TEXT-ADAPTIVE GENERATIVE ADVERSARIAL NETWORKS: MANIPULATIING IMAGES WITH NATURAL LANGUAGE

Seonghyeon Nam et al., NeurIPS, 2019

VISION SEMINAR 2020/04/23

DAVIAN
Data and Visual Analytics Lab

Overview

- **The outputs of paper**
- **Tackling points of the paper**
- **Methods**
- **Experiments**

Original

This flower has **white petals** with a **splash of red coloring** in the middle of each one.

The petals on this flower are **white** with **yellow stamen.**

This flower is **yellow and brown** in color, with petals that are oval shaped.

Tackling Points of the Paper

GOAL: Manipulating an image from a given task description

Position of the paper

(Unconditional) text-to-image generation;

- StackGAN => StackGAN++ => AttnGAN => MirrorGAN

Text conditional image manipulation;

- TAGAN => ManiGAN (ICML 2020 submit)

Segmentation map conditional image manipulation;

- SPADE

Tackling Points of the Paper

Summary

- Most of previous works only focus on generating images from text description without the original image While, a few research addressed a given image manipulation with text description.

- The key idea is to split a single sentence-level discriminator into **a number of word-level discriminators.**

- TAGAN successfully generate a realistic manipulated image, preserving text irrelevant region.

Methods: Overview

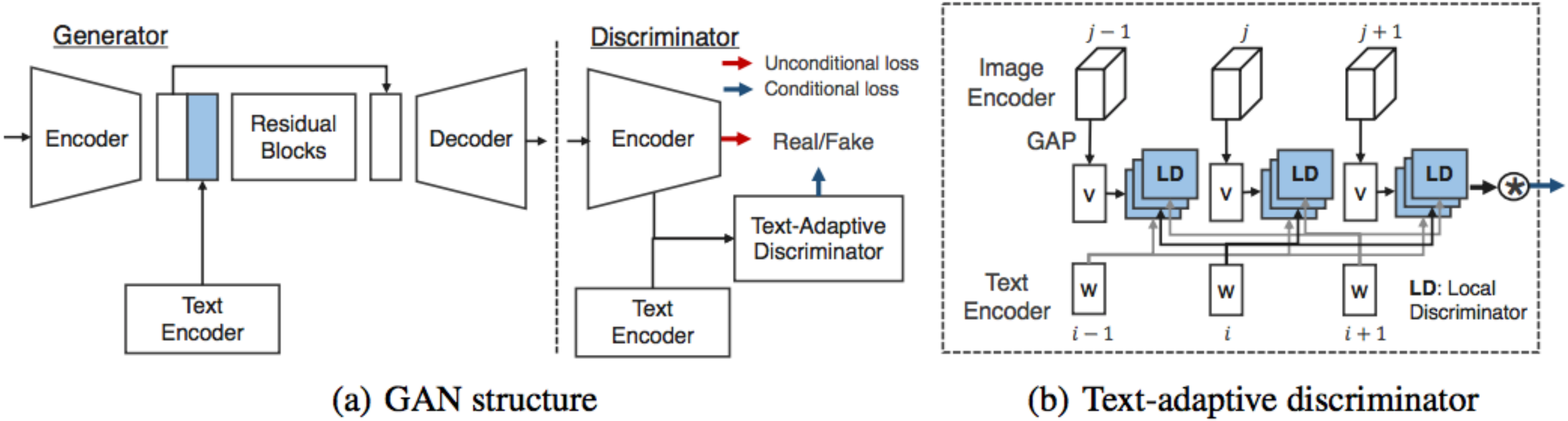


Figure 2: The proposed GAN structure. (a) shows the overall GAN architecture and (b) depicts our text-adaptive discriminator. In (b), the attention and the layer-wise weight are omitted for simplicity.

[INPUT] A text description x $\in R^{3*h*w}$, t: $real\ paired\ text$, $\hat{t}$: $fake\ paired\ text$
[CONCAT ?] A sentence vector $v \in R^{D}$ is broadcasted to fit tensor size.

