Image-to-Image Translation with Text Guidance

Bowen Li[1]
bowen.li@cs.ox.ac.uk
Philip H. S. Torr[1]
philip.torr@eng.ox.ac.uk
Thomas Lukasiewicz[2,1]
thomas.lukasiewicz@cs.ox.ac.uk

[1] University of Oxford
Oxford, UK
[2] TU Wien
Vienna, Austria

Abstract

In this paper, we focus on image-to-image translation with text guidance, where a text description is used to control visual attributes of the synthetic image produced from a given semantic mask. To accomplish this task, we propose a new multi-stage generative adversarial network with three novel components: (1) a discriminator with dual-directional feedback, which provides the generator at the same stage with fine-grained supervisory feedback related to image regions, encouraging it to produce realistic images with finer regional details, and also facilitating generators at following stages to have the ability to complete missing contents and correct inappropriate visual attributes, (2) a compatibility loss guides generators to produce both realistic objects and the background, and also to achieve a good compatibility between them, and (3) a part-of-speech tagging-based spatial attention to better build connection between image regions and corresponding semantic words. Experimental results demonstrate that our model can effectively control the image translation using text descriptions. More importantly, the text input allows our model to produce much diverse results and even new synthetic images that are out-of-distribution of the dataset.

1 Introduction

Image-to-image translation aims to generate photo-realistic images from image conditions, such as coarse sketches [4, 6, 7, 9, 12], simple semantic layouts [10, 31, 38], and fine-grained pixel-level semantic maps [3, 12, 28, 29, 32, 33]. This task will stimulate applications in various areas, including video game creation, automatic art design, and image editing.

However, to satisfy users' preferences, these image conditions may not be rich enough to determine the content and style of the generated results. This means that users cannot freely design fine-grained visual attributes (e.g., color, texture, and style) of synthetic images, which is typically undesirable in real-world applications. As shown in Fig. 1, given only the semantic mask, users cannot determine the category, color, and background of an object. However, in reality, users usually have their preferences when they create an image, e.g., a user may intend to have "a giraffe on green grass" or "a yellow plane in the grey sky". Based on this, it is highly desirable to have a model, which allows users to have the ability to control the image generation process.

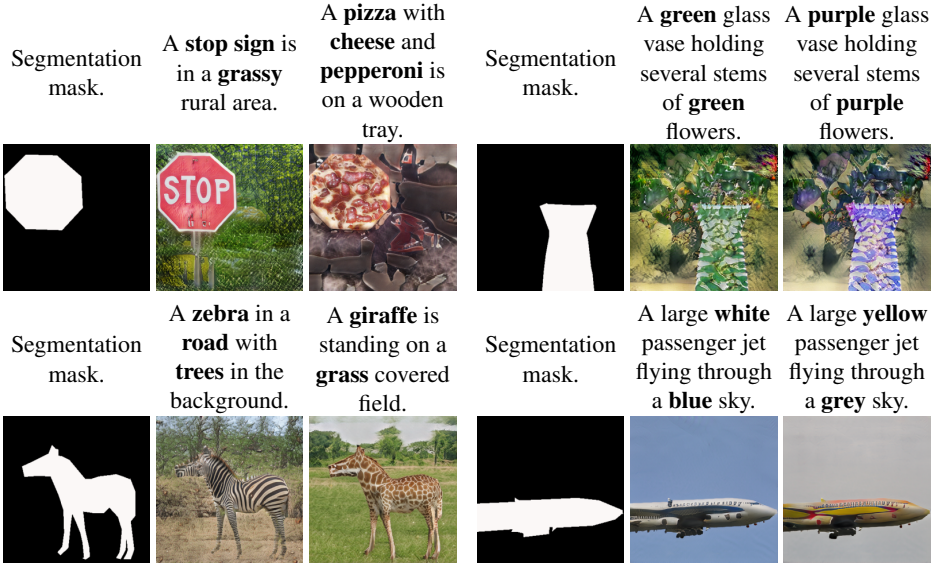

Figure 1: Given a segmentation mask and a text provided by a user that describes desired objects and visual attributes, the goal of this model is to generate realistic images semantically matching the given descriptions with the global structure defined by the masks.

In this paper, we propose to incorporate natural language descriptions into the image-to-image translation framework, where the description is used to allow users to freely determine the visual attributes of generated images. In particular, we focus on translating given masks into realistic images aided by natural language descriptions.

