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 achieves state-of-the-art performance with sufficient experiments.

To Reviewer 1: 1) *Novelty of hierarchically nested discriminators:* Although the idea of 'multi-resolution branch' is used in several other fields mentioned by the reviewer and Related Work, it is completely novel in GANs. Note that training high-resolution GAN is known to be highly challenge. Previous methods rely on 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 to capture the complex image statistics and enable the training of high-resolution GANs in a fully end-to-end manner.

2) *Missing details in Figure 2*: We will 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 key task in GANs. Itself already has enough problems that worth studies in a full paper. To be specific, keeping high diversity and semantic consistency simultaneously of high-resolution samples conditioned one sentence is non-trivial at all, although with provided text charRNN encodings. Besides the explained designs, there are also some technical details to guarantee them in the discriminator designs. We will further explain these merits in the final version. The evaluation of text-to-image synthesis also needs consideration. We propose the visual-semantic metric to alleviate the previously used labor evaluation in complementary of the Inception-score and MS-SSIM for general GANs. But as the reviewer mentioned, our method can be general to other image datasets clearly. We will test them as a future study and open source code to the community for wide validation.

To Reviewer 2: 1) *Insufficient comparison:* Prog.GAN is currently recognized as the most effective method to generate high-resolution images from noises. GANs are developing fast. Outperforming recent Prog.GAN already implies that our method outperforms much earlier LAPGAN (proposed in 2015). Moreover, LAPGAN can only generate up to 96^2 images, while our capability for high-resolution also demonstrates more advantages. The cascaded refinement network is an image-to-image generative model, which is clearly not relevant. Moreover, all of the three are actually not for text-to-image synthesis, so are not main comparable methods. We mentioned them as they share similar high-level motivations with ours. But we tried our best to show the state-of-the-art performance compared with most existing text-to-image synthesis methods on three datasets with three metrics. The sufficiency and solidity of our experiments are highlighted by the other two Reviewers as well.

2) *Comparison to [28]:* We appreciate the recommended new metrics in [28]. We tested our trained 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*: We will discuss the failure cases in the revision. Based on our observation, our method shows generally more realistic results, with clearly less sharp pixel transitions and more photographic colors as well as nature saturability (see the supplementary material). The reason we did not add failure cases in the first version is that it is hard to explain why a certain case looks worse than another. But we thought quantitative results (class-wise comparison in Table 3) are helpful.

3) *Using style transfer loss helps the inception score?:* We agree that adding semantic information in the intermediate outputs (like StackGAN), adding some style losses like [28], or even adding an inception-score loss, can push the inception score. But it not clear if it can enable end-to-end high-resolution GAN training as our method is capable of. Also please note, it is admitted that the inception score is not "perfect" and can not be the sufficient condition/proof of high-quality and semantic consistency images, considering the fact that CNNs (inception) can be "fooled" by only pixel changes without overall visual quality improvement. All-round improvement of qualitative results and other scores is necessary. Our paper has a careful consideration and thorough evaluation for them.

To Reviewer 3: 1) *Writing issues:* Thanks for pointing them out. We will definitely revise the manuscript to better explain 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 supplementary material. Our preliminary experience is that choosing small R_i at low resolution and a bit larger one at high resolution.