# TEXT TO IMAGE SYNTHESIS WITH BIDIRECTIONAL GENERATIVE ADVERSARIAL NETWORK

*Zixu Wang*[1], *Zhe Quan*[1], *Zhi-Jie Wang*[2,3], *Xinjian Hu*[1], *Yangyang Chen*[1]

[1]College of Information Science and Engineering, Hunan University, Changsha, China
[2]College of Computer Science, Chongqing University, Chongqing, China
[3]School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China
{wangzixu, quanzhe, huxinjian, cyy1216}@hnu.edu.cn, cs.zjwang@gmail.com

## ABSTRACT

Generating realistic images from text descriptions is a challenging problem in computer vision. Although previous works have shown remarkable progress, guaranteeing semantic consistency between text descriptions and images remains challenging. To generate semantically consistent images, we propose two semantics-enhanced modules and a novel Textual-Visual Bidirectional Generative Adversarial Network (TVBi-GAN). Specifically, this paper proposes a semantics-enhanced attention module and a semantics-enhanced batch normalization module. These modules improve consistency of synthesized images by involving precisely semantic features. What's more, an encoder network is proposed to extract semantic features from images. During the adversarial process, the encoder could guide our generator to explore corresponding features behind descriptions. With extensive experiments on CUB and COCO datasets, we demonstrate that our TVBi-GAN outperforms state-of-the-art methods.

***Index Terms***— Text-to-Image Synthesis

## 1. INTRODUCTION

Text-to-image generation, which synthesizes images from text descriptions, has become an active research area. Although previous works have made impressive results based on Generative Adversarial Networks (GANs) [4], the uncertainty of natural language [7] makes text-to-image task theoretically an ill-posed problem. Fully understanding the relation between vision and language still has a long way to go.

Leveraging the power of GANs, existing methods make progress on fine-grained image generation by stacking several generators [25, 26, 27], imposing attention guided refinement modules [23, 28, 16, 8] and proposing auxiliary architectures [24, 13]. However, the gap between natural language descriptions and visual contents makes semantic consistency hard to establish. Natural language is ambiguous, so it's nearly impossible to extract precisely semantic features (e.g., texture, colour and shape) from a brief description. In order to overcome this drawback, we explore the possibility of utilizing semantic features behind visual contents to help text-to-image synthesis and propose a novel Textual-Visual Bidirectional Generative Adversarial Network (TVBi-GAN). Our model includes an encoder which maps images to semantic feature space (Figure 1). During the adversarial process, the encoder guides our generator to extract precisely semantic features from descriptions. If we make an analogy between how the generator synthesizes images and how humans imagine images from text descriptions, we could regard the encoding process as how humans comprehend images.

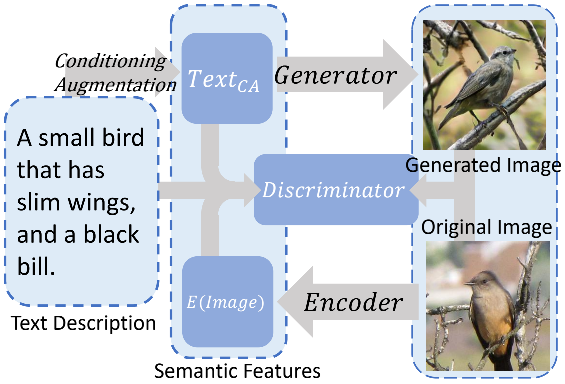


**Fig. 1**. Overview of the TVBi-GAN. First row shows our generation process which bases on a text description. The following row illustrates how we extract semantic features from the original image.

Previous works [13, 23, 24, 8] have shown that text-to-image synthesis could benefit from combining visual contents with sentence or word features. Different from previous works, we pay close attention to precisely semantic features. Specifically, we reformulate attention module from AttnGAN [23], dubbed as Semantics-Enhanced Attention (SEAttn). SEAttn merges semantic features into an adaptive layer, which not only takes in semantic cues but also recalculates the importance of every specific word. After that, we draw inspiration from SD-GAN [24] and propose Semantics-Enhanced Batch Normalization (SEBN). SEBN balances semantic consistency and individual diversity by shifting visual contents to a proper direction. SEAttn and SEBN enable our TVBi-GAN manipulate visual contents toward fine-grained generation and significantly improve the performance. We also note that even without image encoding process, SEAttn and SEBN could adaptively extract valuable semantic cues and still perform better than original modules.

