Image-to-Image Translation with Text Guidance

Bowen Li, Xiaojuan Qi, Philip H. S. Torr, Thomas Lukasiewicz
University of Oxford
{bowen.li, thomas.lukasiewicz}@cs.ox.ac.uk
{xiaojuan.qi, philip.torr}@eng.ox.ac.uk

Abstract

The goal of this paper is to embed controllable factors, i.e., natural language descriptions, into image-to-image translation with generative adversarial networks, which allows text descriptions to determine the visual attributes of synthetic images. We propose four key components: (1) the implementation of part-of-speech tagging to filter out non-semantic words in the given description, (2) the adoption of an affine combination module to effectively fuse different modality text and image features, (3) a novel refined multi-stage architecture to strengthen the differential ability of discriminators and the rectification ability of generators, and (4) a new structure loss to further improve discriminators to better distinguish real and synthetic images. Extensive experiments on the COCO dataset demonstrate that our method has a superior performance on both visual realism and semantic consistency with given descriptions.

1. Introduction

Conditional image synthesis aims to generate realistic images semantically matching given conditions, including text-to-image generation (Reed et al., 2016; Zhang et al., 2017; 2018; Xu et al., 2018; Li et al., 2019a) and image generation from scene graphs (Johnson et al., 2018; Ashual & Wolf, 2019), semantic layout (Isola et al., 2017; Chen & Koltun, 2017; Wang et al., 2018; Mo et al., 2018; Park et al., 2019), or coarse layout (Zhao et al., 2019), which enables numerous potential applications in many areas, including design, video games, art, architecture, and image editing.

The goal of this paper is to produce realistic images from segmentation masks, and also to embed controllable factors, i.e., natural language descriptions, into the generation process to control the visual attributes (e.g., colour, background, and texture) of synthetic images semantically matching the given texts. Unlike current state-of-the-art image-to-image translation models (Isola et al., 2017; Chen & Koltun, 2017; Wang et al., 2018; Park et al., 2019; Pavllo et al., 2019; Tang et al., 2019), which require fine-grained pixel-labelled semantic maps to decide what to generate and usually fail to predict exact visual attributes of synthetic images, our model is able to generate desired images under the control of natural language descriptions, even if the provided masks are simple. As shown in Fig. 1, given a simple circle segmentation mask, our model is able to generate a *stop sign at grassy area* and also a *pizza with cheese and pepperoni*.

To achieve this, the key is to completely disentangle different visual attributes contained in text descriptions and images, and then to build an accurate correlation between semantic words and corresponding visual attributes to achieve effective control. Also, how to effectively generate realistic images involving different modality representations (e.g., natural language) on more difficult datasets (e.g., COCO (Lin et al., 2014)) is a critical issue to address, where each image in the dataset has multiple objects with complicated relationships between each other.

To address the above issues, we propose a novel generative adversarial network, called RefinedGAN, which can effectively generate realistic images from segmentation masks, and also embed controllable factors (i.e., text descriptions) in the generation process, which allows users to determine the exact visual attributes of synthetic images.

We propose four key components in our RefinedGAN: (1) part-of-speech (POS) tagging is implemented to filter out less important (non-semantic) words in the given text description, (2) the affine combination module (Li et al., 2019b) is adopted to effectively fuse different modality representations (text and segmentation mask), and to also build an effective connection between them, (3) to generate high-quality images from segmentation masks involving text, we propose a novel refined multi-stage architecture for discriminators and generators, which strengthens the differential ability of discriminators and in turn encourages generators at lower stages to produce not only the global structure and layout, but also fine-grained details as much as possible. Also, this architecture improves the rectification ability of generators to complete missing details and correct inappropriate attributes produced from lower stages, and (4) a new

