(Supplementary Material) Text-Adaptive Generative Adversarial Networks: Manipulating Images with Natural Language

Seonghyeon Nam, Yunji Kim, and Seon Joo Kim
Yonsei University
{shnnam,kim_yunji,seonjookim}@yonsei.ac.kr

1 Additional Experimental Results

In this supplementary material, we show additional results of our method. Fig. 1 and Fig. 2 show qualitative comparison on CUB and Oxford-102 datasets, and Fig. 3 and Fig. 4 show additional qualitative results of our method.

2 Network Architecture

Table 1 and 2 show the hyperparameters of the proposed network. For the conditional discriminator, we added additional 3×3 convolution layers to `conv3`, `conv4`, and `conv5` layers in the unconditional discriminator. Then, the features from those layers are spatially reduced by global average pooling and classified by local discriminators. The parameters of each local discriminator is generated from each word vector of the RNN. (Conv2d(K, P): 2D convolution with the kernel size K and the padding P, BN: Batch normalization, LeakyReLU(S): LeakyReLU with the negative slope S, NN Upsampling: Nearest neighbor upsampling)

References

[1] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *ICCV*, 2017.

[2] H. Dong, S. Yu, C. Wu, and Y. Guo, "Semantic image synthesis via adversarial learning," in *ICCV*, Oct 2017.

[3] Q. H. H. Z. Z. G. X. H. X. H. Tao Xu, Pengchuan Zhang, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *CVPR*, 2018.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

Table 1: The parameters of the generator.

Module	Layers	Input size	Output size
Text Encoder	Bidirectional GRU	# of words $\times$ 300	# of words $\times$ 512
	Temporal Averaging	# of words $\times$ 512	512
	Linear, LeakyReLU(0.2)	512	256
(a)	Conditioning Augmentation [1]	256	128
Image Encoder	Conv2d(3, 1), ReLU	3×128×128	64×128×128
	Conv2d(4, 2), BN, ReLU	64×128×128	128×64×64
	Conv2d(4, 2), BN, ReLU	128×64×64	256×32×32
(b)	Conv2d(4, 2), BN, ReLU	256×32×32	512×16×16
Concat (a) and (b)	Conv2d(3, 1), BN, ReLU	640×16×16	512×16×16
Residual Blocks	4 $\times$ Residual Block (below)	512×16×16	512×16×16
Residual Block	Conv2d(3, 1), BN, ReLU	512×16×16	512×16×16
(c)	Conv2d(3, 1), BN	512×16×16	512×16×16
	Input + (c)	512×16×16	512×16×16
Decoder	NN Upsampling (2×)	512×16×16	512×32×32
	Conv2d(3, 1), BN, ReLU	512×32×32	256×32×32
	NN Upsampling (2×)	256×32×32	256×64×64
	Conv2d(3, 1), BN, ReLU	256×64×64	128×64×64
	NN Upsampling (2×)	128×64×64	128×128×128
	Conv2d(3, 1), BN, ReLU	128×128×128	64×128×128
	Conv2d(3, 1), Tanh	64×128×128	3×128×128

Table 2: The parameters of the discriminator.

Module	Layers	Input size	Output size
Image Encoder	Conv2d(4, 2), LeakyReLU(0.2)	3×128×128	64×64×64
	Conv2d(4, 2), BN, LeakyReLU(0.2)	64×64×64	128×32×32
`conv3`	Conv2d(4, 2), BN, LeakyReLU(0.2)	128×32×32	256×16×16
`conv4`	Conv2d(4, 2), BN, LeakyReLU(0.2)	256×16×16	512×8×8
`conv5`	Conv2d(4, 2), BN, LeakyReLU(0.2)	512×8×8	512×4×4
Unconditional Discriminator	Conv2d(4, 0), Softmax	512×4×4	1×1×1
Text Encoder	Bidirectional GRU	# of words $\times$ 300	# of words $\times$ 512
β_{ij}	Linear, Softmax	# of words $\times$ 512	# of words $\times$ 3
α_i	See Eq. (3) in the paper	# of words $\times$ 512	# of words $\times$ 1
$f_{\mathbf{w}_i,j}$	Linear (See Eq. (2) in the paper)	N/A	N/A
From `conv3`	Conv2d(3, 1), BN, LeakyReLU(0.2)	256×16×16	256×16×16
(a)	Global Average Pooling	256×16×16	256×1×1
From `conv4`	Conv2d(3, 1), BN, LeakyReLU(0.2)	512×8×8	512×8×8
(b)	Global Average Pooling	512×8×8	512×1×1
From `conv5`	Conv2d(3, 1), BN, LeakyReLU(0.2)	512×4×4	512×4×4
(c)	Global Average Pooling	512×4×4	512×1×1
Conditional Discriminator	See Eq. (5) in the paper with (α_i, β_{ij}, $f_{\mathbf{w}_i,j}$, (a), (b), (c))	N/A	1×1×1

