Memory-Driven Text-to-Image Generation

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

Introduction

A zebra is standing on the grassy field.

A white and blue bus is driving down a street.

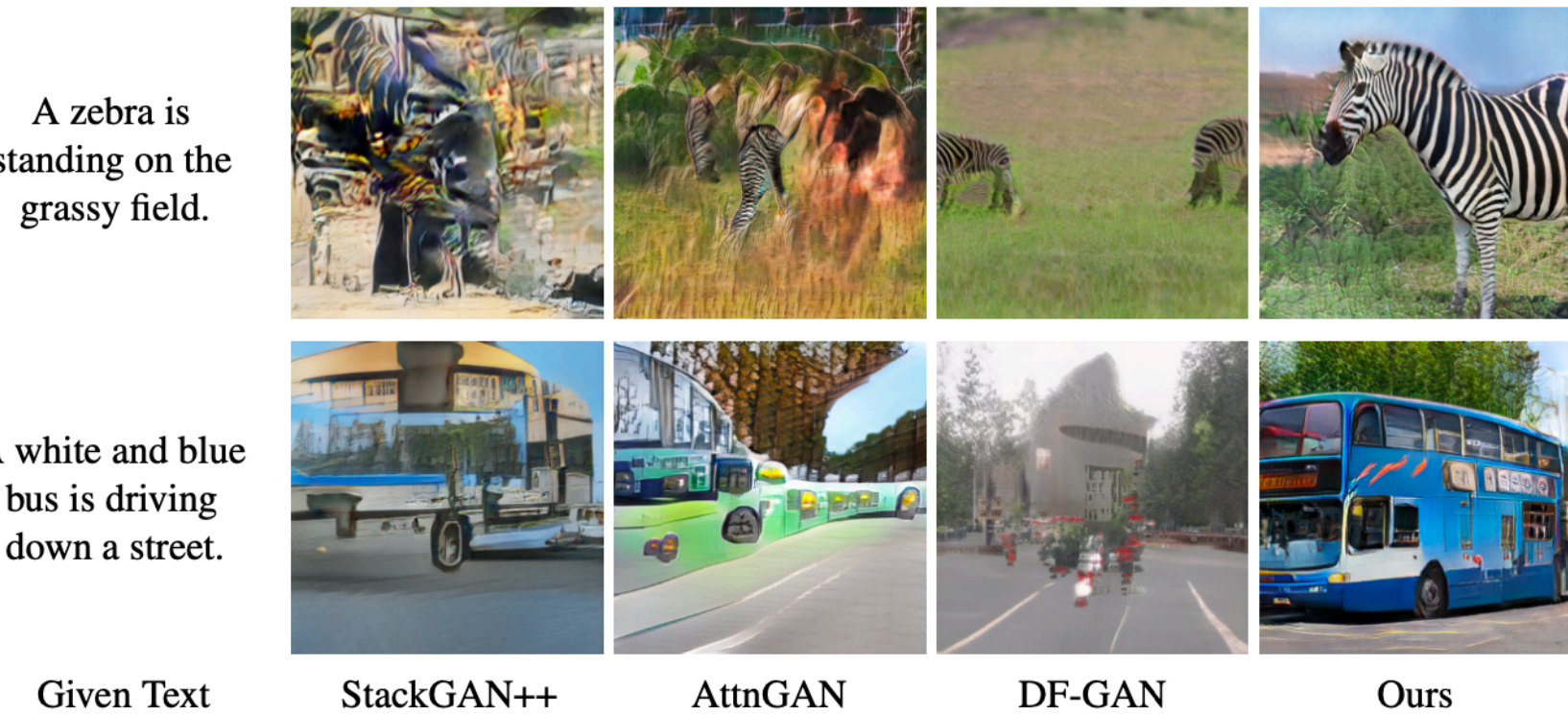Given Text — StackGAN++ — AttnGAN — DF-GAN — Ours