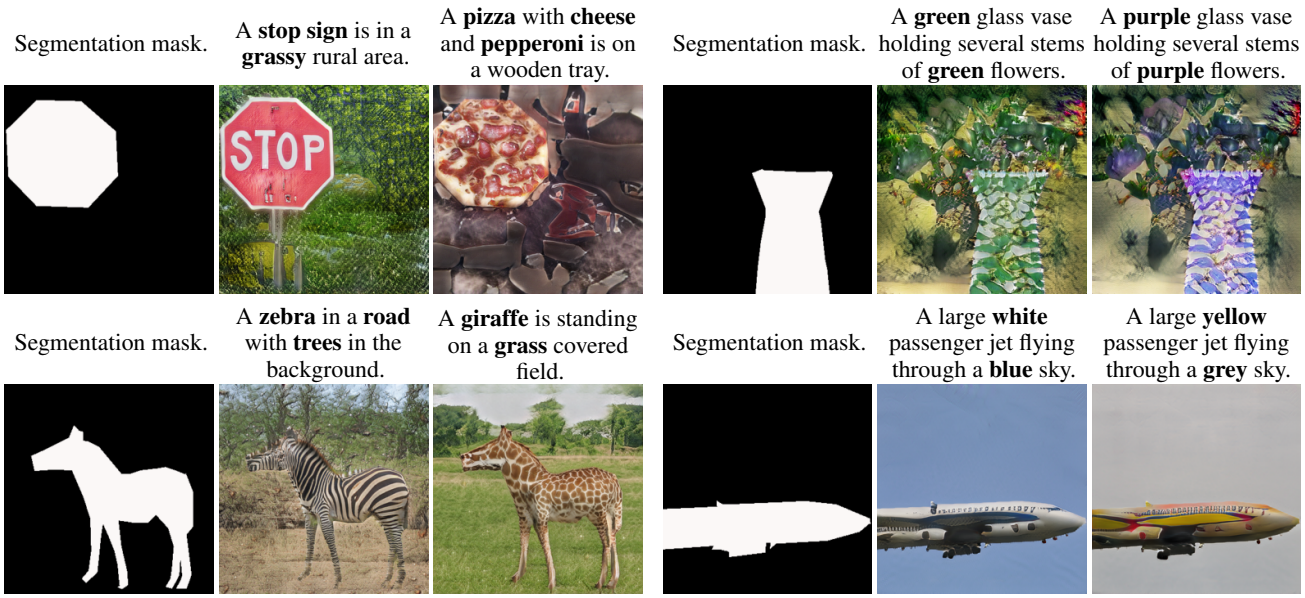

Figure 1. Given a segmentation mask and a text provided by a user that describes desired objects and visual attributes, the goal of this model is to generate realistic images semantically matching the given descriptions with the global structure defined by the masks.

structure loss is proposed to further improve discriminators in order to better distinguish fake images from real ones.

Finally, an extensive analysis is performed, which demonstrates that our method can effectively generate realistic images on the more complex COCO dataset (Lin et al., 2014), and also accurately control the visual attributes of synthetic images using natural language descriptions. Experimental results on the dataset show that our method outperforms existing methods both qualitatively and quantitatively.

2. Related Work

Image-to-image translation is closely related to our work. Chen & Koltun (2017) achieved high-quality image generation using a single feedforward network. Wang et al. (2018) proposed multi-scale generator and discriminator architectures in order to generate high-resolution images. Mo et al. (2018) made use of object segmentation masks to achieve instance transfiguration. Park et al. (2019) implemented affine transformation in conditional normalisation techniques to avoid information loss. However, all these works and others (Isola et al., 2017; Qi et al., 2018; Tang et al., 2019) only focus on generating realistic images from pixel-labelled semantic maps without the ability to determine the visual attributes of synthetic images.

Text-to-image generation has made great progress with the development of GANs (Goodfellow et al., 2014), including image generation from text (Reed et al., 2016; Zhang et al., 2017; 2018; Xu et al., 2018; Li et al., 2019a) and scene graphs (Johnson et al., 2018; Ashual & Wolf, 2019). Also, Hong et al. (2018) and Li et al. (2019c) proposed to generate intermediate layout first (i.e., bounding boxes and segmentation masks) and then convert the semantic layout into an image. However, all the above methods mainly focus on generating realistic images from text descriptions or scene graphs, and usually fail to produce high-quality results on more complicated datasets, such as COCO.

Text-guided image manipulation is about editing given images using texts to achieve semantic consistency. Dong et al. (2017) built an encoder-decoder architecture to get an appropriate modification. Nam et al. (2018) proposed a text-adaptive discriminator to utilise word-level information. Instead of using the simple and coarse concatenation method, Li et al. (2019b) proposed a novel affine combination module to effectively fuse different modality representations. However, these methods mainly aim to edit a given image rather than producing new results.

Multi-stage architectures have been widely adopted in GAN-based models (Denton et al., 2015; Huang et al., 2017; Zhang et al., 2017; 2018; Xu et al., 2018; Shaham et al., 2019; Li et al., 2019a) to produce high-resolution images, which have a generator and a discriminator at each level of an image pyramid, and generate image progressively from coarse-to-fine scale. Differently from them, our model fully makes use of features produced at higher stages by feeding these features to discriminators at lower ones to improve their differential ability, which further benefit generators to improve their rectification ability as well. Our architecture is more suitable for realistic image generation with finer details and under control of natural language descriptions.