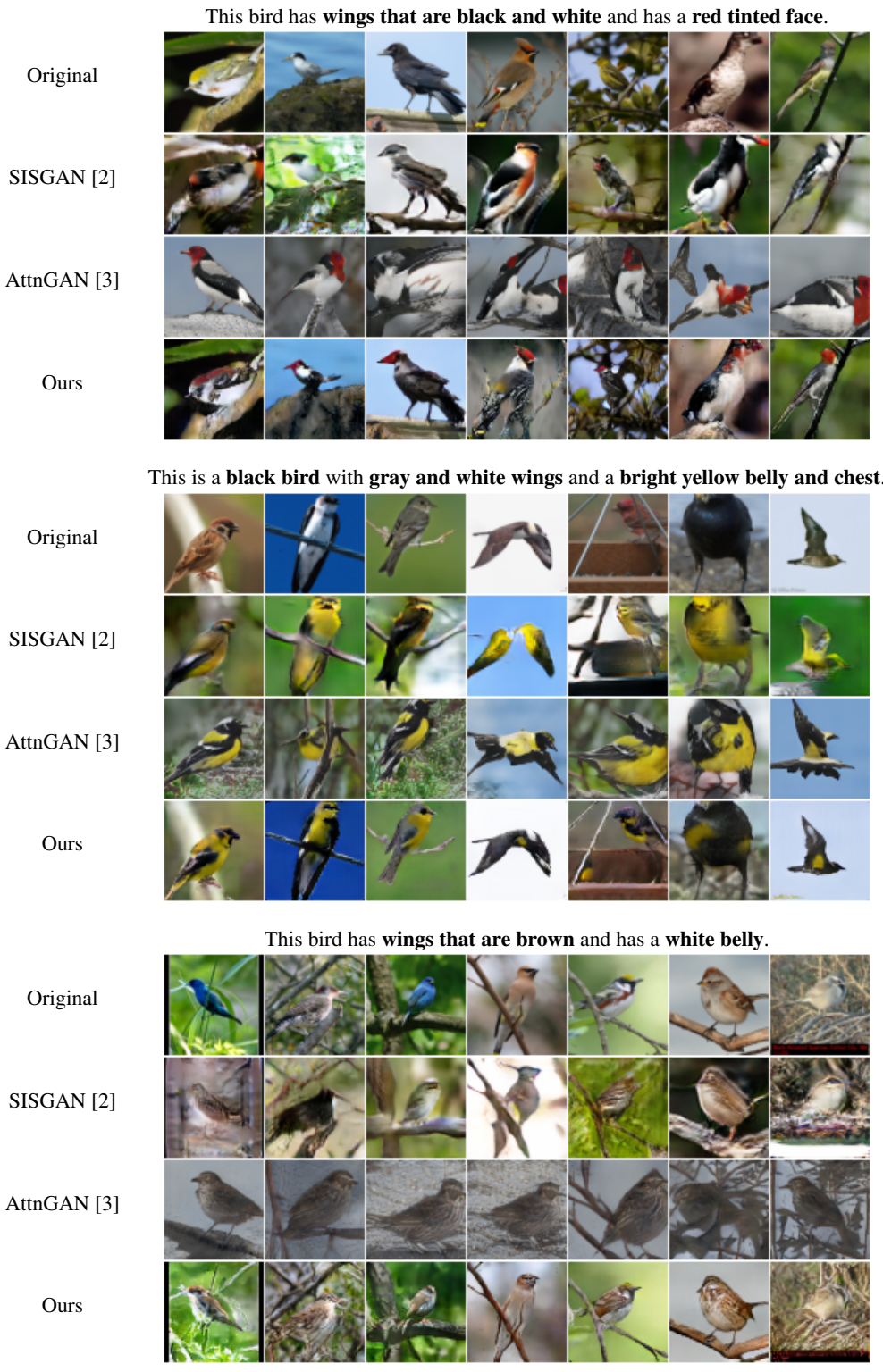

Figure 1: Qualitative comparison on CUB dataset.