To summarize, we make following contributions. (i) We propose a novel GAN model with a bidirectional architecture to combine text descriptions with visual contents. (ii) We reformulate two modules which could adaptively extract detailed semantic cues behind text descriptions and preserve the consistency of generation process. (iii) The experimental results demonstrate that our TVBi-GAN outperforms state-of-the-art methods.

## 2. RELATED WORK

### 2.1. Image Generation Model

Image generation is a fundamental problem in computer vision. Recently, three classes of algorithms have attracted much attention: Variational Autoencoders [6], GANs [4] and autoregressive approaches [17]. Particularly, GANs have become main stream generation models because of its excellent performance. For instance, Brock *et al.* [1] successfully generates impressive images on a large scale using GANs. Moreover, GANs also make great achivements in many other tasks like saliency detection [19, 20, 10, 21] and information recommendation [11].

### 2.2. Bidirectional Generative Adversarial Networks (BiGANs).

BiGANs [3, 2] are proposed as an extension of GANs, which augment standard GANs with an encoder network mapping real data to latent features. It shares many theoretical properties of GANs [4] and proves to be effective in unsupervised learning. For example, Donhahue and Simonyan [2] demonstrate that BiGANs could achieve state-of-the-art performance on unsupervised representation learning. In our work, BiGANs are reconstructed. Specifically, we extend the definition of latent feature space in [3] and project sentence features into this space. As a result, TVBi-GAN enhances the semantic connection between vision and natural language, and improves the performance of generation.

## 3. TVBI-GAN FOR TEXT-TO-IMAGE GENERATION

### 3.1. Image Generation Process

**Text Encoder.** As shown in Figure 2, we firstly employ a pre-trained text encoder developed by Xu *et al.* [23].

$$w, s = RNN(Text) \tag{1}$$

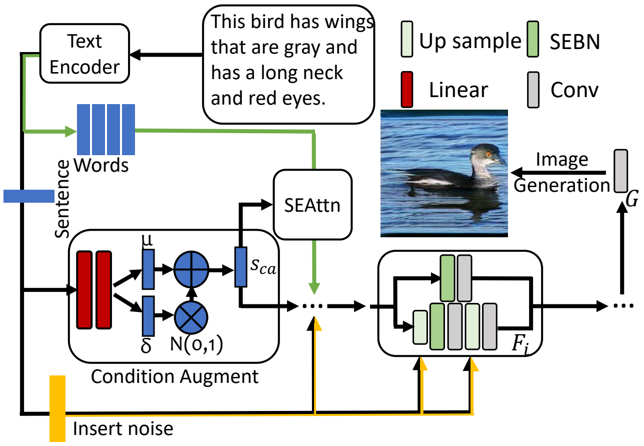


**Fig. 2**. The architecture of the Generator in TVBi-GAN.

where $w = \{w_i | 0 \leq i \leq L-1\} \in R^{D \times L}$ indicates word features, L represents the length of current text description, $s \in R^D$ means current sentence features and $D$ is the dimension of $w_i$ and $s$. Then, we take sentence features into conditioning augmenting function, and resample sentence features from an independent Gaussian distribution $\mathcal{N}(\mu(s), \sigma(s))$. We use $CA$ to represent this process:

$$s_{ca} = CA(s) \tag{2}$$

where $s_{ca} \in R^{D^{'}}$ is semantic features extracted from descriptions, the superscript $D^{'}$ is the dimension of semantic features. $CA$ [25] is used to smooth semantic data manifold in previous works. However, $CA$ plays a greater role in our work and provides discriminator semantic features of current generated images.