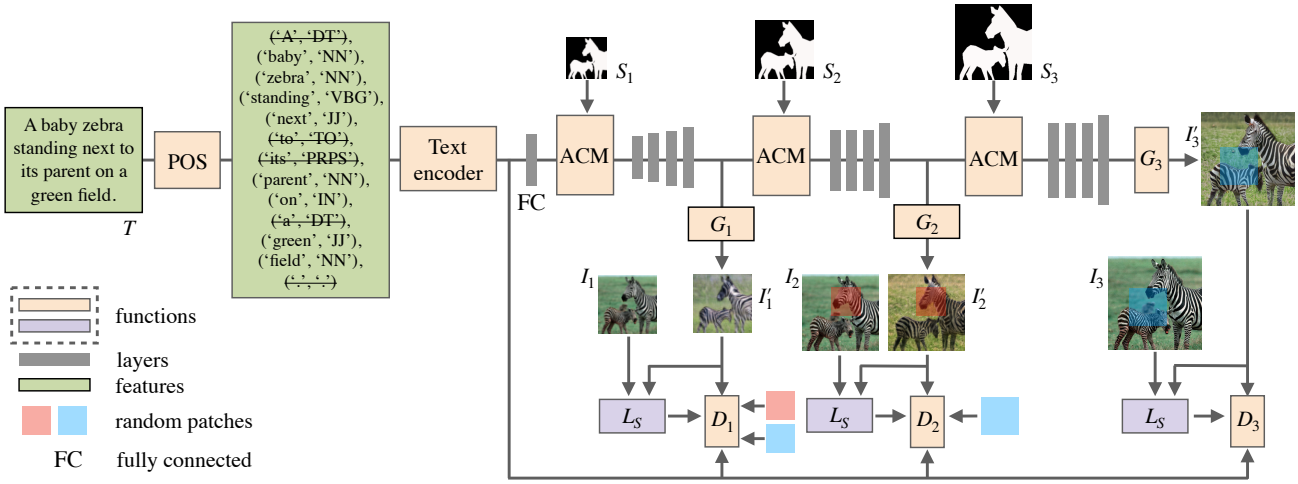

Figure 2. The architecture of RefinedGAN. POS denotes the part-of-speech tagging. ACM denotes the affine combination module. L_S denotes the structure loss defined in Sec. 3.5. The attention is omitted for simplicity. Please see supplementary material for the full architecture. Note that in our paper, we just provide the original text descriptions instead of filtered words in each example for simplicity.

3. Refined Generative Adversarial Networks

Suppose as given a segmentation mask S and a text description T provided by a user. The model aims to generate a realistic image I' semantically matching the given text T with the global structure defined by the segmentation mask S. To achieve this, we propose four components: (1) the implementation of part-of-speech (POS) tagging, (2) the adoption of an affine combination module (Li et al., 2019b), (3) a refined multi-stage architecture for both discriminators and generators, and (4) a novel structure loss.

3.1. Architecture

We adopt the muti-stage ControlGAN (Li et al., 2019a) as our basic framework shown in Fig. 2. We implement part-of-speech (POS) tagging to filter out non-semantic words in the given text description, and then feed the output into a pre-trained RNN (Xu et al., 2018) to generate text representations. Then, we adopt an affine combination module (ACM) 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 and also serve as the input for the next stage to produce images at a higher resolution. 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. Part-of-Speech Tagging

Given a text description, it may contain some less important words that cannot help image generation and even cause negative impact. For example, as shown in Fig. 2, words "a, to, its" in the 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, in order to filter out these words, we implement part-of-speech (POS) tagging to label each word such that each word in the description is marked up corresponding to a particular part of speech, based on both its definition and context, i.e., its relationship with adjacent and related words in the sentence (Bird et al., 2009).

As shown in Fig. 2, 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:

$$\text{NN}^*, \text{IN}^*, \text{VB}^*, \text{and JJ}^*, \tag{1}$$

where asterisk $*$ indicates zero or more occurrences of any element, NN^* represents all nouns in different forms, IN^* represents preposition or subordinating conjunction, VB^* represents all verbs in any form, and JJ^* represents all adjective. 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.

Why does filter out less important words work better than keeping all words? First, not all words are equally important in a given text description, and some words may have no or even negative impact on the generation process, e.g., determiner (a, an, the), or adverb. Also, keeping these

words, the generated word and sentence features from a RNN-based text encoder must contain these less useful information, and then an inappropriate correlation is built between non-semantic words and visual attributes. Thus, the performance of control may be affected, especially when users only want to modify some visual attributes of a synthetic image while preserving the other content. More details are discussed in Sec. 4.2.

3.3. Affine Combination Module

To effectively fuse different modality features (i.e., text and image) together, we adopt the affine combination module (ACM) (Li et al., 2019b) instead of simply concatenating both features along the channel direction. In our model, the ACM is placed before each residual block followed by an upsampling block at the end of each stage, shown in Fig. 2, and is defined as:

$$h' = h \odot W(S) + b(S), \tag{2}$$

where W and b represent the functions that convert the segmentation mask S to learned weights $W(S)$ and biases $b(S)$, h denotes the hidden features encoded from the input text description, or it is the intermediate hidden representation between two stages, h' denotes the fused features containing pieces of information of both language and segmentation mask, and $\odot$ denotes the Hadamard element-wise product. Please see the supplementary material for more details on the architecture.