Fig. 1. Examples of text-to-image generation on COCO. Current approaches only generate low-quality images with unrealistic objects. In contrast, our method can produce realistic images, in terms of both visual appearances and geometric structure.
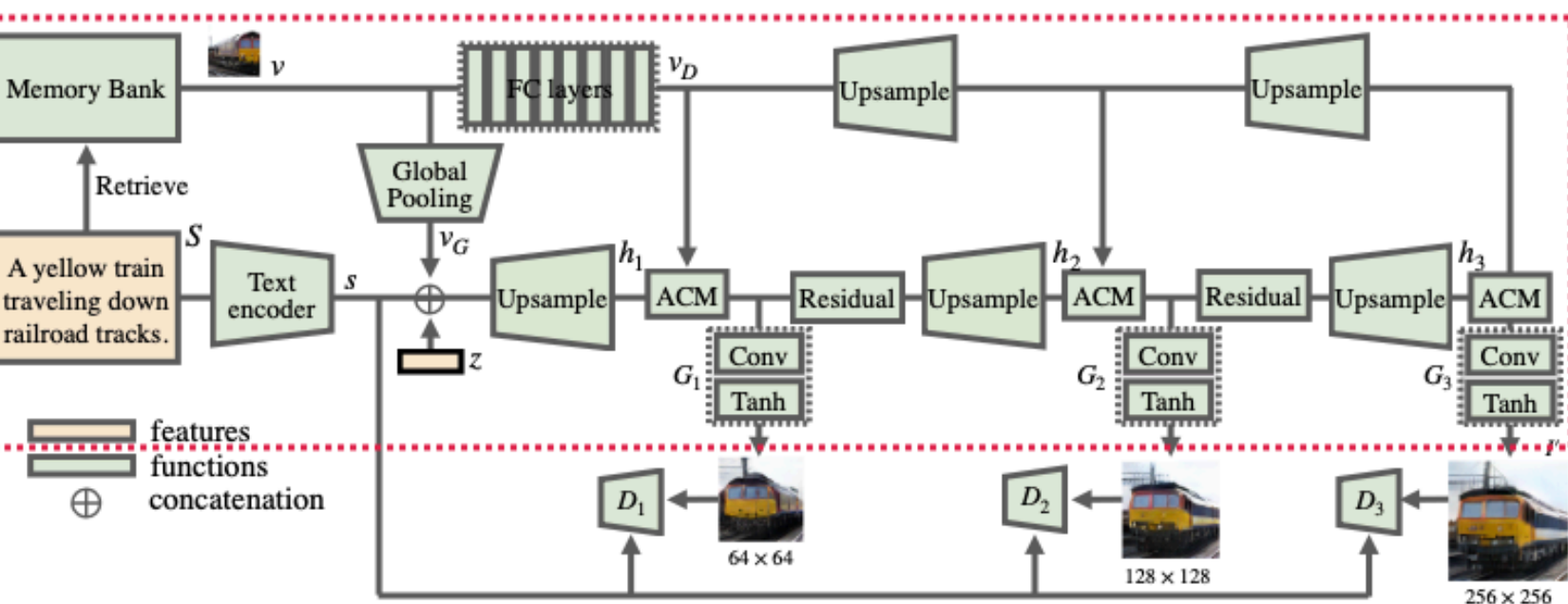
Method

Fig. 2. Top: architecture of our proposed method. The red box indicates the inference pipeline that retrieves image features from the memory bank according to the given text description S; during training, we directly feed image features from the text-paired training image. z is a random vector drawn from the Gaussian distribution.

Experiments

Table 1. Quantitative comparison: IS, FID, R-precision, SOA-C, and SOA-I of current methods and our approach on the CUB and COCO datasets.

Method	CUB			COCO				
	IS	FID score	R-precison	IS	FID score	R-precison	SOA-C	SOA-I
Real Images	25.34	-	89.17	34.88	-	92.71	74.97	80.84
AttnGAN [49]	4.36	23.98	67.82	25.89	32.32	85.47	25.88	39.01
ControlGAN [22]	4.58	13.92	69.33	24.06	33.58	72.43	-	-
MirrorGAN [40]	4.56	-	57.67	26.47	-	74.52	-	-
DM-GAN [56]	4.75	16.09	72.31	32.32	32.64	88.56	33.44	48.03
DF-GAN [46]	5.10	14.81	-	-	21.42	-	-	-
XMC-GAN [52]	-	-	-	30.45	9.33	-	50.94	71.33
LAFITE	**5.97**	**10.48**	-	32.34	**8.12**	-	61.09	74.78
DALL-E [42]	-	-	-	-	~20	-	-	-
GLIDE [35]	-	-	-	-	12.89	-	-	-
CP-GAN [29]	-	-	-	**52.73**	55.82	**93.59**	**77.02**	**84.55**
Obj-GAN [27]	-	-	-	30.29	36.52	87.84	27.14	41.24
OP-GAN [12]	-	-	-	27.88	24.70	89.01	35.85	50.47
Ours	5.91	10.49	**73.87**	29.36	19.47	90.32	47.46	65.83