**Hierarchical Generative Adversarial Network.** Following previous works [27, 25, 26], we put a multi-stage cascaded generator from low-resolution to high-resolution image generation. Specifically, we use $F_0, F_1$ and $F_2$ to denote visual content transformers, and $G_0, G_1$ and $G_2$ to denote image generators from coarse to fine. This way, each stage is expressed as:

$$f_0 = F_0(z, s_{ca}) \tag{3}$$

$$f_i = F_i(f_{i-1}, F_{A_i}(f_{i-1}, w, s_{ca}), z, s_{ca}, s), i \in 1, 2 \tag{4}$$

$$Image_i = G_i(f_i), i \in 0, 1, 2 \tag{5}$$

where $z \sim \mathcal{N}(0, 1)$ denotes random noises and $F_A$ is our attention module. We cut random noises into several pieces and integrate noises with each generation block. Futhermore, we stack the basic residual block from [5], due to its outstanding performance. After generators output images, they are combined with their $s_{ca}$ and $s$ to discriminators. Note that the discriminators after each generator are independent from each other and accept different resolution images.

**Semantics-Enhanced Attention (SEAttn).** We propose a semantics-enhenced attention module to improve generation consistency. Although attention proposed by [23] takes in the word-level attention and synthesizes fine-grained visual details, it is insufficient to count different words to a same level. Inspired by the gate mechanism [28], we compare semantic features with each word features. SEAttn calculates the importance between word features and semantic features before attention. Our gate mechanism could be formulated as follow:

$$Imp_i(s_{ca}, w_i) = \sigma(W_{imp} * concat(s_{ca}, w_i)) \tag{6}$$

where $\sigma$ represents sigmoid function and $w_{imp}$ is a $1 \times (D + D^{'})$ matrix. After that, we refine specific word features:

$$w_i^{'} = Imp_i * M_w(w_i) + (1 - Imp_i) * M_s(s_{ca}) \tag{7}$$

where $M_w(\cdot)$ and $M_s(\cdot)$ denote $1 \times 1$ convolutional operation which map $s_{ca}$ and $w_i$ into the same dimension feature space. After the above gate mechanism, we replace $w_i^{'}$ for $w_i$ as specific word features and utilize the attention layer proposed by [23] to synthesize fine-grained visual details.

**Semantics-Enhanced Batch Normalization (SEBN).** As indicated in SD-GAN [24], modulating the scale-and-shift operation with sentence cues could improve diversity of synthesized images. However, linguistic descriptions are subjective and usually bring inessential features. In order to prevent visual contents shifting to unproper direction too much and to sustain consistency of generation, we integrate semantic features into SEBN. Particularly, we incorporate random noises $z$ into SEBN, which slightly improves the performance of GANs [1]. SEBN is formulated as:

$$\gamma_c = f_\gamma(concat(s, s_{ca}, z)), \beta_c = f_\beta(concat(s, s_{ca}, z)) \tag{8}$$

$$BN(x, s, s_{ca}, z) = \gamma_c \cdot \frac{x - \mu(x)}{\sigma(x)} + \beta_c \tag{9}$$

where $f_\gamma$ and $f_\beta$ are projection layers.

### 3.2. Encoder

Encoding images to semantic feature space is an essential part of TVBi-GAN. We use a deep convolutional network to inverse the image generation process:

$$f_i = F_i(f_{i-1}), i \in 1, 2, ..., t \tag{10}$$

where $F_i$ denotes a residual block which is similar to the generator. But we erase the conditional projection layer in Batch Normalization and resample the outputs of encoder, like the text encoder.

### 3.3. Objective Function

The purpose of TVBi-GAN is to extract precisely semantic features for text-to-image generation. We propose two kinds of adversarial loss: semantic feature loss and conditional semantic feature loss. These functions induce the learnt joint distributions to match at the global optinum. As a result, our generator could learn how to extract semantic features from descriptions. We also employ two adversarial losses to approximate conditional and unconditional distributions. During each stage of training, generator, encoder and discriminator are trained separately. Specifically, in the $i^{th}$ stage, we minimize loss function as follows:

$$\begin{aligned} L_{D_i} = & \underbrace{\mathbb{E}_{x\sim p_{data}} h_1(D_i(x)) + \mathbb{E}_{x\sim p_{G_i}} h_2(D_i(x))}_{unconditional\ loss} \\ & + \underbrace{\mathbb{E}_{x\sim p_{data}} h_1(D_i(x, s)) + \mathbb{E}_{x\sim p_{G_i}} h_2(D_i(x, s))}_{conditional\ loss} \\ & + \underbrace{\mathbb{E}_{x\sim p_{data}} h_1(D_i(x, E(x))) + \mathbb{E}_{x\sim p_{G_i}} h_2(D_i(x, s_{ca}))}_{conditional\ semantic\ feature\ loss} \\ & + \underbrace{\mathbb{E}_{x\sim p_{data}} h_1(D_i(E(x))) + \mathbb{E} h_2(D_i(s_{ca}))}_{semantic\ feature\ loss} \end{aligned} \tag{11}$$

where $h_1(t) = max(0, 1 - t)$ and $h_2(t) = max(0, 1 + t)$ are ”hinge” losses, $x$ denotes images from the image data distribution or from the generated distribution. We use hinge loss here to improve stability and prevent our model from gradient vanish. For the loss function of generator, we add conditioning augmentation loss and DAMSM loss [25, 23], because of their excellent performance towards text-to-image synthesis.

## 4. EXPERIMENT

### 4.1. Experiment Settings

**Datasets.** TVBi-GAN is evaluated on CUB bird dataset [18] and COCO dataset [9], following previous text-to-image synthesis methods [25, 23].

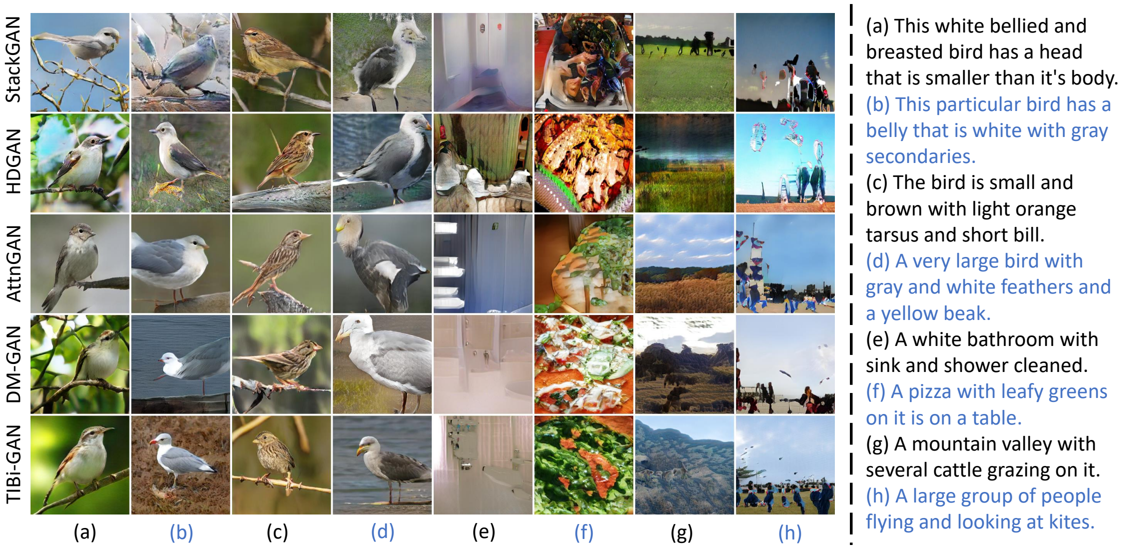
**Evaluation metrics.** How to evaluate the performance of generative models is still a hard problem. We follow previous works [25, 23, 28] on this task. We use Inception Score (IS) and Fréchet Inception Distance (FID) [22] for quantitatively evaluation. Besides that, we also conduct Human Perceptual Test to evaluate whether the generated images are conditioned on the text descriptions.

### 4.2. Main Results.

#### *4.2.1. Quantitative results*

To evaluate our TVBi-GAN, we compare our results with state-of-the-art methods [25, 26, 14, 23, 24, 27, 28, 13, 12, 15, 8, 16].