3.4. Refined Multi-Stage Architecture

Generating realistic images involving different modality representations (e.g., natural language) on more difficult datasets (e.g., COCO) is a big challenge for generative models (Reed et al., 2016; Dong et al., 2017; Nam et al., 2018), even with a multi-stage architecture (Zhang et al., 2018; Xu et al., 2018; Li et al., 2019a), 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 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.

In order to address the above issues, we propose a novel refined multi-stage architecture for both discriminators and generators, which can fully explore the internal distribution of patches within a single image to strengthen the differential ability of discriminators at lower stages and the rectification ability of generators.

As shown in Fig. 2, our model has a multi-stage architecture and each stage has a generator and a discriminator, $\{G_1, D_1; G_2, D_2, ...\}$. Different-scale images are generated progressively, $\{I'_1, I'_2, ...\}$, 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 a better basic image for the following stages, we feed patches of real and fake images 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, especially for the generator at the first stage, see Fig. 7. Thus, the extra unconditional adversarial loss $\mathcal{L}_{Z_{Di}}$ for the discriminator at stage i is defined as:

$$\begin{aligned}\mathcal{L}_{Z_{Di}} = -(\sum_{k=i+1}^{K} (&E_{I_k \sim P_{\text{data}}} [\log(D_i(P_k))] + \\ &E_{I'_k \sim PG_k} [\log(1 - D_i(P'_k))])),\end{aligned} \tag{3}$$

where K is the total number of stages, P'_k and P_k are random patches (detached) 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 these refined discriminators at lower ones in order to strengthen the rectification ability of their following generators, which can complete missing details and correct inappropriate visual attributes, shown in Fig. 8. Thus, the extra unconditional adversarial loss $\mathcal{L}_{Z_{Gi}}$ for the generator at stage i is defined as:

$$\mathcal{L}_{Z_{Gi}} = -(\sum_{k=1}^{i-1} E_{I'_i \sim PG_i} [\log(D_k(P'_k))]), \tag{4}$$

where $i > 1$, P'_k is a random patch (non-detached) 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? This architecture aims to refine discriminators and generators by feeding patches from higher stages to lower ones, where these patches contain an unseen internal distribution with fine-grained details. By doing this, discriminators at lower stages can better identify fake images based on not only the global distribution, but also regional features, which

in turn encourages generators at the same stages to produce realistic images with finer details. Also, these enhanced discriminators can provide regional feedback to generators at their following stages, refining the generators to produce realistic regions and have the abilities of completing missing contents and rectifying inappropriate visual attributes.

3.5. Structure Loss

To further improve the differential ability of discriminators, we propose a novel structure loss, which can also be used to stabilise the training, since generators have to produce natural statistics for both objects and background. More specifically, we use the provided segmentation mask to separate objects and background on both synthetic and real images. Then, we create new compositions with different objects and background, i.e., fake objects + real background and real objects + fake background, and feed these new images to discriminators to improve their differential ability, identifying a fake image if there exist some unrealistic regions only in the foreground or background. In turn, generators can be encouraged to produce finer details everywhere without preference, instead of focusing only on the generation of realistic objects or the background. Thus, the structure loss $\mathcal{L}_{S_i}$ at stage i is defined as:

$$\mathcal{L}_{S_i} = -E_{(X_i^1, X_i^2)} \left[\log(D_i(X_i^1))\right] + \left[\log(D_i(X_i^2))\right], \tag{5}$$

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.

3.6. Objective Functions

To train the model, we follow the ControlGAN (Li et al., 2019a) and add extra unconditional adversarial losses ($\mathcal{L}_{Z_{Di}}$, $\mathcal{L}_{Z_{Gi}}$) in Eqs.3 and 4 and the structure loss ($\mathcal{L}_{S_i}$) in Eq. 5 at each stage. Generators and discriminators are optimised alternatively by minimising their objective functions. Please see the supplementary material for complete objectives. We only the highlight differences compared to the ControlGAN.

4. Experiments

The model is evaluated on the COCO dataset (Lin et al., 2014), generating different-scale images progressively. We are unaware of any previous end-to-end methods for generating images from segmentation masks with embedded controllable factors, e.g., natural language descriptions, so we compare our method with AttnGAN (Xu et al., 2018) and ControlGAN (Li et al., 2019a), state-of-the-art methods for generating images from texts. To have a fair comparison, we slightly modify both models by implementing ACM (Li et al., 2019b) to incorporate segmentation masks, and call the modified models S-AttGAN and S-ControlGAN, respectively. Note that we do not choose the models introduced in (Johnson et al., 2018; Ashual & Wolf, 2019) as baselines, because input scene graphs do not contain descriptions of visual attributes. Also, we do not compare our work with studies (Hong et al., 2018; Li et al., 2019c), because models proposed in both studies require bounding boxes as intermediate input to produce synthetic images.

