low, as the synthetic image cannot perfectly keep a semantic consistency with the given text description.

4.1. Comparison with state-of-the-art approaches

Quantitative comparison. As mentioned above, our model can generate high-quality images compared with the state-of-the-art methods. To demonstrate this, we adopt the inception score (IS) [28] as a quantitative evaluation measure. Besides, we adopt the manipulative precision (MP) to evaluate manipulation results. In our experiments, we evaluate IS on a large number of manipulated samples generated from mismatched pairs, i.e., randomly chosen input images manipulated by randomly selected text descriptions.

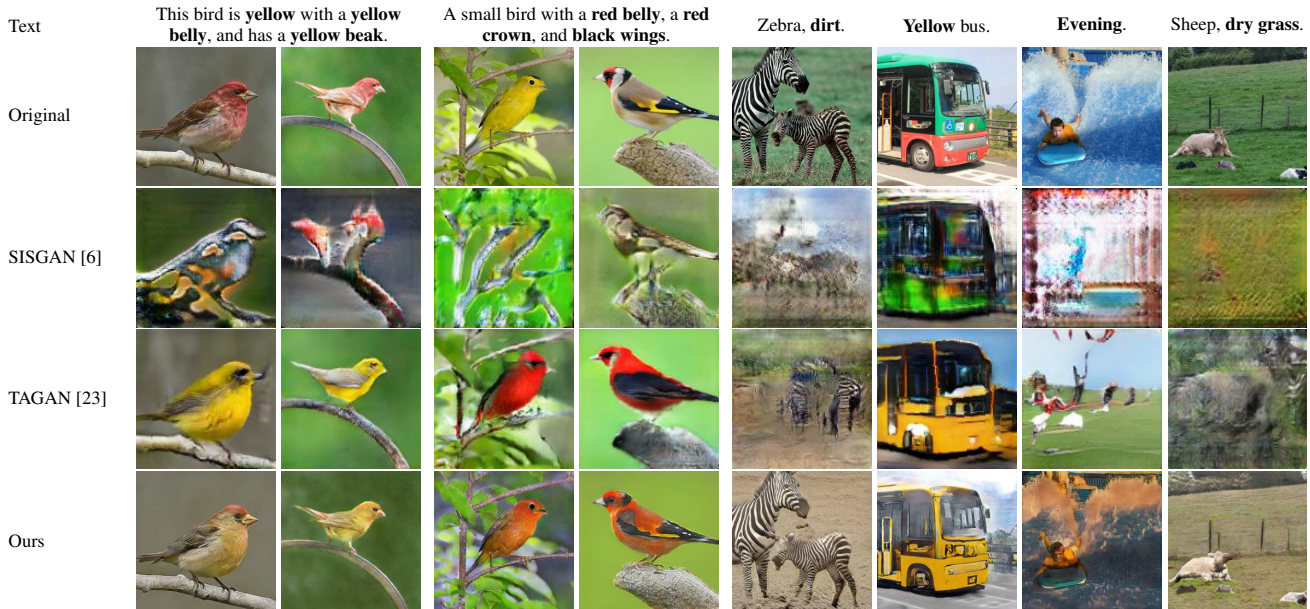


Figure 6: Qualitative comparison of three methods on the CUB bird and COCO datasets.

As shown in Table 1, our method has the highest IS and MP values on both the CUB and COCO datasets compared with the state-of-the-art approaches, which demonstrates that (1) our method can produce high-quality manipulated results, and (2) our method can better generate new attributes matching the given text, and also effectively reconstruct text-irrelevant contents of the original image.

Qualitative comparison. Fig. 6 shows the visual comparison between our ManiGAN, SISGAN [6], and TAGAN [23] on the CUB and COCO datasets. It can be seen that both state-of-the-art methods are only able to produce low-quality results and cannot effectively manipulate input images on the COCO dataset. However, our method is capable of performing an accurate manipulation and also keep a highly semantic consistency between synthetic images and given text descriptions, while preserving text-irrelevant contents. For example, shown in the last column of Fig. 6, SISGAN and TAGAN both fail to achieve an effective manipulation, while our model modifies the *green grass* to *dry grass* and also edits the *cow* into a *sheep*.

Note that as birds can have detailed descriptions (e.g., colour for different parts), we use a long sentence to manipulate them, while the descriptions for COCO are more abstract and focus mainly on categories, thus we use words (i.e., object + desired attributes) to do manipulation for simplicity, which has the same effect as using a sentence.

4.2. Ablation studies

Ablation experiments of the affine combination module.

