SUPPLEMENTARY MATERIAL FOR PHRASEGAN: PHRASE-BOOST GENERATIVE ADVERSARIAL NETWORK FOR TEXT-TO-IMAGE GENERATION

Fei Fang[1], *Ziqing Li*[1], *Fei Luo*[1*], *Chengjiang Long*[2], *Shenghong Hu*[3], *and Chunxia Xiao*[1*]

1. School of Computer Science, Wuhan University, Wuhan 430072, China;
2. JD Finance American Corporation, Mountain View, CA, USA;
3. Information Engineering School, Hubei University of Economics, Wuhan 430205.
fangfei369@163.com, thalialee@163.com, luofei_whu@126.com, cjfykx@gmail.com,
wuhanhush@126.com, cxxiao@whu.edu.cn

ABSTRACT

In this supplementary material, we add some implementation details and additional experimental results. We describe the memory features added in word embedding and summarize the objective function in terms of implementation details. In terms of experimental results, we first provide user studies to evaluate the quality of the generated images more comprehensively. Second, we illustrate the existing limitations through some failure results. Third, we visualize the attention maps to show that the proposed method correctly captures the phrase-object relationship. Finally, we compare more methods for qualitative experiments and show the qualitative results of the ablation study.

1. IMPLEMENTATION DETAILES

1.1. Memory feature used in word embedding

As mentioned in Section 3.1.1 of the main paper, we concatenate the memory feature m_r constructed by [1] with the initial word embedding before the transformer encoding. Fig. 1 shows the memory construction method of CPGAN [1]. To construct the memory m_r, they considered salient regions from all relevant images across the training dataset to capture full semantic correspondence. Please refer to [1] for more details.

1.2. Objective function

The final objective function of the generators in our method mainly include three terms:

$$\mathcal{L} = \mathcal{L}_G + \lambda_1 \mathcal{L}_{OAIE} + \lambda_2 \mathcal{L}_{GTISM}, \tag{1}$$

where $\mathcal{L}_G = \sum_{i=0}^{m-1} \mathcal{L}_{G_i}$, $\mathcal{L}_{G_i}$ is the adversarial loss for the i-th generator, and m is the number of generators. For each generator, the adversarial loss $\mathcal{L}_{G_i}$ is the sum of the discriminators' unconditional and conditional losses:

$$\begin{aligned}
\mathcal{L}_{G_i} = & \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{x}\sim p_{G_i}}[\log(D_i^{obj}(\hat{x}_i))]}_{\text{POD uncond loss}} \\
& \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{x}\sim p_{G_i}}[\log(D_i^{obj}(\hat{x}_i, P, S))]}_{\text{POD cond loss}} \\
& \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{x}\sim p_{G_i}}[\log(D_i^{pat}(\hat{x}_i))]}_{\text{FGCD uncond loss}} \\
& \underbrace{-\frac{1}{2}\mathbb{E}_{\hat{x}\sim p_{G_i}}[\log(D_i^{pat}(\hat{x}_i, E, S))]}_{\text{FGCD cond loss}}
\end{aligned} \tag{2}$$

where $\hat{x}_i$ is the generated scene image from the i-th generator, D_i^{obj} is the i-th POD, D_i^{pat} is the i-th FGCD for the equally divided grid regions. The second term $\mathcal{L}_{OAIE}$ in Equation 1 is imported from [1] to compute the text-image matching score of the unimportant regions. The last term $\mathcal{L}_{GTISM}$ in Equation 1 is the loss computed using the contextual phrase and subregion features from our GTISM. In our experiment, we set $\lambda_1 = 50$ and $\lambda_2 = 30$ to make the two terms play the

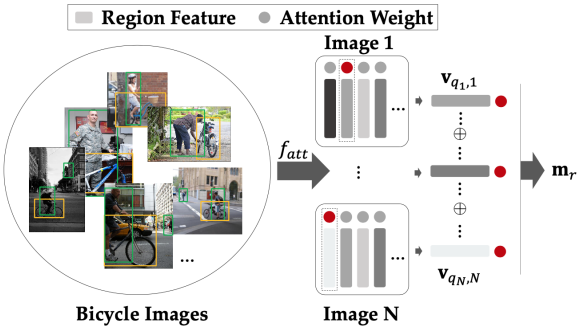
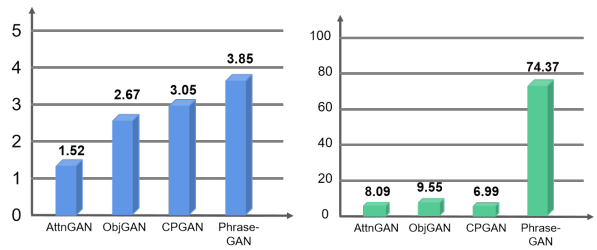


Fig. 1. The scheme of memory construction in CPGAN [1].