Firstly, we propose a novel discriminator with dual-directional feedback. According to feeding patches from higher-resolution fake/real images produced at higher stages to discriminators at lower stages, it encourages the discriminators to provide generators at the same stage with fine-grained supervisory feedback, related to image regions, encouraging generators to be aware of not only the global structure, but also the quality of image regions, which promotes the model to generate more realistic images with fine-grained regional details even at lower stages. This is actually in contrast to the widely adopted consensus in multi-stage training that lower stages are only responsible for coarse results. Besides, the improved discriminator can further provide feedback to higher-stage generators, with respect to the quality of image regions, encouraging them to complete missing details and to correct inappropriate visual attributes.

Secondly, we further introduce a new compatibility loss guided by a semantic mask. Given real/fake objects and background, we construct new fake images by combing real objects with the fake background or the real background with fake objects. The new combined fake images can enforce generators to produce realistic images with a better visual compatibility between objects and the background, i.e., both the generated objects and the background should be realistic and also fit the corresponding real background and objects without visual conflicts, respectively.

Finally, a new part-of-speech tagging-based (POS) attention is proposed, where POS filters out non-semantic words, and then allows the generator to capture detailed relations between image regions and corresponding semantic words to enable a translation with fine-grained regional details and accurate controllability. Extensive experiments demonstrate

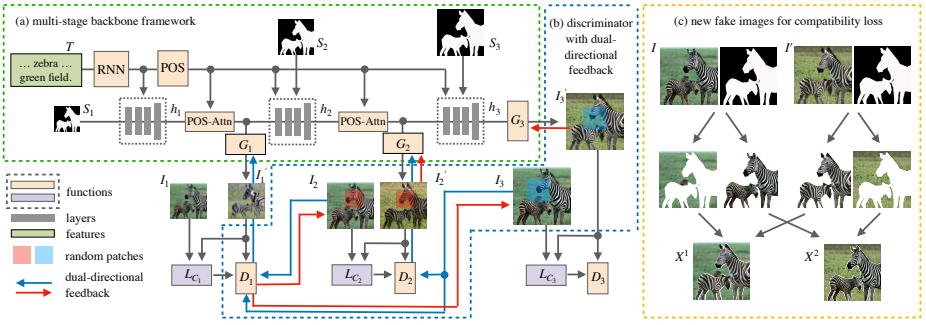

Figure 2: Architecture of our network. POS-Attn denotes the part-of-speech tagging-based attention. L_C denotes the compatibility loss.

that, given a semantic mask, our method can generate high-quality and diversified images, strictly controlled by given text descriptions, and even produce new images that are out-of-distribution of the given dataset, e.g., objects with an unusual color or a novel composition between objects and the background.

2 Related Work

Image-to-image translation is closely related to our work. Chen & Koltun [3] achieved a high-quality image generation using a single feedforward network. Wang et al. [33] proposed multi-scale generator and discriminator architectures in order to generate high-resolution images. Mo et al. [26] made use of object semantic masks to achieve instance transfiguration. Qi et al. [29] proposed a semi-parametric approach for image-to-image translation. Park et al. [28] implemented an affine transformation in conditional normalization techniques to avoid information loss. However, all these works and others [12, 32] only focus on generating realistic images from pixel-labeled semantic maps, and fail to have the ability to determine visual attributes of the synthetic images.

Text-guided image generation and manipulation have made great progress with the development of GANs [8, 14, 15, 17, 37], including image generation from text [1, 11, 13, 18, 19, 22, 23, 24, 30, 34, 35, 36], and image modification using text [5, 20, 21, 27]. Text-to-image generation aims to generate an image from a given text with text-image semantic alignment, and text-guided image manipulation is about editing given images using text descriptions to achieve semantic consistency.

3 Generative Adversarial Networks with Text Guidance

Given a semantic mask S and a text description T, we aim to translate the mask S into a realistic image I' with the global layout defined by the mask S. Meanwhile, the synthetic image I' should semantically match the description T, containing all required visual attributes. To achieve this, we propose three novel components: (1) a discriminator with dual-directional feedback, (2) a compatibility loss, and (3) a part-of-speech (POS) tagging-based attention.