Fig. 3. Qualitative results on CUB and COCO.
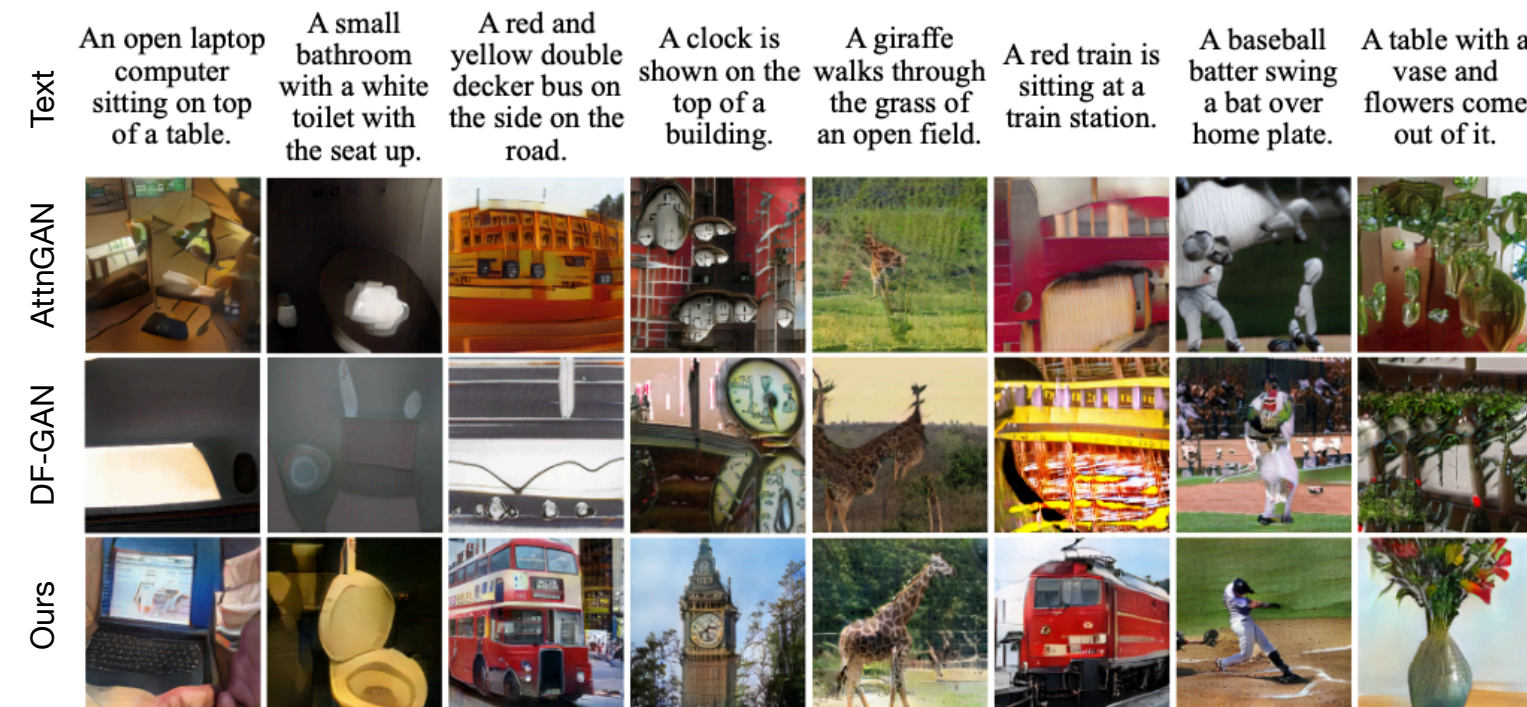
Fig. 4. Qualitative comparison between AttnGAN, DF-GAN, and Ours on COCO.