This work is partially supported by the Key Technological Innovation Projects of Hubei Province (2018AAA062) and NSFC (No. 61972298). * Chunxia Xiao and Fei Luo are corresponding authors.



(a) Realism (b) Text-image consistency

Fig. 2. The user study results of our PhraseGAN. The realism evaluation result is shown in (a) and text-image consistency result is shown in (b).

same role.

2. ADDITIONAL EXPERIMENTAL RESULTS

2.1. User study

Although we use three evaluation metrics to measure the quality of the generated images, they can not replace human visual perception. We invite 50 human users to evaluate the realism of the generated images and the text-image consistency between the generated images and the input text. We choose four different methods for comparison, namely AttnGAN [2], ObjGAN [3], CPGAN [1], and PhraseGAN (ours). All these methods are trained and tested on the MSCOCO14 dataset. For each method, we randomly select 25 generated images from their testing results.

For the realism evaluation, we require the users to score the realism for the 100 generated scene images from 1 to 5, and a higher score means a better sense of reality. We calculate the average scores of all selected images for each method. For text-image consistency, the 100 generated images are divided into 25 groups, and each group contains four images generated by different methods. We require the users to select one most consistent image with the corresponding given caption for each group. Note that the text-image consistency includes the correctness of the relative positional relationships between pairs of objects in the generated scene images. Finally, we count the percentage of each method selected by the users. The user study results are shown in Fig. 2.

From the result in Fig. 2 (a), we can find that users think our generated scene images are more realistic than the images generated by the other three methods. The result of Fig. 2 (b) shows that 74.37% of the users think that the images generated by our PhraseGAN are more consistent with the input text.

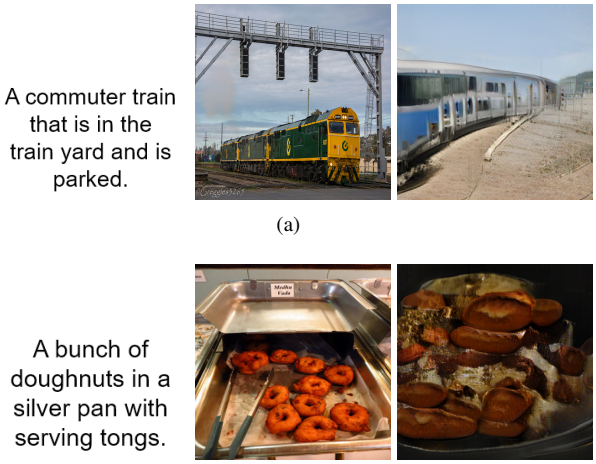


Fig. 3. The two main limitations of our method. In subfigures (a) and (b), the input texts are on the left, the images in the middle are the ground truth images, and the images on the right are the corresponding generated images.

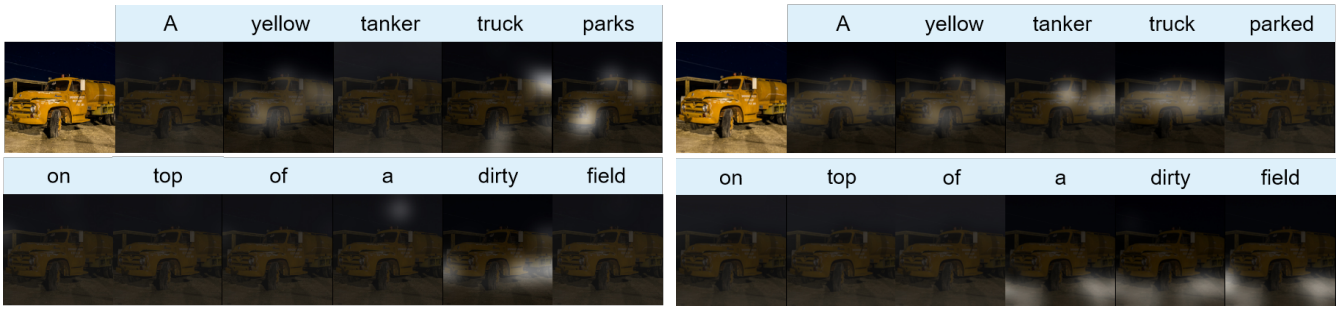
2.2. Limitations

On the one hand, the implicit attributes of the objects in the generated images are occasionally different from the ground truth images. The implicit attributes are the ones that are not mentioned in the input text but appear in the ground truth images. As we all know, the ground truth images in the dataset may contain much more information than their captions, which is called the information gap. As shown in Fig. 3, the subfigure (a) shows that the generated train has a different color from the ground truth one since the input text does not mention the color of the train.