3.1 Architecture

Our architecture is shown in Fig. 2. Given a text description T, we feed it into a pre-trained RNN (e.g., LSTM [34]) to generate text features. Then, we adopt an affine combination

module [20] at each stage to fuse text features (generated from the previous stage) with the segmentation mask, which can build an accurate correlation between words and the corresponding semantic parts of the mask, and thus embed text information into the generation process enabling an effective controllable ability. Next, the fused features are refined by a residual block followed by an upsampling block to produce hidden features, which are fed into a generator to output synthetic images I' and also serve as the input for the next stage to produce images at a higher resolution. Meanwhile, we use POS-based attention to capture correlation between image regions and corresponding semantic words. The whole framework generates high-quality images progressively, matching the global structure defined by the segmentation mask, and gradually produces regional visual attributes semantically aligned with the given description.

3.2 Discriminator with Dual-Directional Feedback

Generating realistic images involving different modality representations (e.g., natural language) on difficult datasets (e.g., COCO) is a big challenge for generative models [5, 27, 30], even with a multi-stage architecture [19, 34, 36], which generates a coarse image at the first stage, and then progressively increases its resolution with finer details. The ineffective generation is mainly because: (1) these models fail to produce a complete basic structure at lower stages, especially at the first one, which means that some parts of the synthetic image generated at the first stage are unrealistic, and (2) generators lack the ability to complete missing details or rectify inappropriate visual attributes. Thus, due to the flawed basic image and less efficient generators, the models fail to generate high-quality images with realistic details everywhere. This coincides with the observation shown in [39], where the quality of initial image features can greatly affect the quality of output images.

To address the above issues, we propose a novel discriminator with dual-directional feedback, which can fully explore the internal distribution of patches within a single image to strengthen the differential ability of discriminators and also the rectification ability of generators. As shown in Fig. 2, our network has a multi-stage architecture, and each stage has a generator and a discriminator, $\{G_1, D_1; G_2, D_2, \dots\}$. Different-scale images are generated progressively, $\{I'_1, I'_2, \dots\}$, and the resolution of the synthetic image is 4 times of the previous one. The generation of an image starts at the coarsest scale with the smallest resolution and sequentially passes through higher stages to the finer scale with larger resolution. To generate a complete structure at lower stages with finer details and thus to provide better basic image features for the following stages, we feed patches of real and fake images produced at higher stages to discriminators at lower ones, where the internal distribution of patches within images at higher stages contains unseen but finer pieces of information, which can be used as extra information to help to train and refine discriminators at lower stages to improve their differential ability, which in turn encourages generators at the same stages to produce a complete basic structure with fine-grained details (see Fig. 2, blue lines). The extra unconditional adversarial loss $\mathcal{L}_{Z_{Di}}$ for the discriminator at stage i is defined as:

$$\mathcal{L}_{Z_{Di}} = -\Big(\sum_{k=i+1}^{K} \big(E_{I_k \sim P_{\text{data}}}\big[\log(D_i(P_k))\big] + E_{I'_k \sim PG_k}\big[\log(1 - D_i(P'_k))\big]\big)\Big), \tag{1}$$

where K is the total number of stages, P'_k and P_k are random patches of the synthetic image I'_k and the real image I_k at a higher stage k, respectively. The size of patches P'_k and P_k matches the input requirement of the discriminator D_i.

Besides, we further feed the informative patches of fake images produced at higher stages to the improved discriminators at lower ones, in order to strengthen the rectification ability of generators, which can complete missing details and correct inappropriate visual attributes (see Fig. 2, red lines). Thus, the extra unconditional adversarial loss $\mathcal{L}_{Z_{Gi}}$ for the generator at stage i is defined as:

$$\mathcal{L}_{Z_{Gi}} = -(\sum_{k=1}^{i-1} E_{I'_i \sim PG_i}\left[\log(D_k(P'_k))\right]), \tag{2}$$

where $i > 1$, P'_k is a random patch of the i^{th} stage synthetic image I'_i, and the cropped size of P'_k matches the input requirement of the discriminator D_k.

Why does the refined multi-stage architecture work better? Patches from higher-stage real/fake images contain rich region-level fine-grained details that discriminators at lower stages are unfamiliar with. Taking these new informative patches as input, the low-stage discriminators can learn to distinguish details. In the generator training phase, discriminators can promote generators at the same stages to produce not only the coarse global structure, but also fine-grained regional details as much as possible. Moreover, the enhanced lower-stage discriminator can provide regional supervisory feedback to generators at higher stages, encouraging generators to complete missing contents and rectify inappropriate visual attributes produced from lower stages.