Dataset. The COCO (Lin et al., 2014) contains $82,783$ training images and $40,504$ validation images. Each image has a ground truth semantic mask and 5 descriptions. In our task, we only use binary segmentation masks instead of fine-grained pixel-labelled semantic maps. We preprocess the dataset according to the method in (Zhang et al., 2017).

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 optimiser (Kingma & Ba, 2014) with the learning rate 0.0002. The hyperparameters controlling the extra losses $\mathcal{L}_{Z_D}$, $\mathcal{L}_{Z_G}$ and $\mathcal{L}_S$ are set to 1.

4.1. Comparison with State-of-the-Art Approaches

Table 1. Quantitative comparison: Inception Score (IS) and R-precision (R-prcn) of state-of-the-art methods and RefinedGAN on the COCO dataset. "w/o POS" denotes without part-of-speech tagging; "w/ Concate." denotes using a concatenation method to combine text and image features instead of using the affine combination module; "w/o Refined" denotes without using refined multi-stage architectures on discriminators and generators; "w/o SL" denotes without structure loss; "w/o POS*" denotes the model is trained without using POS, but the POS is implemented on the testing. For IS and R-prcn, higher is better.

Method	IS	R-prcn (%)
Real Images	27.41 ± 0.59	-
S-AttnGAN	12.09 ± 0.28	75.24 ± 3.39
S-ControlGAN	11.56 ± 0.16	80.43 ± 2.79
Ours w/o POS	$\mathbf{16.49 \pm 0.18}$	$\mathbf{84.01 \pm 1.59}$
Ours w/ Concat.	8.50 ± 0.15	44.11 ± 3.99
Ours w/o Refined	12.16 ± 0.20	80.13 ± 2.20
Ours w/o SL	14.72 ± 0.32	81.43 ± 1.21
Ours w/o POS*	14.74 ± 0.13	83.03 ± 1.15
Ours	15.96 ± 0.16	83.23 ± 1.37

Quantitative comparison. We adopt the Inception Score (IS) (Xu et al., 2018) 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) (Xu et al., 2018), which is an evaluation metric for ranking retrieval results. IS and R-prcn are evaluated on a large number of matched text-mask pairs sampled from the dataset.