To better understand what has been learned by our ACM, we ablate and visualise the learned feature maps shown in

Fig. 4. As we can see, without W , some attributes cannot be perfectly generated (e.g., white belly in the first row and red head in the second row), and without b , the text-irrelevant contents (e.g., background) are hard to preserve, which verifies our assumption that W behaves as a regional selection function to help the model focus on attributes corresponding to the given text, and b helps to complete missing text-irrelevant details of the original image. Also, the visualisation of the channel feature maps of $W(v)$, $h \odot W(v)$, and $b(v)$ shown in the last three columns of Fig. 4 validates the regional selection effect of the multiplication operation.

Effectiveness of the affine combination module. To verify the effectiveness of the ACM, we use the concatenation method to replace all ACMs, which concatenates hidden features h and regional features v along the channel direction, shown in Fig. 7 (d). As we can see, with the concatenation method, the model generates structurally different birds on CUB, and fails to do manipulation on COCO, which indicates that it is hard for the concatenation method to achieve a good balance between generation and reconstruction. The results on CUB is an example of the generation effect surpassing the reconstruction effect, while results on COCO show the domination of the reconstruction effect. In contrast, due to the regional selection effect of ACM that can distinguish which parts need to be generated or to be reconstructed, our full model synthesises an object having the same shape, pose, and position as the one existing in the original image, and also generates new visual attributes aligned with the given text description.

Also, to further validate the effectiveness of ACM, we conduct an ablation study shown in Fig. 7 (c). In “Our

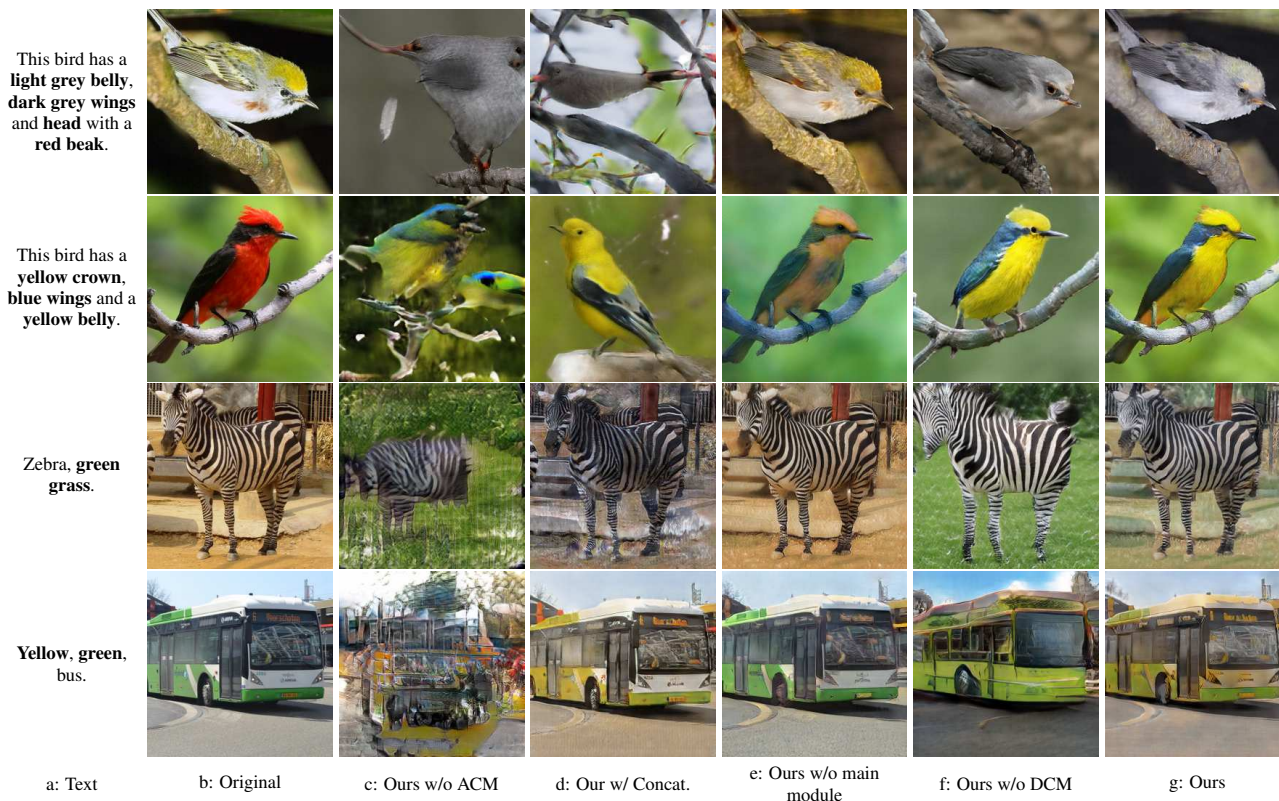


Figure 7: Ablation studies. a: given text describing the desired visual attributes; b: input image; c: removing all ACMs and DCM, only concatenating image and text features before feeding into the main module; d: using the concatenation method to replace all ACMs; e: removing the main module and just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