3.3 Compatibility Loss

To generate both realistic objects and the background with a better visual compatibility between them, we propose a novel compatibility loss. More specifically, we use the provided semantic mask to extract objects from the background on both fake and real images. Then, we create new compositions with different objects and background, and feed the fake composed images to discriminators. By doing this, the discriminator can be enhanced to have the ability to check the image quality of objects and the background, and also the visual compatibility between them, which in turn encourages generators to produce both realistic objects and the background with a better compatibility. The loss $\mathcal{L}_{C_i}$ at stage i is defined as:

$$\mathcal{L}_{C_i} = -E_{(X_i^1, X_i^2)}\left[\log(D_i(X_i^1))\right] + \left[\log(D_i(X_i^2))\right], \tag{3}$$

where X_i^1 represents the new image composed of fake objects with real background, and X_i^2 denotes real objects with fake background at stage i.

3.4 Part-of-Speech Tagging Based Spatial Attention

Given a text description, it may contain some less important words that cannot help image generation. For example, words "a, to, its" in a description do not have any semantic meaning, but if we keep these words, they may be connected with some visual attributes in the synthetic image, which may harm the ability of accurate control. Therefore, to ensure an accurate control of visual attributes, we propose a part-of-speech (POS) tagging-based attention, which first labels each word based on its definition and context, i.e., its relationship with adjacent and related words in the sentence [2], and produces attention weights to build correct relations between visual attributes and corresponding semantic words.

POS takes the text description as input and then labels each word with corresponding tags. In our model, we only keep words with specific tags: NN*, IN*, VB*, and JJ*. NN*

represents all nouns in different forms, IN* represents preposition or subordinating conjunction, VB* represents all verbs in any form, and JJ* represents all adjectives. We only keep these specific words, because nouns, prepositions, and verbs already capture the main meaning of a sentence, and adjectives contain the major descriptions of visual attributes of an image. Then, similarly to [34], the POS-based spatial attention weights are obtained by the following equations:

$$\beta_{i,j} = \frac{\exp(a_{i,j})}{\sum_{l=0}^{L-1} \exp(a_{i,l})}, \quad \text{where } a = v^T * w_{\text{pos}}, \tag{4}$$

where T is the transpose, $v \in \mathbb{R}^{D\times(H*W)}$ are intermediate visual hidden features, $w_{\text{pos}} \in \mathbb{R}^{D\times L}$ are filtered word embeddings containing desired semantic words, H is the height, W is the width, D is the feature dimension, and L represents the number of left semantic words in a sentence. So, $\beta_{i,j}$ denotes the correlation between the ith visual location and the jth word. Then, the weighted visual hidden features can be obtained by $v' = w_{\text{pos}} * \beta^T$, containing the information of the corresponding semantic words.

3.5 Objective Functions

To train the model, we add extra unconditional adversarial losses ($\mathcal{L}_{Z_{Di}}$, $\mathcal{L}_{Z_{Gi}}$) shown in Eqs. 1 and 2, and the compatibility loss ($\mathcal{L}_{C_i}$) shown in Eq. 3 to traditional conditional GANs' objectives at each stage. Generators and discriminators are optimized alternatively by minimizing their objective functions.

Discriminator objective. The loss function for the discriminator follows those used in the ControlGAN [19], but we introduce an extra unconditional adversarial loss (Eq. 1) at each stage to strengthen the differential ability of discriminators at lower stages, which, in turn, encourages the generators at the same stage to produce a complete structure with finer details.

Furthermore, we introduce a new compatibility loss (Eq. 3) in discriminators to better improve their differential ability, which can encourage generators to produce realistic details on both objects and the background. Thus, the complete loss function for the discriminator D_i at stage is i is defined as:

$$\begin{aligned}\mathcal{L}_{D_i} = & \underbrace{-\frac{1}{2}E_{I_i\sim P_{\text{data}}}\left[\log(D_i(I_i))\right] - \frac{1}{2}E_{I_i'\sim PG_i}\left[\log(1-D_i(I_i'))\right]}_{\text{unconditional adversarial loss}} \\ & \underbrace{-\frac{1}{2}E_{I_i\sim P_{\text{data}}}\left[\log(D_i(I_i,T))\right] - \frac{1}{2}E_{I_i'\sim PG_i}\left[\log(1-D_i(I_i',T))\right]}_{\text{conditional adversarial loss}} \\ & + \lambda_1((1-\mathcal{L}_{\text{corre}}(I_i,T)) + \mathcal{L}_{\text{corre}}(I_i,T')) + \lambda_2\mathcal{L}_{Z_{Di}} + \lambda_3\mathcal{L}_{C_i},\end{aligned} \tag{5}$$

