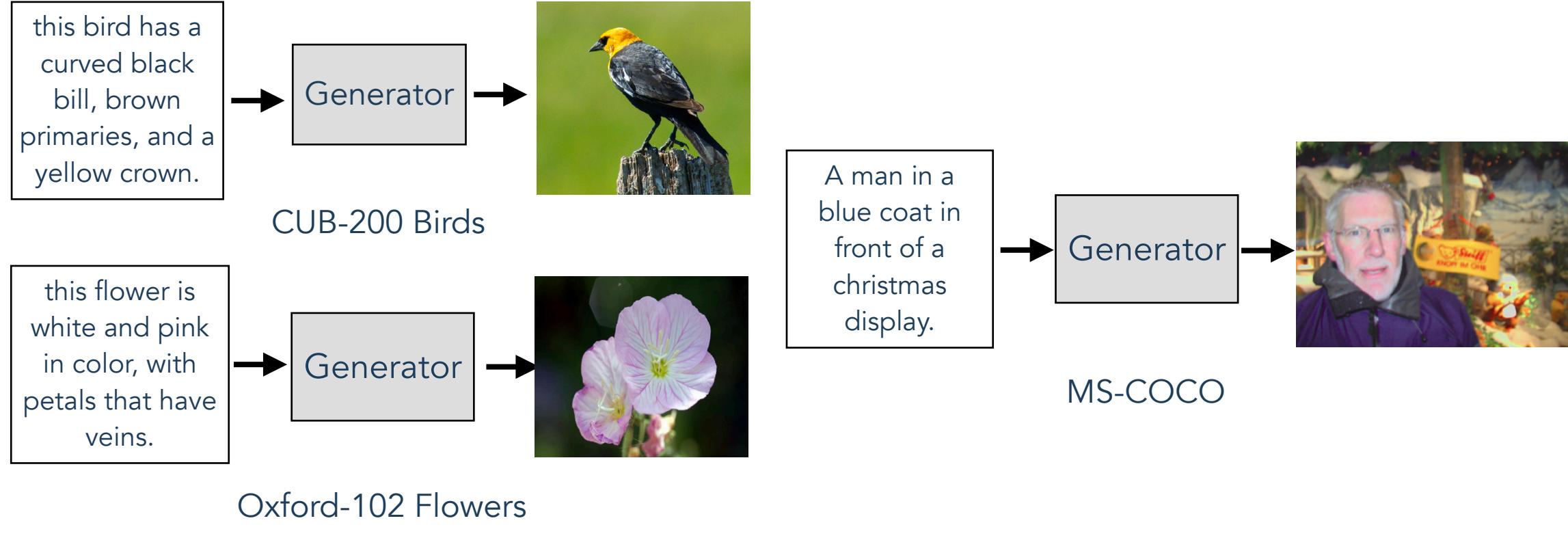




1 – Overview

Text-to-Image Synthesis (*)



Evaluation of Text-to-Image Models

Evaluating the text-to-image models is a **challenging** task!

Caption	DM-GAN	CPGAN	AttnGAN++	Real Images
Several plates of food include fry dough and salad.				
There are people standing on the sand at the beach.				
Inception Score R-Precision SOA-C SOA-I	32.43 92.23 33.44 48.03	52.90 93.59 77.02 84.55	40.13 96.39 48.33 67.19	37.71 67.35 74.97 80.84
	IS = 5.12	IS* = 13.05	IS = 4.78	IS* = 15.13

But do not better quality than real images!

Better metric scores than real images

Contribution: We identify several pitfalls in the existing evaluation pipeline of text-to-image synthesis models, and propose a set of metrics combining the improved versions of existing metrics and new ones to form a unified evaluation toolbox, so-called **TISE**, for benchmarking text-to-images models fairly, robustly, and consistently.