w/o ACM”, we fully remove ACM in the main module and remove DCM as well. That is the main module without ACM, and we only concatenate original image features with text features at the beginning of the model and do not further provide additional original image features in the middle of the model. This method is used in both state-of-the-art SISGAN [6] and TAGAN [23]. It can be seen that our model without ACM fails to produce realistic images on both datasets. In contrast, our full model better generates attributes matching the given text, and also reconstructs text-irrelevant contents shown in (g). Table 1 also verifies the effectiveness of our ACM, as the values of IS and MP increase significantly when we implement ACM.

Effectiveness of the detail correction module and main module. As shown in Fig. 7 (f), our model without DCM misses some attributes (e.g., the bird missing the tail in the second row, the zebra missing the mouth in the third row), or generates new contents (e.g., new background in the first row, different appearance of the bus in the fourth row), which indicates that our DCM can correct inappropriate attributes and reconstruct text-irrelevant contents. Fig. 7 (e) shows that without main module, our model fails to do im-

age manipulation on both datasets, which just achieves an identity mapping. This is mainly because the model fails to correlate words with corresponding attributes, which has been done in the main module. Table 1 also illustrates the identity mapping, as our model without main module gets the lowest L_1 pixel difference value.

5. Conclusion

We have proposed a novel generative adversarial network for image manipulation, called ManiGAN, which can semantically manipulate input images using natural language descriptions. Extensive experimental results demonstrate the superiority of our method, in terms of both the effectiveness of image manipulation and the capability of generating high-quality results.

Acknowledgements. This work was supported by the Alan Turing Institute under the UK EPSRC grant EP/N510129/1, the AXA Research Fund, the ERC grant ERC-2012-AdG 321162-HELIOS, EPSRC grant Seebibyte EP/M013774/1 and EP-SRC/MURI grant EP/N019474/1. We would also like to acknowledge the Royal Academy of Engineering and FiveAI.

References

- [1] Andrew Brock, Theodore Lim, James M. Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks. *arXiv preprint arXiv:1609.07093*, 2016.
- [2] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. Language-based image editing with recurrent attentive models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8721–8729, 2018.
- [3] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1511–1520, 2017.
- [4] Yu Cheng, Zhe Gan, Yitong Li, Jingjing Liu, and Jianfeng Gao. Sequential attention gan for interactive image editing via dialogue. *arXiv preprint arXiv:1812.08352*, 2018.
- [5] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C. Courville. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2017.
- [6] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5706–5714, 2017.
- [7] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016.
- [8] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W Taylor. Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10304–10312, 2019.
- [9] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.
- [10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.
- [11] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.
- [12] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (TOG)*, 35(4):110, 2016.
- [13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.
- [14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 694–711. Springer, 2016.
- [15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [16] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc’Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pages 5967–5976, 2017.
- [17] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. In *Advances in Neural Information Processing Systems*, pages 2063–2073, 2019.
- [18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014.
- [19] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.
- [20] Ming-Yu Liu and Oncl Tuzel. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 469–477, 2016.
- [21] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018.
- [22] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. InstaGAN: Instance-aware image-to-image translation. *arXiv preprint arXiv:1812.10889*, 2018.
- [23] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. In *Advances in Neural Information Processing Systems*, pages 42–51, 2018.
- [24] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.
- [25] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.
- [26] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [27] Scott Reed, Zeynep Akata, Xinchun Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.
- [28] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

- [29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.
- [31] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016.
- [32] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection GAN with cascaded semantic guidance for cross-view image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2417–2426, 2019.
- [33] Hao Tang, Dan Xu, Yan Yan, Philip Torr, and Nicu Sebe. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. *arXiv preprint arXiv:1912.12215*, 2019.
- [34] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. 2011.
- [35] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018.
- [36] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [37] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2018.
- [38] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5907–5915, 2017.
- [39] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1947–1962, 2018.
- [40] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision*, pages 649–666. Springer, 2016.
- [41] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *Proceedings of the European Conference on Computer Vision*, pages 597–613. Springer, 2016.
- [42] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.