As shown in Table. 1, our model achieves a better IS value,

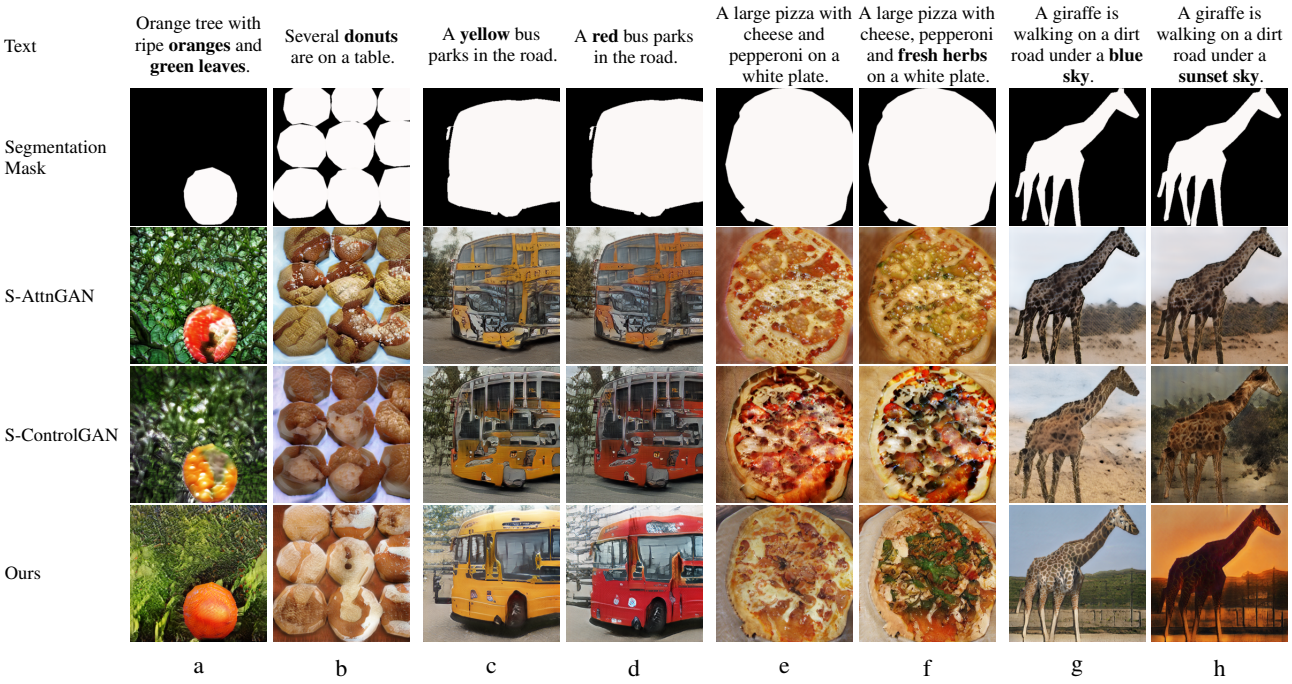

Figure 3. Qualitative comparison of three methods on the COCO dataset. (1) *a* and *b* represent the generation of objects belonging to different categories on similar segmentation masks; (2) *c* and *d* illustrate the controllable ability of internal visual attributes of objects; (3) *e* and *f* show the capability of adding new visual attributes on synthetic images while preserving other text-unmodified contents; and (4) *g* and *h* show that the model can also control the global style of the generated results.

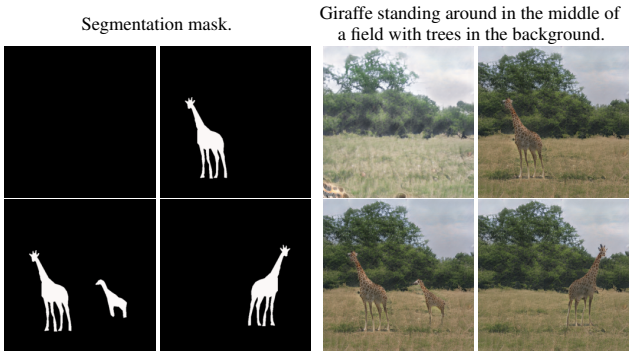

Figure 4. Disentanglement of objects and background.

which demonstrates that our model can generate more realistic images with high diversity. Also, the better R-prcn value indicates that the synthetic images generated by our model highly match the given text descriptions. Note that compared to "Ours w/o POS", the IS and R-prcn values of "Ours w/o POS*" decrease, which illustrates that unnecessary connections are built between non-semantic words and visual attributes, and these useless bonding can harm the quality of synthetic results, examples shown in Fig. 6.

Qualitative comparison. Fig. 3 shows the visual comparison between the RefinedGAN, AttnGAN-Seg, and ControlGAN-Seg on the COCO dataset (Lin et al., 2014). We can easily observe that both methods are only able to generate low-quality images with unrealistic objects, e.g., the buses produced by both methods have distorted textures (columns *c* and *d*), and oranges have unusual black or brown colour (column *a*). Also, the synthetic results generated from both methods do not have a perfect semantic consistency with given text descriptions, i.e., both methods fail to produce the new global style *sunset* at the column *h*, and the new attribute *fresh herbs* at the column *f*.

Failing to generate realistic images by both methods is mainly because: (1) there exist unnecessary connections between non-semantic words and visual attributes, which can potentially constrain the control of attributes (e.g., change of colour to *red* at *d*); and (2) both methods fail to produce a complete structure at lower stages and also do not have an effective rectification ability, e.g., no *white plate* at the column *e*, no *blue sky* at column *g*. However, our model addresses above problems by implementing POS tagging, the refined multi-stage architecture and structure loss. More details are discussed in Sec.4.2.

Besides, our model can effectively disentangle the foreground objects with background, shown in Fig. 4. As we can see that if there is no segmentation mask being provided, only background is generated by our model, but the result still semantically matches the given description. Also, the generation of objects has almost no impact on the generation

Figure 5. Ablation studies. *a*: given text description with the desired objects and visual attributes; *b*: segmentation mask; *c*: using the concatenation method to replace the affine combination module; *d*: without implementing the refined multi-stage architecture; *e*: without structure loss; *f*: our full model.

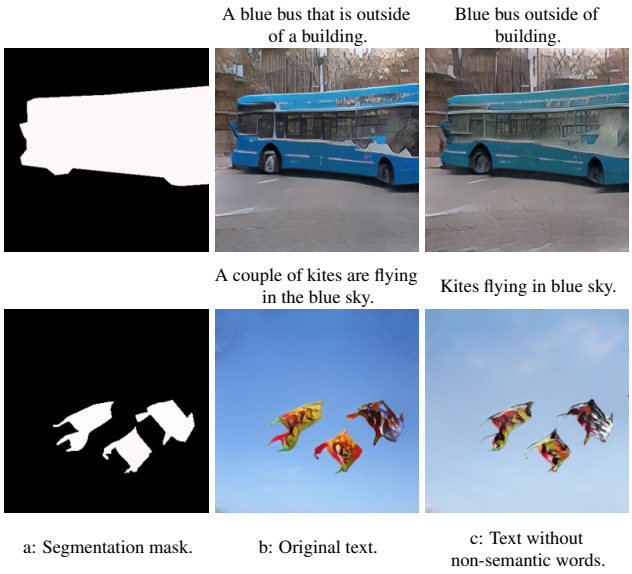

Figure 6. Existence of unnecessary connections between non-semantic words and visual attributes. *b* and *c* show images that are generated from full text descriptions and filtered descriptions by the model without POS, respectively.