References

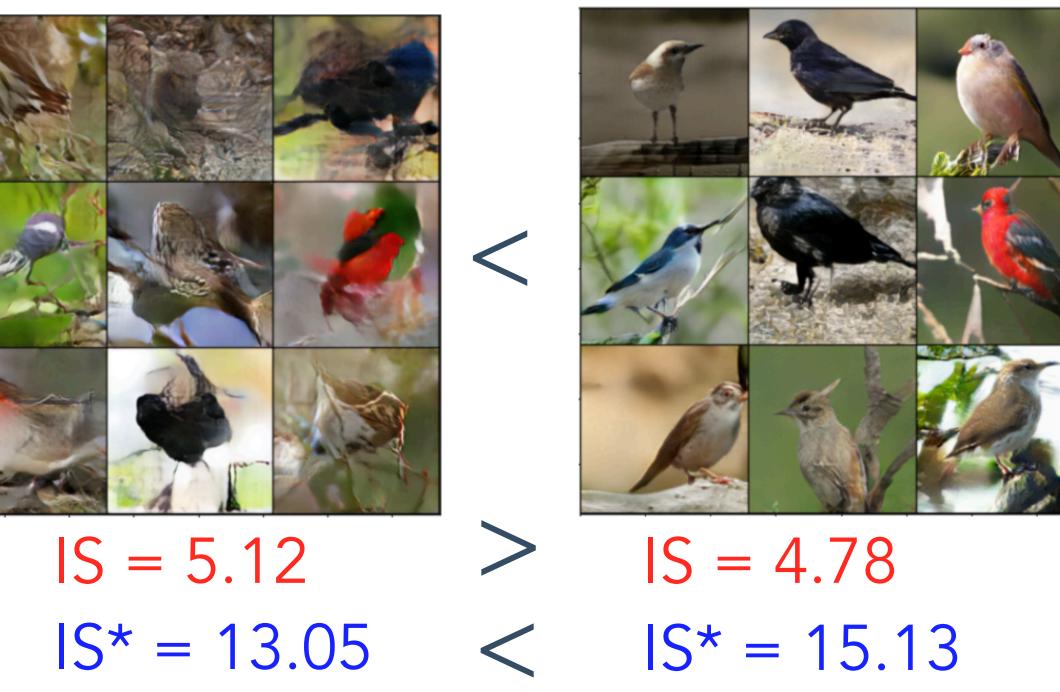
- [1] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: NeurIPS (2016)
 - [2] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
 - [3] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial net- works. In: CVPR (2018)
 - [4] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML (2017)
 - [5] Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks.
 - [6] Hinz, T., Heinrich, S., Wermter, S.: Semantic object accuracy for generative text- to-image synthesis. In: TPAMI 2020
- [*] Source: Example captions and images are randomly taken from each dataset.

2 – Single-Object Text-to-Image Synthesis

Evaluation aspects

- ◆ Image Realism
 - Inception Score (IS) [1]
 - Fréchet Inception Distance (FID) [2]
- ◆ Text Relevance
 - R-precision (RP) [3]

Inconsistent issue of IS and our IS*



Benchmark Results

Method	IS (\uparrow)	IS* (\uparrow)	FID (\downarrow)	RP (\uparrow)
GAN-INT-CLS	2.73	7.51	194.41	3.83
StackGAN++	4.10	12.69	27.40	13.57
AttnGAN	4.32	13.63	24.27	65.30
AttnGAN + CL	4.45	14.42	17.96	60.82
DM-GAN	4.68	15.00	15.52	76.25
DF-GAN	4.77	14.70	16.46	42.95
DM-GAN + CL	4.77	15.08	14.57	69.80
AttnGAN++ (ours)	4.78	15.13	15.01	77.31

Our AttnGAN++ is a strong baseline!