Methods: Text-adaptive discriminator

- The discriminator classifies each **attribute (word) independently** using word-level local discriminators.
- 1D sigmoid local discriminator f_{w_i}, which determines whether a visual attribute related to w_i exists in the image.

$$f_{\mathbf{w}_i}(\mathbf{v}) = \sigma(\mathbf{W}(\mathbf{w}_i) \cdot \mathbf{v} + \mathbf{b}(\mathbf{w}_i)),$$

- To reduce the impact of less important words to the final score, where u is a temporal average of w_i and

$$\alpha_i = \frac{\exp(\mathbf{u}^T \mathbf{w}_i)}{\sum_i \exp(\mathbf{u}^T \mathbf{w}_i)}, \qquad D(\mathbf{x}, \mathbf{t}) = \prod_{i=1}^{T} [f_{\mathbf{w}_i}(\mathbf{v})]^{\alpha_i}.$$

Methods: Text-adaptive discriminator

- The discriminator classifies each **attribute (word) independently** using word-level local discriminators.
- 1D sigmoid local discriminator f_{w_i}, which determines whether a visual attribute related to w_i exists in the image.

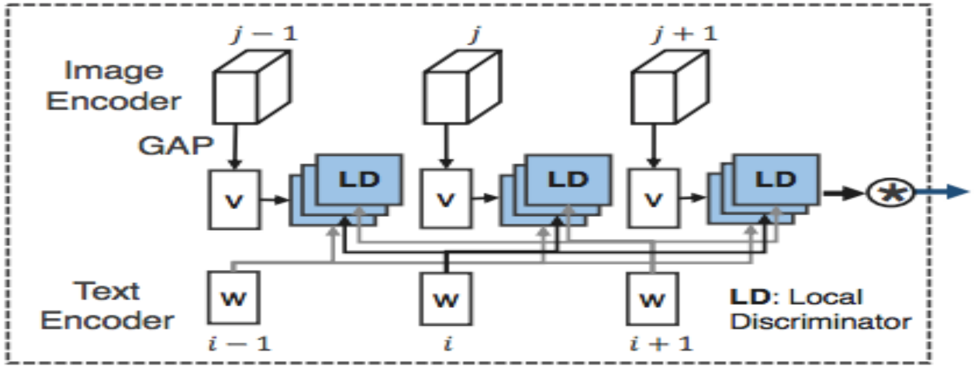
$$f_{\mathbf{w}_i}(\mathbf{v}) = \sigma(\mathbf{W}(\mathbf{w}_i) \cdot \mathbf{v} + \mathbf{b}(\mathbf{w}_i)),$$

- To reduce the impact of less important words to the final score, where u is a temporal average of w_i and

$$\alpha_i = \frac{\exp(\mathbf{u}^T \mathbf{w}_i)}{\sum_i \exp(\mathbf{u}^T \mathbf{w}_i)}, \qquad D(\mathbf{x}, \mathbf{t}) = \prod_{i=1}^{T} [f_{\mathbf{w}_i}(\mathbf{v})]^{\alpha_i}.$$

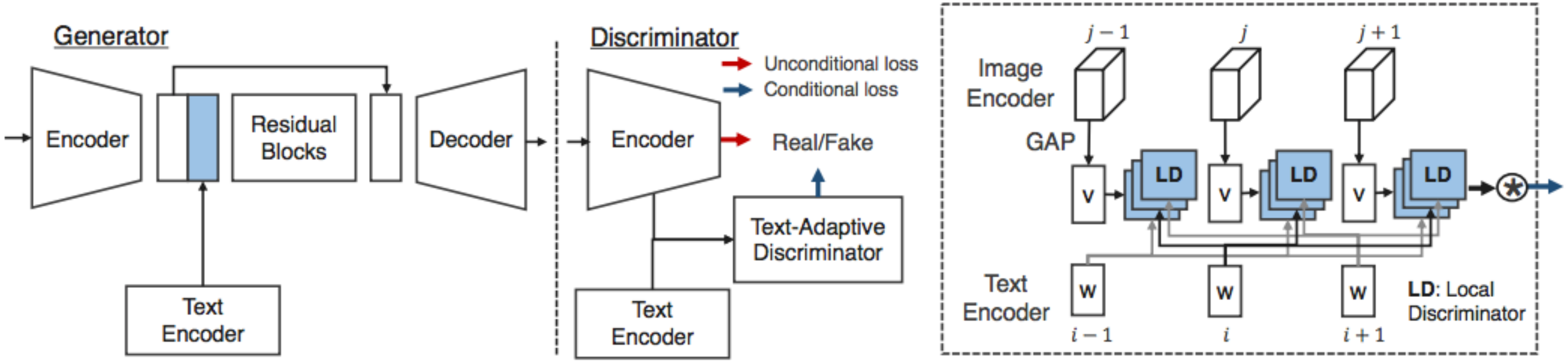
- Considering multi-scale image features, the authors
- enforce word to determine where to concentrate
- (small scale features or large-scale features),

$$D(\mathbf{x}, \mathbf{t}) = \prod_{i=1}^{T} [\sum_j \beta_{ij} f_{\mathbf{w}_i, j}(\mathbf{v}_j)]^{\alpha_i},$$