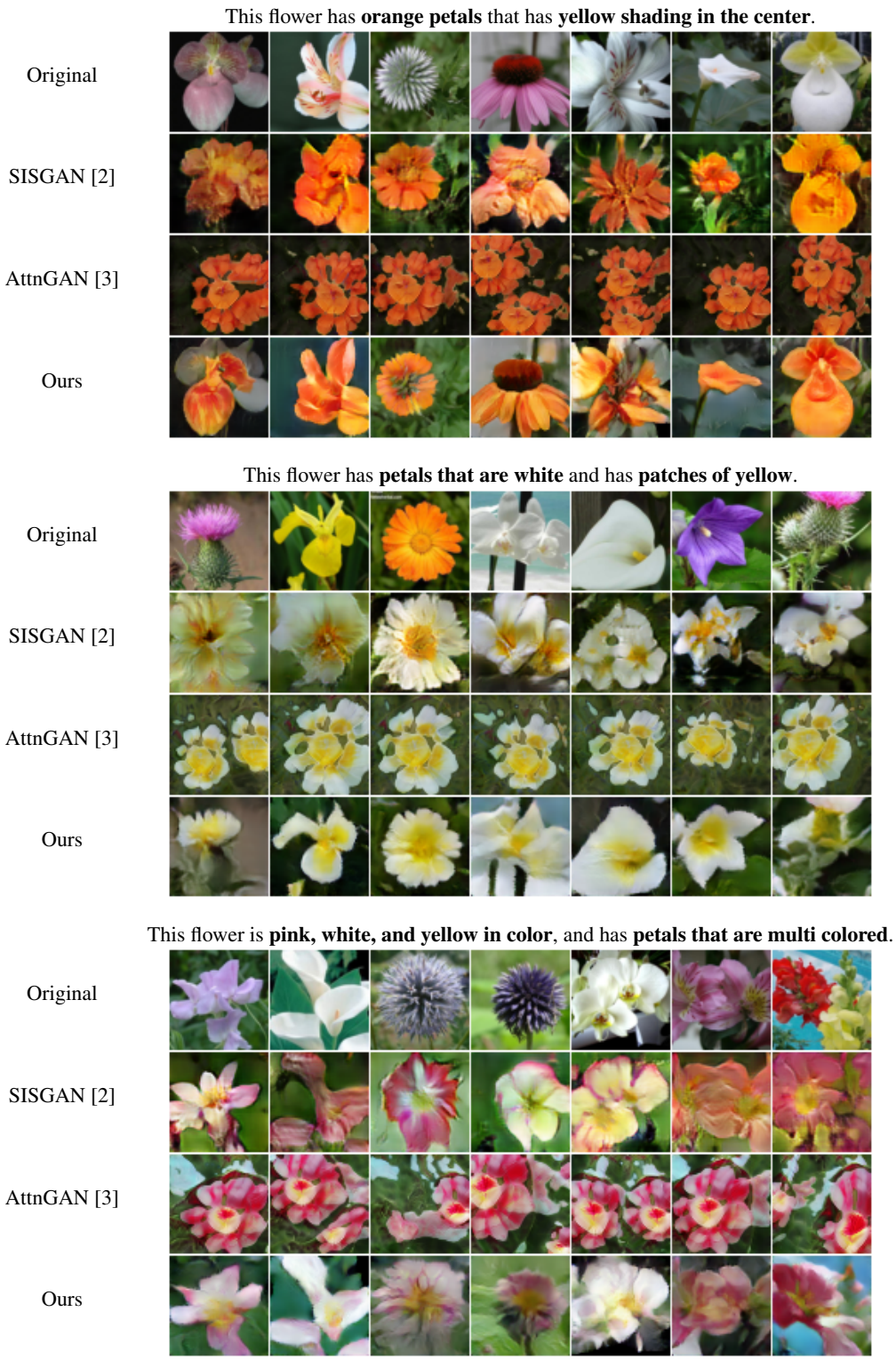

Figure 2: Qualitative comparison on Oxford-102 dataset.