where I is the real image sampled from the true image distribution P_{data}, T is the corresponding matched text that correctly describes I, I' is the generated image sampled from the model distribution PG, and T' is a mismatched text description randomly sampled from the dataset. The unconditional adversarial loss makes the synthetic image I' indistinguishable from the real image I, while the conditional adversarial loss aligns the generated image I' with the given text description T. $\mathcal{L}_{\text{corre}}$ [19] determines whether word-related visual attributes exist in the image. λ_1, λ_2, and λ_3 are hyperparameters controlling the importance of additional losses, and all are set to 1.

Generator objective. The loss function for the generator follows those used in the ControlGAN [19] with an extra unconditional adversarial loss (Eq. 2) at each stage to strengthen the rectification ability of generators, which can complete missing details and correct inappropriate visual attributes. Thus, the loss function of the generator G_i at stage i is defined as:

$$\mathcal{L}_{G_i} = \underbrace{-\frac{1}{2}E_{I'_i\sim PG_i}\left[\log(D_i(I'_i))\right]}_{\text{unconditional adversarial loss}} \underbrace{-\frac{1}{2}E_{I'_i\sim PG_i}\left[\log(D_i(I'_i,T))\right]}_{\text{conditional adversarial loss}} + \lambda_4\mathcal{L}_{\text{DAMSM}} + \lambda_5\mathcal{L}_{Z_{Gi}}, \tag{6}$$

where $\mathcal{L}_{\text{DAMSM}}$ [34] measures the text-image similarity at the word-level to provide fine-grained feedback for image generation. λ_4 and λ_5 are hyperparameters controlling the importance of the additional losses $\mathcal{L}_{\text{DAMSM}}$ and $\mathcal{L}_{Z_{Gi}}$, which are set to 5 and 1, respectively.

4 Experiments

We are unaware of any previous image-to-image translation work with text guidance, so we compare our method with the text-to-image generation methods AttnGAN [34] and ControlGAN [19], and image-to-image translation SPADE [28]. To have a fair comparison, we slightly modify AttnGAN and ControlGAN by implementing an affine combination module [20] to incorporate segmentation masks, denoted as AttnGAN-Seg and ControlGAN-Seg. As the input for SPADE is a pixel-labeled semantic map, for a fair comparison, we slightly modify the code released by authors, where we only keep the label for desired objects and set the rest to 0.

Dataset. COCO [25] contains 82,783 training images and 40,504 validation images. Each image has a ground-truth semantic mask and 5 descriptions. We only use binary segmentation masks instead of fine-grained pixel-labeled semantic maps. We preprocess the dataset according to the method in [35].

Implementation. Our model has three stages, and each stage has a generator and a discriminator. Three different-scale images (64×64, 128×128, and 256×256) are generated progressively. The model is trained for 120 epochs on the COCO dataset using the Adam optimizer [16] with the learning rate 0.0002. The hyperparameters controlling the importance of extra losses $\mathcal{L}_{Z_D}$, $\mathcal{L}_{Z_G}$, and $\mathcal{L}_C$ are set to 1. All experiments are conducted on a single Quadro RTX 6000 GPU.

4.1 Quantitative and Qualitative Comparison

Quantitative comparison. We adopt Fréchet Inception Distance (FID) [10] and Inception Score (IS) [34] to evaluate the quality and diversity of synthetic images. Also, to measure the semantic consistency between the generated images and the corresponding text descriptions, we adopt the R-precision (R-prcn) [34], which is an evaluation metric for ranking retrieval results. As shown in Table 1, our model achieves better FID and IS scores, which demonstrates that our model can generate realistic images with high diversity. Also, the better R-prcn value indicates that the synthetic images generated by our model are highly semantically matching the given text descriptions.