(b) Text-adaptive discriminator

Methods: Total losses



(a) GAN structure (b) Text-adaptive discriminator

$$L_D = \mathbb{E}_{\mathbf{x},\mathbf{t},\hat{\mathbf{t}}\sim p_{data}}[\log D(\mathbf{x}) + \lambda_1(\log D(\mathbf{x},\mathbf{t}) + \log(1 - D(\mathbf{x},\hat{\mathbf{t}})))]$$
$$+ \mathbb{E}_{\mathbf{x},\hat{\mathbf{t}}\sim p_{data}}[\log(1 - D(G(\mathbf{x},\hat{\mathbf{t}})))],$$

$$L_G = \mathbb{E}_{\mathbf{x},\hat{\mathbf{t}}\sim p_{data}}[\log D(\mathbf{x}) + \lambda_1 \log D(G(\mathbf{x},\hat{\mathbf{t}}),\hat{\mathbf{t}})] + \lambda_2 L_{rec},$$

Experiments: Qualitative results (1/3)

Original

This bird has **wings that are blue** and has a **white belly**.

A small bird with **white base** and **black stripes** throughout its belly, head, and feathers.

Original

The petals of the flower have **yellow and red stripes**.

This flower has petals of **pink and white color** with **yellow stamens**.

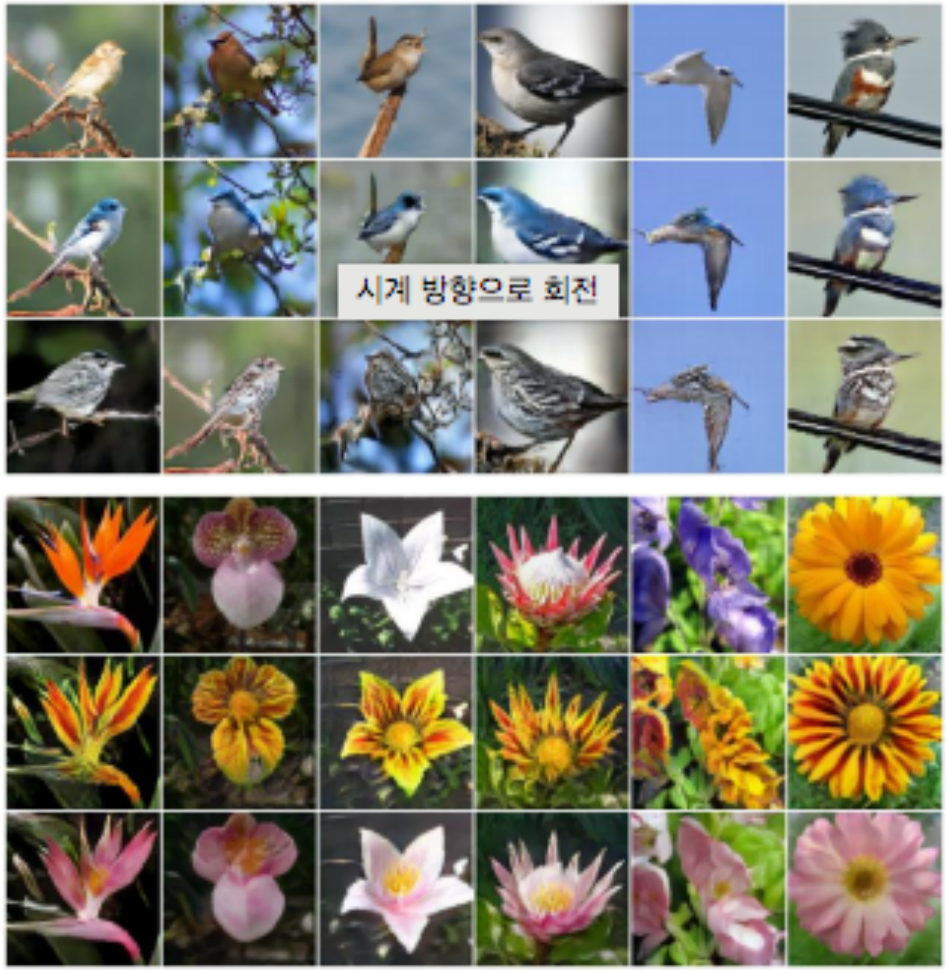


Figure 3: Qualitative results of our method on CUB and Oxford-102 datasets.

