Appendices

A. Training Details

The Adam optimizer is used for all experiments. The initial learning rate is set as 0.0002 and decreased by half for every 100 epochs (50 for COCO). The CUB and Oxford-102 datasets are trained for 500 epochs in total (200 epochs for COCO). For the local image loss, we set $R = 1$ and for 64^2 and 128^2 side outputs and $R = 5$ for 256^2 and 512^2 outputs.

All intermediate conv layers except from the specified ones for both generators and discriminators use 3×3 kernels (with reflection padding). We also experimented other normalization (i.e. instance normalization [3] and layer normalization [1]) used by recent advances [4, 2]. Both are not satisfactory.

We use 1-repeat residual block for the generator till 256^2 resolution. The input of the generator is a $1024\times4\times4$ tensor. As the feature map size increases, the number of feature maps is downsampled by 2 at $8, 32, 128, 256$ sizes in generator. To generate 512^2 images, we pre-train the generator to 256^2 due to the limitation of GPU memory. We use 3-repeat res-block followed by the stretching and linear compression layer. Since the 256^2 image already captures the overall semantics and details, to encourage the 512^2 maintain these information, we use a l1 reconstruction loss to self-regularize the generator in case the discriminator confuses the expected output.

B. Results on the CUB bird dataset

References

[1] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. *ICCV*, 2017.

[3] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[4] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *ICCV*, 2017.