AttnGAN++ = AttnGAN + Spectral Normalization [5]
+ Tune with different hyperparameters

Inception score (IS)

Estimated by the pre-trained Inception-v3 classifier

$$IS = \exp(\mathbb{E}_x D_{KL}(p(y|x) \| p(y)))$$

where x is the generated image; y is the class label

Consideration:

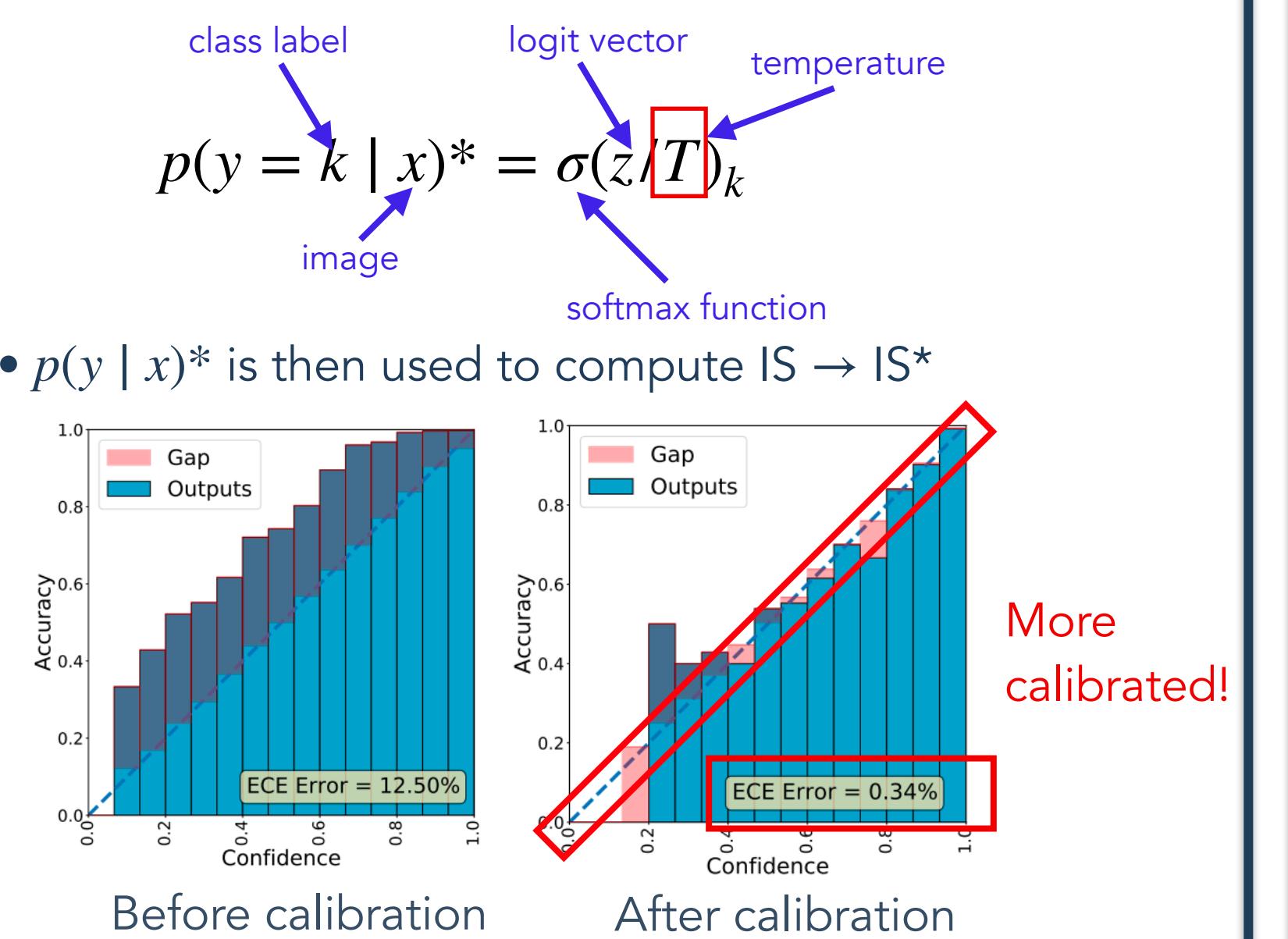
- The objects should be **distinct recognizable** (1)
→ $p(y|x)$ must have low entropy
- Having **diverse** number of object classes (2)
→ $p(y)$ must have high entropy