Experiments: Qualitative results (2/3)

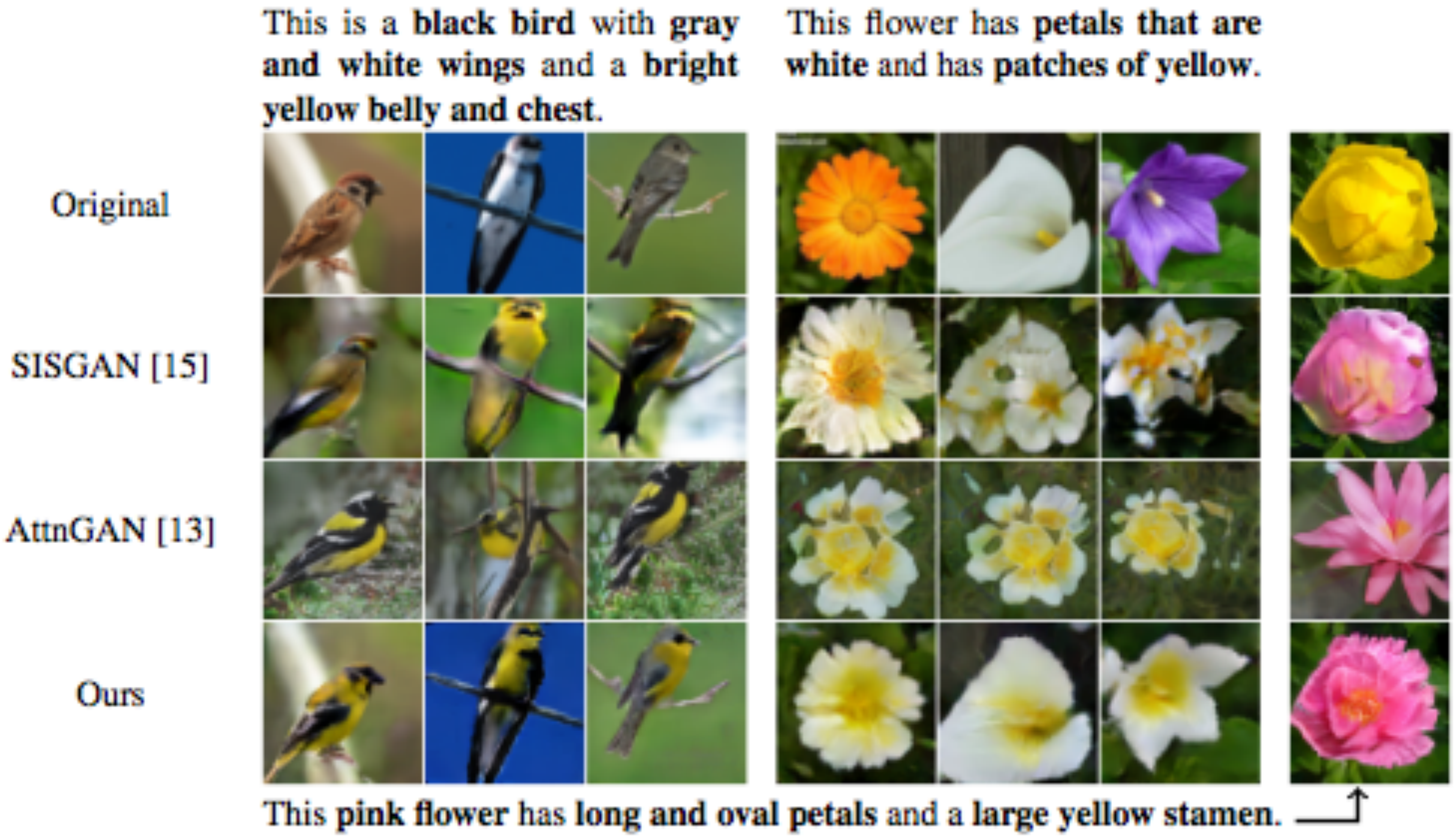


Figure 4: Qualitative comparison of three methods. In most cases, our method outperforms baseline methods qualitatively. The rightmost column shows a failure case using our method.