Experiments

A yellow bird has brown wings, and a yellow belly.

A yellow and green train is traveling through a station.

A red double decker bus driving down a street.

Fig. 5. Diversity of our approach, where we only change the value of noise vector z.

Given Sentence: A zebra is grazing on green grass. — A zebra is walking in an open grassy filed. — A bus is driving down a street. — A bus is parking on the side of the road.
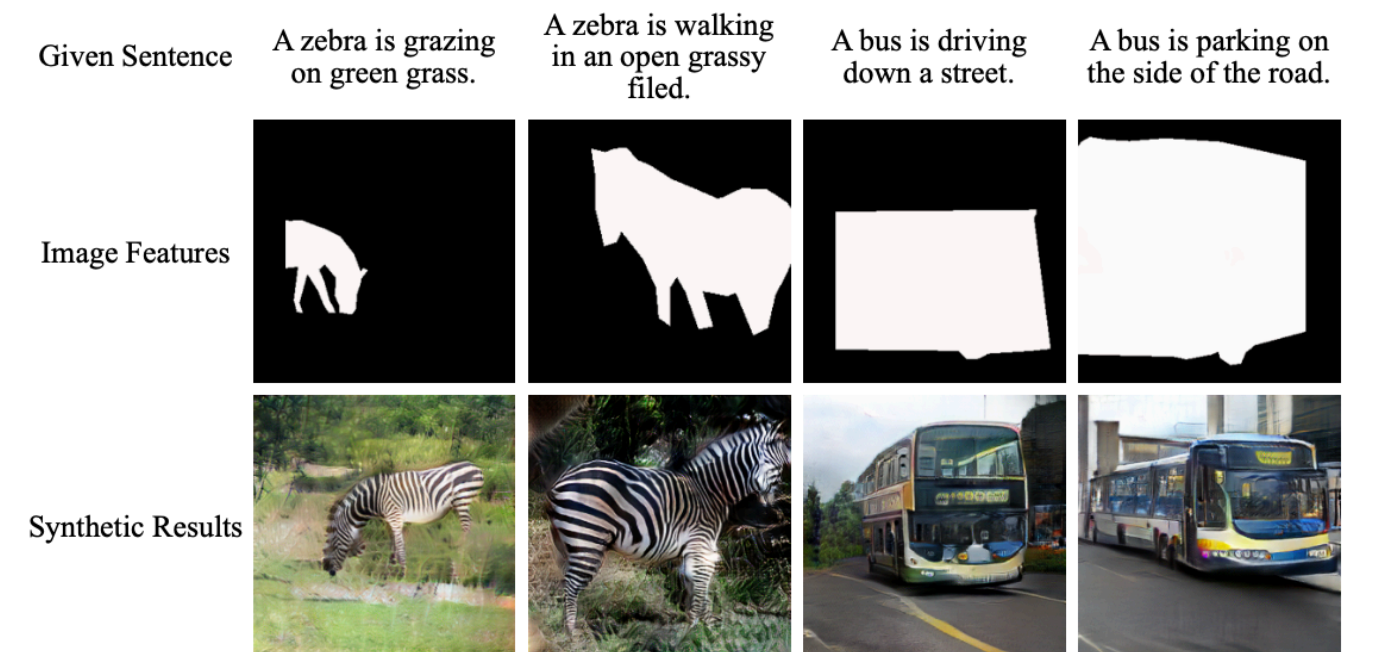
Image Features

Synthetic Results

Fig. 6. Semantic information exploration by feeding semantic masks into our network to generate realistic images.

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

Method	FID	R-psr
Ours w/o Feature	22.20	84.63
Ours w/o Disen.	18.82	92.17
Ours w/o Disen.*	18.80	67.05
Ours w/o Content	20.96	88.95
Ours w/o Reg.	27.12	82.97
Ours w/ Max	26.12	83.11
Ours w/ Aver (Full Model)	**19.47**	**90.32**