(1) + (2) → KL-divergence between $p(y)$ and $p(y|x)$ should be large

→ Higher value of IS is better

Improved IS (IS*): Calibrating Image Classifier

- Using **temperature scaling** [4] to calibrate the classifier



3 – Multiple-Object Text-to-Image Synthesis

Demanding evaluation aspects

- Image Realism
- Text Relevance
- Object Accuracy
- Object Fidelity
- Positional Alignment
- Counting Alignment
- Paraphrase Robustness
- Explainable

Existing metrics: Improved Version

Current problems: overfitting

- **RP** [3]: Training T2I models and computing RP used the same text-encoder module (DAMSM)
- **SOA** [6]: CPGAN and SOA used the same pre-trained YOLO-v3

Solution: Replace them by other independent, better models

- **RP** [3]: DAMSM → CLIP
- **SOA** [6]: YOLO-v3 → Mask-RCNN

Object Fidelity

- Crop the object set from generated images
 - Use off-the-shelf object detector
- Compute IS* on the cropped objects → **O-IS**
- Compute FID on the cropped objects → **O-FID**

Counting Alignment

- **Constructing evaluation data**
 - Define a set of **positional** words, called W
 - $D_w = \{(P_{wi}, Q_{wi})\}_{i=1}^{N_w}\}$ for each $w \in W$
- **Positional Alignment (PA) metric**
 - Define P_{wi} → (i) replace w with its antonym; (ii) keeping other words → Q_{wi}
 - P_{wi} : A man is **in front of** the blue car
 - Q_{wi} : A man is **behind** the blue car
- **Counting Alignment (CA) metric**
 - Use T2I model to generate the image from caption
 - Count the number of objects in generated image
 - Use an off-the-shelf object counter
 - Measure **error(GT count, predicted count)**

$$CA = \frac{1}{|D|} \sum_{i=1}^{|D|} \sqrt{\frac{1}{N_{ic}} \sum_{j=1}^{N_{ic}} (\hat{c}_{ij} - c_{ij})^2}$$

c_{ij} , \hat{c}_{ij} : GT and predicted count in image i for object class j
 N_{ic} : number of GT object types in image i

Benchmark Results

Method	Image Realism	Text Relevance	Object Accuracy	Object Fidelity	Counting Alignment	Positional Alignment
GAN-CLS	1.0	2.0	1.0	1.0	1.0	1.0
StackGAN	2.5	1.0	2.0	2.0	2.0	2.0
AttnGAN	5.0	5.0	5.5	4.5	6.0	3.0
DM-GAN	6.5	7.0	7.0	7.5	8.0	5.0
CPGAN	7.5	8.0	10.0	7.5	4.0	6.0
DF-GAN	7.0	3.0	4.0	8.5	5.0	4.0
AttnGAN + CL	6.5	6.0	5.5	5.0	7.0	7.0
DM-GAN + CL	8.5	9.0	8.0	7.0	9.0	10.0
DALLE-mini (zero-shot)	2.5	4.0	3.0	3.0	3.0	8.0
AttnGAN++ (Ours)	9.0	10.0	9.0	9.0	10.0	9.0
Real Images	10.0	11.0	11.0	11.0	11.0	11.0

Ranking scores for each evaluation aspect

Our TISE toolbox is **available!**

<https://github.com/VinAIResearch/tise-toolbox>

✓ We provide the comprehensive benchmarks for T2I synthesis models for both single- and multi- object case.

✓ We propose the improved version of existing metrics + new ones to evaluate many vital evaluation aspects.

✓ We release a Python assessment toolbox called **TISE** to advocate fair comparisons and reproducible results for future T2I synthesis research.