Experiments: Qualitative results (3/3)

- *CAM results of each word*

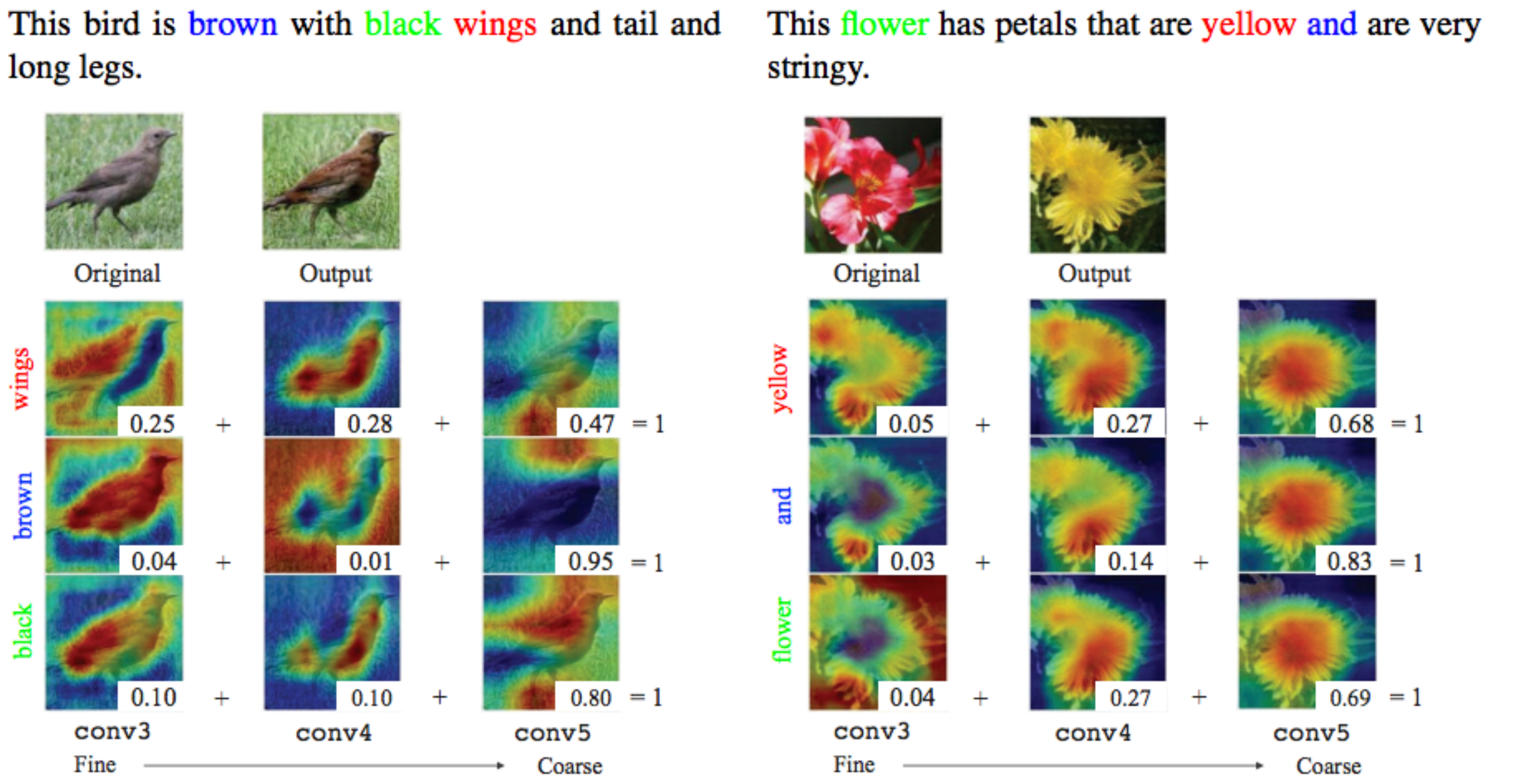


Figure 5: Visualization of the text-adaptive discriminator. From top to bottom, the top-3 word attentions are shown. From left to right, the saliency maps of 3 layer-wise local discriminators are visualized. Each fractional number is β_{ij}. Note that $\sum_j \beta_{ij} = 1$.

Experiments: Quantitative result

Table 1: Quantitative comparison. Accuracy and Naturalness were evaluated by users, and the values indicate the average ranking. L_2 reconstruction error was additionally compared.

	CUB			Oxford-102		
Method	Accuracy	Naturalness	L_2 error	Accuracy	Naturalness	L_2 error
SISGAN [15]	2.33	2.34	0.30	2.67	2.28	0.29
AttnGAN [13]	2.19	2.11	0.25	2.21	2.10	0.32
Ours	**1.49**	**1.56**	**0.11**	**1.52**	**1.62**	**0.11**