of background, even when we provide different segmentation masks, which illustrates an effective disentanglement between foreground and background. Based on this, our model can generate diverse results by adding objects without changing the background, and also enable us to modify visual attributes of synthetic images, while preserving content that is not required in the modified text descriptions. For example, shown in the columns *e* and *f* in Fig. 3, the backgrounds *white plate* are almost the same when a new attribute *fresh herbs* is added.

4.2. Ablation Studies

Necessity of part-of-speech tagging. As discussed in Sec. 3.2, 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 achieve a better controllable performance.

Thus, we conduct a study to verify the existence of those useless or even harmful connections via the model without POS, shown in Fig. 6. As we can see, when we remove non-semantic words from the description, some regions of the synthetic image become unrealistic, i.e., there is an obvious white patch shown on the bus. Moreover, those useless words can even decrease the quality of synthetic results, shown in the bottom of Fig. 6, which reduces the brightness of the blue sky and affects the texture of colourful kites.

Effectiveness of affine combination module. As shown in Fig. 5 *c* and *f*, we conduct an ablation study to show the effectiveness of the affine combination module (ACM). As we can see, the adoption of concatenation instead of ACM to combine different modality features, the model cannot produce realistic images with finer details (see top of *c*), and even fails to keep a semantic consistency with the given description (see bottom of *c*), while our model is able to produce high-quality results with fine-grained visual attributes under the control of the texts. This is mainly because the simple concatenation cannot build an accurate connection between semantic words with corresponding regions of the image, and also fails to effectively encode the controllable text description into the generation process.

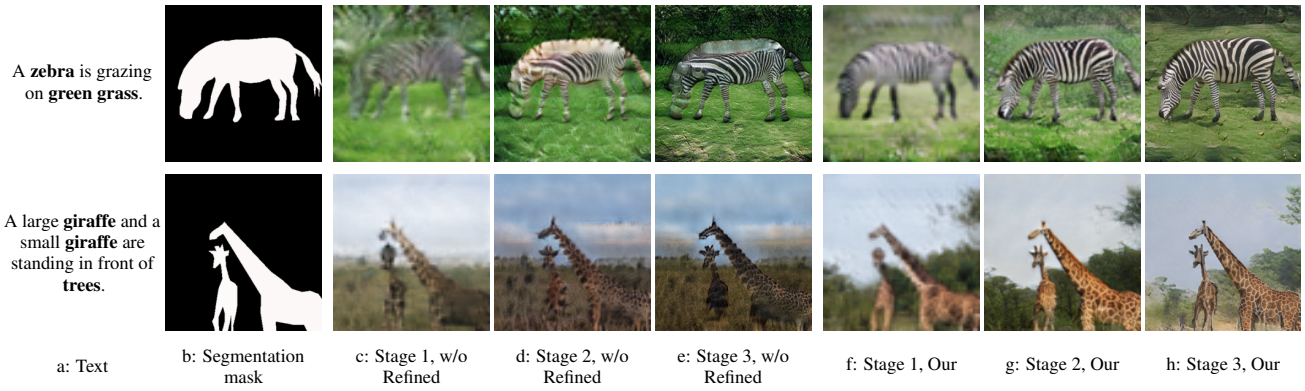

Figure 7. Effectiveness of refined multi-stage architecture. *c*, *d*, and *e* show the synthetic images produced at each stage by the model without refined multi-stage architecture. *f*, *g*, and *h* show the synthetic images generated at each stage by our model.

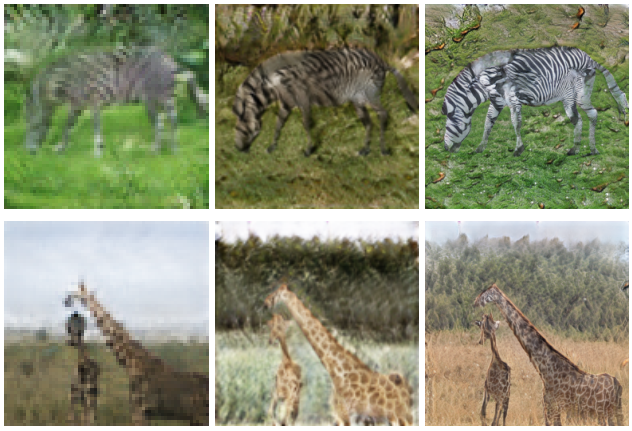

Figure 8. Rectification ability of generators in RefinedGAN. *a* denotes images generated at the first stage by the model without refined architecture. *b* and *c* denote our model takes this flawed features and feed through stages 2 and 3 progressively, producing the corresponding images shown at *b* and *c*.

