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 on three public datasets.

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. 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 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 plenty of 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 char-RNN encodings. There are essential technical details to guarantee them in the discriminator designs. The evaluation of text-to-image synthesis also needs rethinking. We propose the visual-semantic metric to alleviate labor evaluation in [44] in complementary of the Inception-score and MS-SSIM for general GANs. The used datasets are also large (especially for COO) and representative. Nevertheless, as the reviewer indicated, our method is definitely general to other datasets. We will test them in the next version and open 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 significant 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 essential comparable methods. We discussed them as they share similar high-level motivations and solutions with ours. We tried our best to show the state-of-the-art performance compared with most existing text-to-image synthesis methods on three datasets (including the best inception score on the large COCO dataset) with three 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*: 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, more photographic colors, and nature saturability (see the supplementary material). The reason we did not add failure cases is that it is hard to explain why a certain case looks worse than another. But Table 3 left compares class-wise scores to StackGAN, which is helpful for the concerns.

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 directly, can push the inception score. But it not clear these approaches can enable end-to-end high-resolution GAN training as our method is capable of. And Investigation of them is out of our focus. 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 metrics is necessary. Our paper has a careful consideration and thorough demonstration for them.

To Reviewer 3: 1) *Writing issues:* Thanks for pointing them out. We will definitely revise the manuscript to better explain the intuition of 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.