Image-to-Image Translation with Text Guidance

Bowen Li, Philip H.S. Torr, and Thomas Lukasiewicz

UNIVERSITY OF OXFORD

Introduction

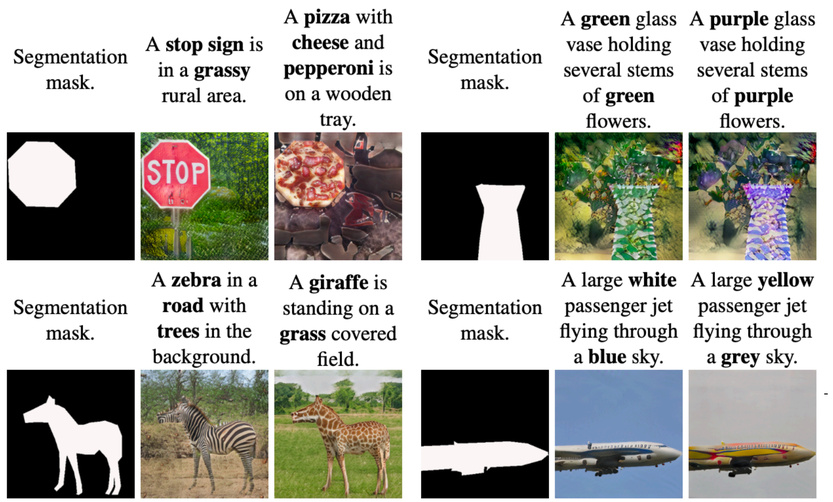


Fig. 1. Given a segmentation mask and a text provided by a user that describes desired objects and visual attributes, the goal of this model is to generate realistic images semantically matching the given descriptions with the global structure defined by the masks.

Method

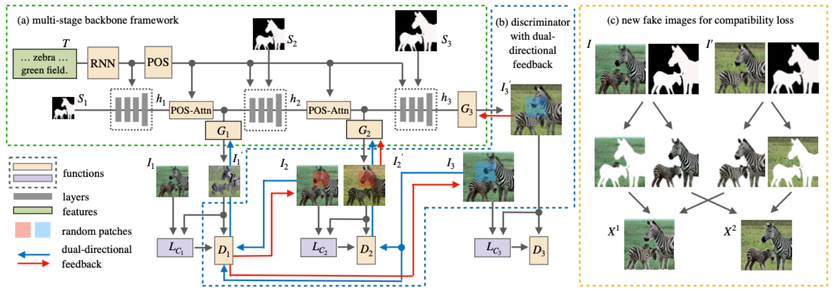
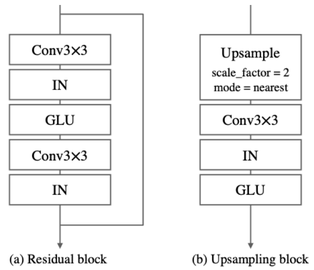


Fig. 2. Architecture of our network. POS-Attn denotes the part-of-speech tagging-based attention. Lc denotes the compatibility loss.

Method



Experiments

Table 1. Quantitative comparison: FID, IS, and R-prcn of ours and baselines on the COCO dataset. For FID, lower is better, for IS and R-prcn, higher is better.

Method	FID	IS	R-prcn (%)
SPADE	42.74	11.69 ± 0.26	-
AttnGAN-Seg	32.39	12.09 ± 0.28	75.24 ± 3.39
ControlGAN-Seg	31.41	11.56 ± 0.16	80.43 ± 2.79
Ours	**28.30**	$\mathbf{15.96 \pm 0.16}$	$\mathbf{83.23 \pm 1.37}$

Fig. 3. Qualitative comparison of SPADE and ours on the COCO dataset.

Experiments

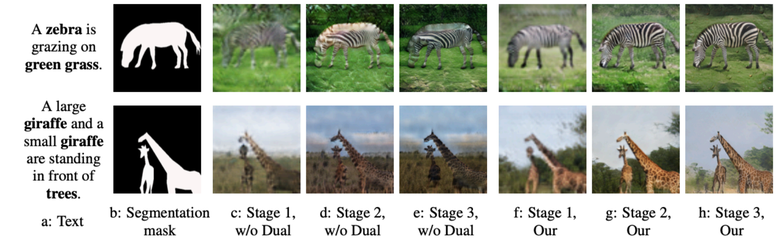


Fig. 4. Effectiveness of dual-directional feedback. c, d, and e show the synthetic images produced at each stage by the model without adopting the discriminator with dual-directional feedback. f, g, and h show the synthetic images generated at each stage by our full model.

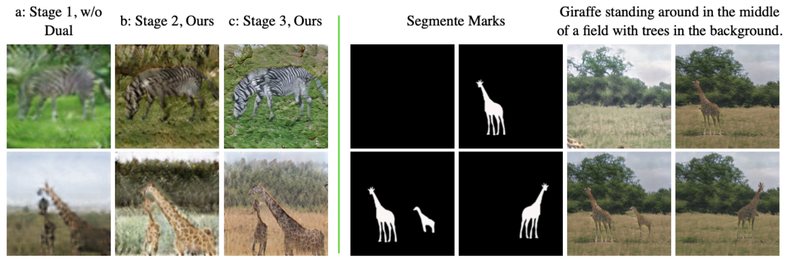


Fig. 5. Left: rectification ability of our generators. a denotes images generated at the first stage by the model without dual-directional feedback. In b and c, denote our model takes these flawed features and feeds them through stages 2 and 3 progressively, producing the corresponding images shown at b and c. Right: disentanglement of objects and background.

Table 2. Ablation studies of different components used in our approach.

Method	FID	IS	R-prcn (%)
Ours w/o Dual	32.14	12.16 ± 0.20	80.13 ± 2.20
Ours w/o Compatibility	29.47	14.72 ± 0.32	81.43 ± 1.21
Ours w/o POS-Attn	32.72	12.77 ± 0.21	81.07 ± 1.60
Ours w/ WSA [32]	30.14	14.49 ± 0.15	82.03 ± 1.03
Ours	**28.30**	$\mathbf{15.96 \pm 0.16}$	$\mathbf{83.23 \pm 1.37}$