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

We appreciate the insightful comments and instructive suggestions by three reviewers. In brief, our paper presents a simple yet effective generative adversarial network (GAN) with hierarchically-nested discriminators to perform photographic text-to-image generation, which is end-to-end trainable and achieves state-of-the-art performance on three public datasets.

To Reviewer 1: 1) *Novelty of hierarchically nested discriminators:* Although the idea of 'multi-resolution branch' has been exploited 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 effectively regularize mid-level representations and assist generator training of high-resolution GANs in a fully end-to-end manner, and 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) *Beyond text-to-image synthesis:* Text-to-image synthesis is a relative new but challenging task in GANs. Itself already has plenty of problems that worth studies in a full paper. Nevertheless, as the reviewer indicated, our method looks universal to other type of image generation problems, but in this work, we focus on the text-to-images synthesis, 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. We believe that exploiting other image generation problems using our method will also be an interesting research area left open for the community. We will test them in the future work and open source our code to the community for future study. The evaluation of text-to-image synthesis also needs rethinking. We propose the visual-semantic metric to alleviate labor evaluation used in [44] in complementary of the Inception-score and MS-SSIM for general GANs.

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 as they share similar high-level motivations and solutions 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 inception score 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% (To be careful, 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*: Thanks for pointing this out. We will show more diverse qualitative results and discuss failure cases (especially in COCO dataset) in the revision. Based on our observation, our method shows generally more realistic results, with clearly less sharp pixel transitions, more photographic colors, and nature saturability (see the supplementary material). And the evaluation results with three metrics also confirm our conclusion.

3) *Using style transfer loss helps the inception score?:* 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. We will definitely revise the manuscript to better explain the intuition behind the proposed visual similarity metric 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.