Qualitative comparison. Fig. 3 shows a visual comparison between our method and SPADE on COCO. For SPADE, we randomly generate several synthetic images for the same semantic mask input by sampling different random vectors. For our method, given the semantic

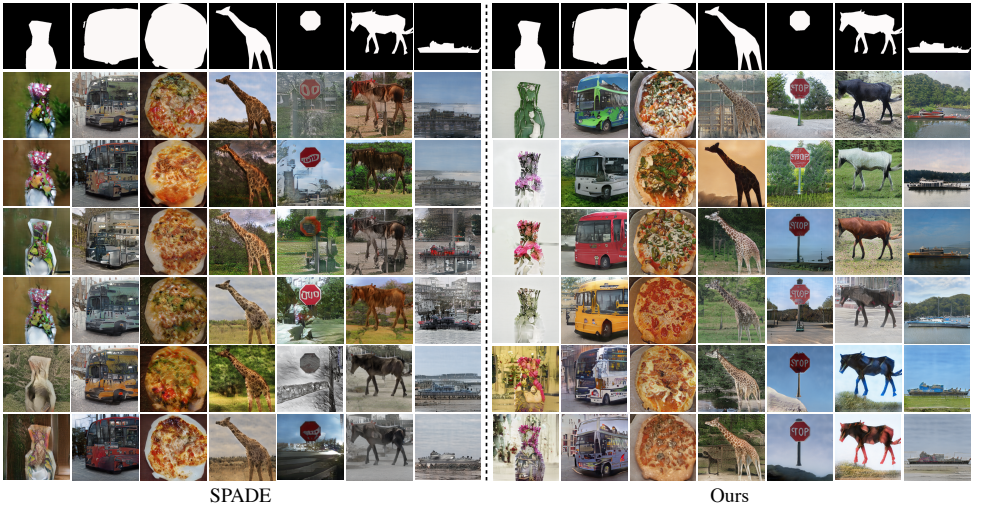

Figure 3: Qualitative comparison of SPADE and ours on the COCO dataset. For simplicity, we omit corresponding text descriptions for our approach.

Method	FID	IS	R-prcn (%)
SPADE	42.74	11.69 ± 0.26	-
AttnGAN-Seg	32.39	12.09 ± 0.28	75.24 ± 3.39
ControlGAN-Seg	31.41	11.56 ± 0.16	80.43 ± 2.79
Ours	**28.30**	$\mathbf{15.96 \pm 0.16}$	$\mathbf{83.23 \pm 1.37}$

Table 1: Quantitative comparison: FID, IS, and R-prcn of ours and baselines on the COCO dataset. For FID, lower is better, for IS and R-prcn, higher is better.

mask, we randomly sample text descriptions and then use our model to produce various synthetic images under the control of these descriptions. We can easily observe that our method attains a much better image quality and diversity, and also flexibly controls the visual attributes of the generated images as well.

Besides, as shown in Fig. 3, last two rows, our method can generate novel images with a good visual compatibility that are out-of-distribution of the given dataset, e.g., the stop sign floating in the sky, red/blue horses, and different color boats on the green grass or the dirt. Such unusual colors or compositions between objects and the background demonstrate that our method effectively disentangles different visual attributes, and accurately builds connections with corresponding semantic words.

In Fig. 5, right, we further verify the disentanglement ability of our method. As we can see, if there is no segmentation mask being provided to the network, only the background is generated by our model, but the result still semantically matches the given text description. Besides, the generation of objects has almost no impact on the generation of the background, even when we provide different segmentation masks, which illustrates an effective disentanglement between fore- and background. Based on this, our model is able to generate diverse synthetic results by adding objects without changing the background, and also enables us to modify visual attributes of synthetic images, while preserving content that is not required in the given text description.

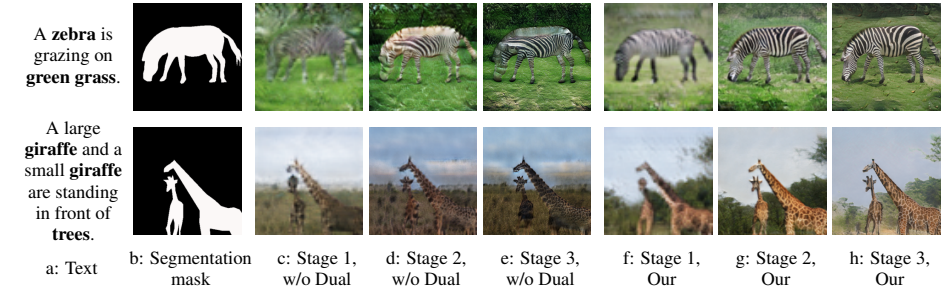

Figure 4: Effectiveness of dual-directional feedback. *c*, *d*, and *e* show the synthetic images produced at each stage by the model without adopting the discriminator with dual-directional feedback. *f*, *g*, and *h* show the synthetic images generated at each stage by our full model.