Effectiveness of refined multi-stage architecture. To demonstrate the effectiveness of refined multi-stage architecture, an ablation study is conducted, shown in Fig. 7. The model without the refined multi-stage architecture, i.e., without feeding patches of images at higher stages to lower ones, 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, the generators at the following stages fail to complete the missing content or rectify inappropriate attributes, and just leave it without any correction (see columns *d* and *e* of Fig. 7). To further verify the rectification ability of the refined multi-stage architecture, we feed the flawed features generated by the model without refined multi-stage architecture at the first stage to our full model. As we can see in Fig. 8, 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.

Also, more examples in Fig. 5 *d* show that images produced by the model without refined multi-stage architecture keeps flawed and incorrect details without rectification, e.g., the inappropriate green window of the bus, and there is a stripe of texture missing on the head of airplane at bottom of *d*.

Effectiveness of structure loss. To verify the effectiveness of structure loss, we remove it from our model, shown in Fig. 5 *e*. As we can see, the synthetic images produced by the model without the structure loss contain some unrealistic regions, e.g., there are some unrealistic patches on the bus, and the appearance of the airplane is far from satisfactory. However, our model can generate not only realistic objects but also a high-quality background, which potentially demonstrates that the structure loss can improve the differential ability of discriminators to identify fake images if there exist small unreasonable areas only in objects or the background, and in turn improve generators to produce higher-quality images with finer details everywhere.

5. Conclusion

We have proposed a novel generative adversarial network, called RefinedGAN, which effectively embeds controllable factors, i.e., natural language descriptions, into image-to-image translation to control the generation of objects and visual attributes. Also, our model can disentangle objects from the background, produce complete images at lower stages and enable a great rectification ability. Extensive experimental results demonstrate the advantages of our method, with respective to both high-quality image generation and the effectiveness of control of local visual attributes.

References

Ashual, O. and Wolf, L. Specifying object attributes and relations in interactive scene generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4561–4569, 2019.

Bird, S., Klein, E., and Loper, E. *Natural language processing with Python: Analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

Chen, Q. and Koltun, V. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1511–1520, 2017.

Denton, E. L., Chintala, S., Fergus, R., et al. Deep generative image models using a Laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems*, pp. 1486–1494, 2015.

Dong, H., Yu, S., Wu, C., and Guo, Y. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5706–5714, 2017.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Hong, S., Yang, D., Choi, J., and Lee, H. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7986–7994, 2018.

Huang, X., Li, Y., Poursaeed, O., Hopcroft, J., and Belongie, S. Stacked generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5077–5086, 2017.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, 2017.

Johnson, J., Gupta, A., and Fei-Fei, L. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1219–1228, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Li, B., Qi, X., Lukasiewicz, T., and Torr, P. Controllable text-to-image generation. In *Advances in Neural Information Processing Systems*, pp. 2063–2073, 2019a.

Li, B., Qi, X., Lukasiewicz, T., and Torr, P. Manigan: Text-guided image manipulation. *arXiv preprint arXiv:1912.06203*, 2019b.

Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S., and Gao, J. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12174–12182, 2019c.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Mo, S., Cho, M., and Shin, J. Instagan: Instance-aware image-to-image translation. *arXiv preprint arXiv:1812.10889*, 2018.

Nam, S., Kim, Y., and Kim, S. J. Text-adaptive generative adversarial networks: manipulating images with natural language. In *Advances in Neural Information Processing Systems*, pp. 42–51, 2018.

Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2337–2346, 2019.

Pavllo, D., Lucchi, A., and Hofmann, T. Controlling style and semantics in weakly-supervised image generation. *arXiv preprint arXiv:1912.03161*, 2019.

Qi, X., Chen, Q., Jia, J., and Koltun, V. Semi-parametric image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8808–8816, 2018.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

Shaham, T. R., Dekel, T., and Michaeli, T. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4570–4580, 2019.

Tang, H., Xu, D., Yan, Y., Torr, P. H., and Sebe, N. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. *arXiv preprint arXiv:1912.12215*, 2019.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8798–8807, 2018.

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., and He, X. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1316–1324, 2018.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5907–5915, 2017.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1947–1962, 2018.

Zhao, B., Meng, L., Yin, W., and Sigal, L. Image generation from layout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8584–8593, 2019.