Photographic Text-to-Image Synthesis with a Hierarchically-nested Adversarial Network

We would like to thank the three reviewers for their thoughtful review of the this manuscript and for their insightful comments and instructive suggestions, which help to improve quality of the manuscript. Our response follows.

To Reviewer 1: 1) *Novelty of hierarchically nested discriminators:* Although the concept of 'multi-resolution branch' has been used in several other fields mentioned by the reviewer and Related Works, but it's effectiveness has not been studied in GANs. Training high-resolution GAN is known to be highly challenging. Previous methods resort to stacking a set of small GANs and train them progressively (see paragraph 4, Section 2). Differently, our method firstly shows that incorporating such hierarchical discriminators can overcome unsolved difficulties and assist the training of high-resolution GANs in an end-to-end manner. Extensive experiment results demonstrate the superiority of our method.

2) *Additional citations and Missing details in Figure 2*: Thanks for pointing out the additional related works, we will cite them properly. We will also carefully revise the figure to make it intuitive and clear.

3) *Scope of text-to-image synthesis:* We focus on text-to-image synthesis as it is a relative new but key task in GANs, which already has plenty of difficulties that worth studies in a full paper. To be specific, achieving high diversity and semantic consistency simultaneously in high-resolution samples generation is a highly difficult problem even with provided charRNN text encodings. Nevertheless, as the reviewer indicated, our method looks universal to other type of image generation problems; We believe that exploiting other image generation tasks using our method will also be an interesting research area left open for the community. We will test them in the next version and open source our code to the community for future study. we also want to emphasis that the used datasets are also large (especially for COO), representative and are capable of proving the effectiveness of our method. In addition, the evaluation of text-to-image synthesis also needs rethinking. We propose the visual-semantic metric to alleviate labor evaluation used in [44] for text-to-image synthesis evaluation.

To Reviewer 2: 1) *Insufficient comparison:* Prog.GAN is currently recognized as one of the most effective methods to generate high-resolution images from noises. Outperforming recent Prog.GAN already implies that our method outperforms much earlier LAPGAN (proposed in 2015). Moreover, LAPGAN only reports images of resolution up to 96^2 , while our capability for high-resolution also demonstrates significant advantages. The cascaded refinement network is an image-to-image generative model with pixelwise semantic layout input, which is not relevant. Moreover, all of the three are actually not for text-to-image synthesis,we discussed them in the related work as they share similar high-level motivations with ours. We have showed the state-of-the-art performance compared with most recent existing text-to-image synthesis methods on three datasets (including the best performance on the large COCO dataset) with three evaluation metrics. The sufficiency and solidity of our experiments are highlighted by other two Reviewers as well.

2) *Comparison to [28]:* We appreciate the recommended new metrics. We tested our model on the CUB bird dataset following the procedure advised in [28] (i.e. query text, the inception model, and a bird word list used to match the ImageNet bird categories). Our method achieves a high (top-1) accuracy of 98.7 (256^2 images) compared with [28]'s 85% (for a few bird category names in ImageNet do not have an exact match in the bird word list, we manually checked image-by-image). Particularly, less than ten images completely failed and the rest misclassified images still look like birds visually. We will complete the evaluation with this metric in the final version. The author ordering of [28] will be corrected. All results with source code will be released for wide test.

4) *Qualitative results of failed cases*: We definitely agree with the reviewer that it is better to show some failure cases for completeness. We will try our best to compare and discuss this in the revision. In addition, Table 3 left compares class-wise MS-SSIM scores to StackGAN, which illustrates that our method does not outperform StackGAN 100% of the time. We hope this will also be helpful for the concerns.

3) *About the style transfer loss:* We agree with the nice study in [28] that the perpetual loss is helpful to improve the classification accuracy. But we also have the concern (as confirmed by our early experiment using intermediate representation of discriminator) that adding such strong loss might lead to model collapse and impede the diversity of the generated images, since balancing the generator and discriminator can be a very tricky task. However, we agree that finding an effective way to utilize the mentioned style transfer loss can be a future research topic that worth studies.

To Reviewer 3: 1) *Writing issues:* Thanks for pointing them out. The intuition behind VS-similarity is to increase the score c (cosine similarity) between matching pairs while decrease these between the mismatching pairs. We will definitely revise the manuscript to better explain the intuition and correct the misspelling of MS-SSIM.

2) *Choosing R_i:* The actual size of R_i for discriminators is described in the Training and Architecture Details section in the supplementary material. Our preliminary experience is that choosing small R_i at low resolution and a bit larger one at high resolution.