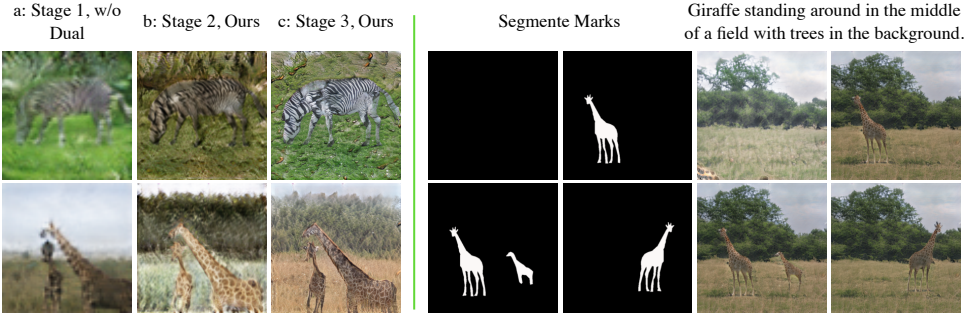

Figure 5: Left: rectification ability of our generators. *a* denotes images generated at the first stage by the model without dual-directional feedback. In *b* and *c*, denote our model takes these flawed features and feeds them through stages 2 and 3 progressively, producing the corresponding images shown at *b* and *c*. Right: disentanglement of objects and background.

4.2 Ablation Studies

To evaluate the effectiveness of each proposed component adopted in our network, we conducted the ablation studies shown in Table 2.

Discriminator with dual-directional feedback. When the model adopts our proposed discriminator with dual-directional feedback, the scores on all evaluation metrics improve significantly. We attribute this improvement to the generation of high-quality visual features at lower stages, which contain both global information and fine-grained regional details, and thus further improve the final synthetic results. This is also supported by the observation found in text-to-image generation [39], where the quality of initial image features can considerably affect the quality of output images in such a sequential upsampling generation pipeline (see Fig. 2). In Fig. 4, we further visualize this improvement. Without dual-directional feedback, the model fails to produce completed images with appropriate regional details at the lower stages, especially the first stage, e.g., "the zebra misses the back and head" at the top of columns *c* and *d*, and "there is no tree background", and "the smaller giraffe misses legs" at the bottom of columns *c* and *d*.

Besides, we think that the improvement is also because the discriminator with dual-directional feedback enables generators to have an effective rectification ability, where the

Method	FID	IS	R-prcn (%)
Ours w/o Dual	32.14	12.16 ± 0.20	80.13 ± 2.20
Ours w/o Compatibility	29.47	14.72 ± 0.32	81.43 ± 1.21
Ours w/o POS-Attn	32.72	12.77 ± 0.21	81.07 ± 1.60
Ours w/ WSA [34]	30.14	14.49 ± 0.15	82.03 ± 1.03
Ours	**28.30**	$\mathbf{15.96 \pm 0.16}$	$\mathbf{83.23 \pm 1.37}$

Table 2: Ablation studies. "w/o Dual" denotes without using the proposed discriminator with dual-directional feedback; "w/o Compatibility" denotes without compatibility loss; "w/o POS-Attn" denotes without part-of-speech tagging-based attention; "w/ WSA" denotes using the attention in [34] to replace our proposed POS-based attention.

generator at the following stages can complete the missing content or rectify inappropriate visual attributes. For example, as shown in Fig. 4, without using the proposed discriminator, the model leaves the missing areas without any correction (see columns *d* and *e*). To further verify the rectification ability, we feed the flawed features generated at the first stage by the model without adopting the discriminator with dual-directional feedback to our full model, shown in Fig. 5, left. Even if there are missing parts in the given images, the generators in our full model are able to complete the missing attributes, e.g., "adding back and head for the zebra" at the top row, and to correct inappropriate visual attributes, e.g., "change the background with trees" at the bottom row.