On the other hand, the object detector of YOLOv4 is not always reliable. Occasionally, it will miss the objects that we do not use to pre-train the object detection networks. Therefore, the generation quality of the undetected objects cannot be enhanced by GTISM and POD. These objects will look worse than other objects in the generated images or even be lost. Fig. 3 (b) shows that we lost the tongs mentioned in the text due to the pre-trained YOLOv4 detector cannot detect them in the dataset images.

2.3. Attention maps

We visualize attention maps generated by using phrases and words in the attention mechanism of multi-step generation, respectively. From Fig. 4 we can find that in the attention map generated by adding phrases to the attention mechanism, descriptive words that belong to a phrase tends to pay attention to the same region. Whereas in the attention map generated by using only words, a region may miss the attention of some related words, which will gain a negative impact on the scene image generation process.



(a) Attention map by word attention mechanism.

(b) Attention map by phrase attention mechanism.

Fig. 4. The comparison of attention maps between using words (a) and phrases (b).

2.4. Additional qualitative comparison with some state-of-the-art methods

Fig. 5 shows the additional qualitative comparison results between our method and three other methods, namely AttnGAN [2], MirrorGAN [4], DMGAN [5]. Through the results in Fig. 5, we can observe that the scenes in the generated image by our model are also more consistent with the given text.

2.5. Visualization of ablation study

The scene images generated by each ablation study are shown in Fig. 6. In the ablation study of PhraseGAN-TTE, we use the Transformer-based text encoder (TTE) to encode the input text into the sentence and word embeddings. The qualitative results in Fig. 6 demonstrate that the TTE module can help generate better images than the baseline method. It proves that the TTE module has better performance than the traditional LSTM-based text encoder because the TTE can better extract the semantic features in the input text and fully exploit the relevance between different words.

In the ablation study of PhraseGAN-GTISM, we use the proposed GTISM module to evaluate the phrase-object similarities and the accuracy of relative positions between object pairs. PhraseGAN-GTISM generates better quality scene images shown in the third row of Fig. 6 than the baseline method and PhraseGAN-TTE. This experiment demonstrates that the GTISM module can effectively promote the realism and diversity of the generated images.

In the final ablation study of PhraseGAN-POD, we use the proposed POD as the discriminator to train the image generator. The results demonstrate that the proposed POD can effectively improve the semantic consistency of the input text and generated images.

3. REFERENCES

[1] Jiadong Liang, Wenjie Pei, and Feng Lu, "Cpgan: Content-parsing generative adversarial networks for text-to-image synthesis," in *European Conference on Computer Vision*. Springer, 2020, pp. 491–508.

[2] 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*, 2018, pp. 1316–1324.

[3] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12174–12182.

[4] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao, "Mirrorgan: Learning text-to-image generation by redescription," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1505–1514.

[5] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5802–5810.