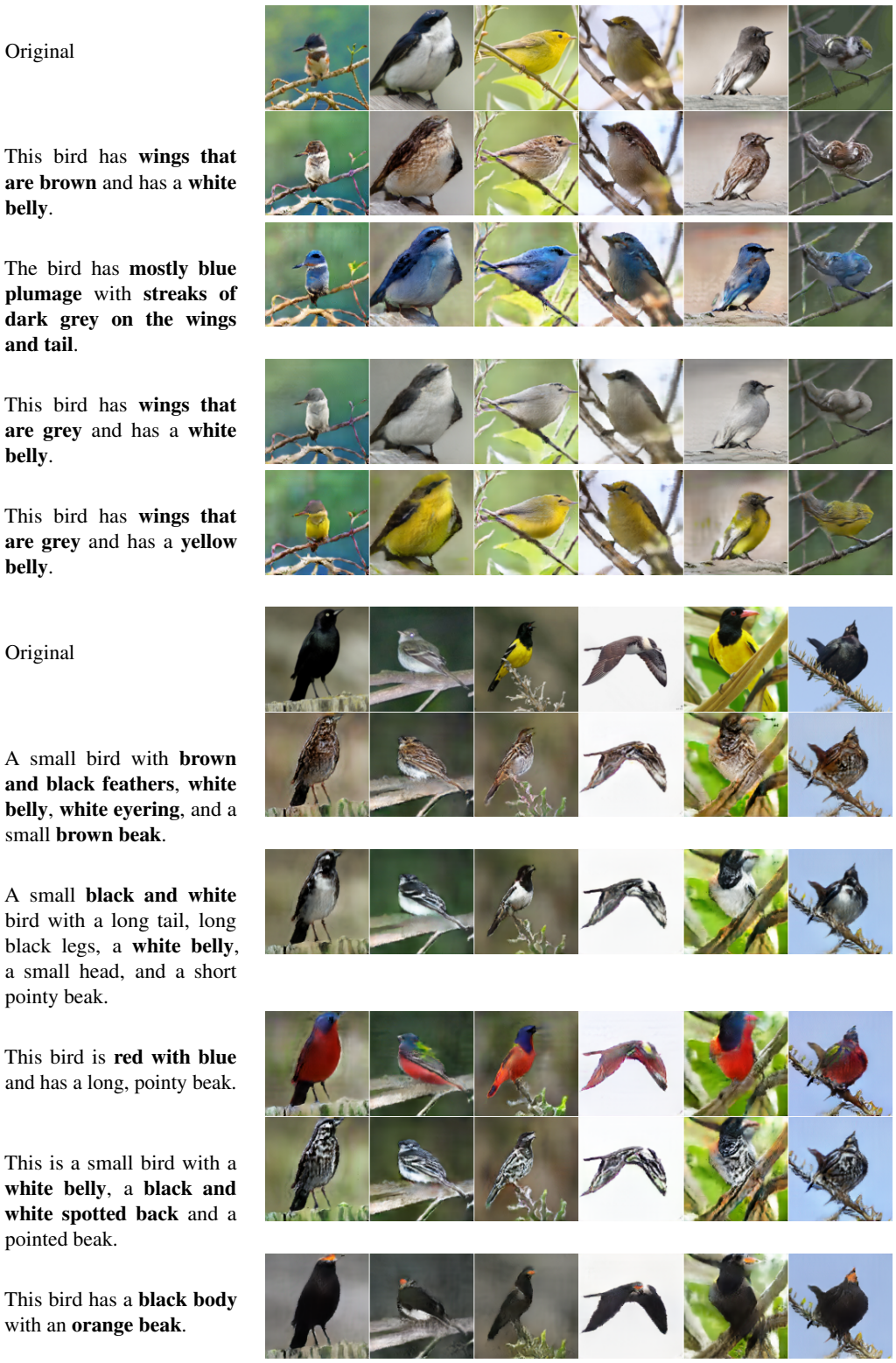

Figure 3: Additional qualitative results of our method on CUB dataset.

Figure 4: Additional qualitative results of our method on Oxford-102 dataset.