Compatibility loss. To verify the effectiveness of the compatibility loss, we removed it from our model and then checked all evaluation metrics, shown in Table 2. Without it, the scores on all metrics degrade. We think that, without the compatibility loss, the model may fail to produce realistic objects or the background in the synthetic images, and the compatibility between them may be far from satisfactory.

Part-of-speech tagging-based spatial attention. As discussed in Section 3.4, the implementation of part-of-speech (POS) tagging can help to filter out specific words, especially less important ones, which can effectively prevent less useful information being contained in word and sentence features, and also avoid building inappropriate connections between non-semantic words and visual attributes, such that the model can produce high-quality images with finer regional details, and also achieve a better controllable performance.

To evaluate its effectiveness, we first remove it from the model, and then replace it with word-level spatial attention [34], shown in Table 2. As we can observe, the worse performance on "w/o POS-Attn" shows its effectiveness on high-quality image generation. Compared to "w/ WSA", our full model achieves better scores on all metrics, which demonstrates that unnecessary connections can be built between non-semantic words and visual attributes, and these useless bonding can harm the quality of the synthetic results.

5 Conclusion

We have proposed a novel generative adversarial network, which effectively embeds controllable factors, i.e., text descriptions, into image-to-image translation to control the generation of objects and visual attributes. Also, our model disentangles objects from the background, produces high-quality image features at lower stages, and has a rectification ability. Extensive experiments demonstrate the advantages of our method, with respective to both high-quality image generation and the effectiveness of control of local visual attributes.

Acknowledgments

This work was supported by the UKRI Turing AI Fellowship EP/W002981/1 and the EPSRC/MURI grant EP/N019474/1. We also thank the Royal Academy of Engineering and FiveAI. This work was also supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1, by the AXA Research Fund, and by the EPSRC grant EP/R013667/1. We also acknowledge the use of the EPSRC-funded Tier 2 facility JADE (EP/P020275/1) and GPU computing support by Scan Computers International Ltd.

References

[1] Oron Ashual and Lior Wolf. Specifying object attributes and relations in interactive scene generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4561–4569, 2019.

[2] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: Analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[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] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. Sketch2photo: Internet image montage. *ACM Transactions on Graphics (TOG)*, 28(5):1–10, 2009.

[5] 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.

[6] Mathias Eitz, Ronald Richter, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. Photosketcher: Interactive sketch-based image synthesis. *IEEE Computer Graphics and Applications*, 31(6):56–66, 2011.

[7] Arnab Ghosh, Richard Zhang, Puneet K Dokania, Oliver Wang, Alexei A Efros, Philip HS Torr, and Eli Shechtman. Interactive sketch & fill: Multiclass sketch-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1171–1180, 2019.

[8] 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.

[9] David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[11] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7986–7994, 2018.

[12] 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.

[13] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1219–1228, 2018.

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Bowen Li and Thomas Lukasiewicz. Lightweight long-range generative adversarial networks. *arXiv preprint arXiv:2209.03793*, 2022.

[18] Bowen Li and Thomas Lukasiewicz. Word-level fine-grained story visualization. *arXiv preprint arXiv:2208.02341*, 2022.

[19] 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.

[20] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. ManiGAN: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020.

[21] Bowen Li, Xiaojuan Qi, Philip Torr, and Thomas Lukasiewicz. Lightweight generative adversarial networks for text-guided image manipulation. *Advances in Neural Information Processing Systems*, 33:22020–22031, 2020.

[22] Bowen Li, Philip HS Torr, and Thomas Lukasiewicz. Clustering generative adversarial networks for story visualization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 769–778, 2022.

[23] Bowen Li, Philip HS Torr, and Thomas Lukasiewicz. Memory-driven text-to-image generation. *arXiv preprint arXiv:2208.07022*, 2022.

[24] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12174–12182, 2019.

[25] 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.

[26] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Instagan: Instance-aware image-to-image translation. *arXiv preprint arXiv:1812.10889*, 2018.

[27] 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.

[28] 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.

[29] Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. Semi-parametric image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8808–8816, 2018.

[30] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[31] Wei Sun and Tianfu Wu. Learning layout and style reconfigurable gans for controllable image synthesis. *arXiv preprint arXiv:2003.11571*, 2020.

[32] Hao Tang, Dan Xu, Yan Yan, Philip HS 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363. PMLR, 2019.

[38] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8584–8593, 2019.

[39] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019.