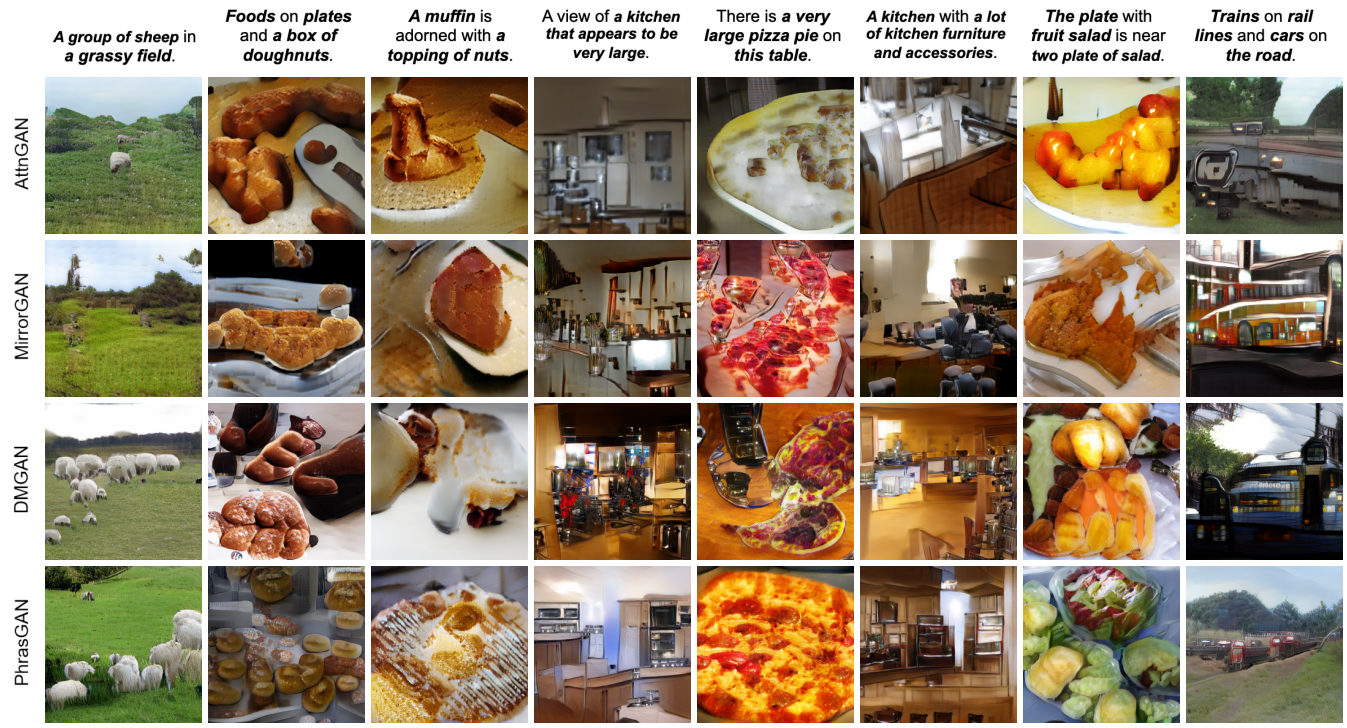


Fig. 5. Qualitative comparison between our method (last row) and three other methods, namely AttnGAN [2] (first row apart from the input texts), MirrorGAN [4] (second row), DMGAN [5] (third row).

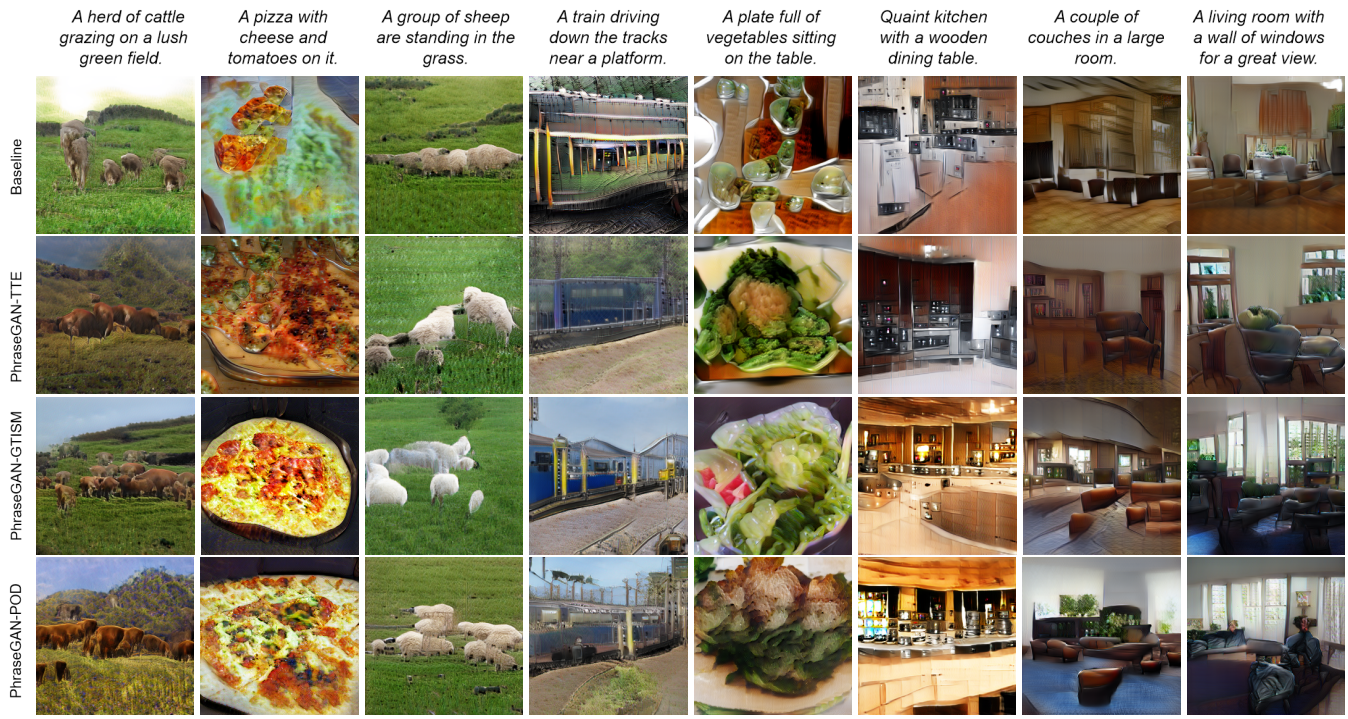


Fig. 6. The ablation study results of the baseline method and the proposed three modules of our PhraseGAN.