As shown in Table 1, our TVBi-GAN achieves 5.03 IS on CUB dataset and 31.01 IS on COCO dataset. Although TVBi-GAN performs worse than SD-GAN on COCO dataset, SD-GAN [24] has a serious defect. Specifically, SD-GAN makes use of the siamese network to extract semantic commons from a pair of descriptions, which highly relies on the



(a) This white bellied and breasted bird has a head that is smaller than it's body.
(b) This particular bird has a belly that is white with gray secondaries.
(c) The bird is small and brown with light orange tarsus and short bill.
(d) A very large bird with gray and white feathers and a yellow beak.
(e) A white bathroom with sink and shower cleaned.
(f) A pizza with leafy greens on it is on a table.
(g) A mountain valley with several cattle grazing on it.
(h) A large group of people flying and looking at kites.

**Fig. 3**. Random generated examples by TVBi-GAN, StackGAN [25], HDGAN [27], AttnGAN [23] and DM-GAN [28] on CUB (four left columns) and COCO (four right columns) test sets.

| Methods | CUB | COCO |
|---|---|---|
| GAN-INT-CLS [14] | $2.88 \pm .04$ | $7.88 \pm .07$ |
| GAWWN [15] | $3.62 \pm .07$ | - |
| StackGAN [25] | $3.70 \pm .04$ | $8.45 \pm .03$ |
| StachGAN++ [26] | $4.04 \pm .05$ | - |
| PPGN [12] | - | $9.58 \pm .21$ |
| HDGAN [27] | $4.15 \pm .05$ | $11.86 \pm .18$ |
| AttnGAN [23] | $4.36 \pm .03$ | $25.89 \pm .47$ |
| MirrorGAN [13] | $4.56 \pm .05$ | $26.47 \pm .41$ |
| ControlGAN [8] | $4.58 \pm .09$ | $24.06 \pm .60$ |
| SEGAN [16] | $4.67 \pm .04$ | $27.86 \pm .31$ |
| DM-GAN [28] | $4.75 \pm .07$ | $30.49 \pm .57$ |
| SD-GAN [24] | $4.67 \pm .09$ | $35.69 \pm .50$ |
| TVBi-GAN (backbone network) | $4.95 \pm .06$ | $31.33 \pm .41$ |
| TVBi-GAN | $5.03 \pm .03$ | $31.01 \pm .34$ |

**Table 1**. The performance of IS for TVBi-GAN comparing with others on CUB and COCO test sets. For IS, higher means better. Red, blue and green are corresponding to first, second and third top result. Backbone network doesn't have an encoder network proposed in Section 3.2.

| Methods | CUB | COCO |
|---|---|---|
| AttnGAN [23] | 23.98 | 35.49 |
| DM-GAN [28] | 16.09 | 32.64 |
| TVBi-GAN (backbone network) | 12.78 | 32.50 |
| TVBi-GAN | **11.83** | **31.97** |

**Table 2**. The FID scores of the TVBi-GAN, AttnGAN [23] and DM-GAN [28] on CUB and COCO test sets. The bold results are the best. For FID, lower means better.

diversity of text descriptions. As a result, it's impossible for SD-GAN to widely applicate. Interestingly, we find that the backbone network, which doesn't include the encoder network, performs slightly better than TVBi-GAN on COCO. A possible reason is that IS mainly focuses on the major sets of data distribution and doesn't give enough punishment on the mismatching overall distribution. On the other hand, previous work [22] has investigated that FID is a better measurement compared with IS in terms of discriminability and roubustness. In our experiments, we also test FID scores (table 2) and find that TVBi-GAN achieves better FID than the backbone network. These results imply that our encoder network could help generator distinguish different kinds of parts and help generator synthesize images corresponding to the groud truth image distribution. Moreover, Table 2 also compares FID scores of TVBi-GAN with AttnGAN and DM-GAN on CUB and COCO datasets. Our TVBi-GAN decreases the FID from 16.09 to 11.83 on CUB dataset and from 32.64 to 31.97 on COCO dataset. Our TVBi-GAN achieves excellent performance on both datasets, which means the efficiency of our method and indicates TVBi-GAN could not only generate divers images but also high-quality images.

### 4.2.2. *Qualitative results*

**Human Perceptual Test.** Although existing measurements of GANs have revealed its correspondence to humam perception, it's still hard to evaluat the semantic consistency between images and text descriptions. In order to thoroughly compare our method with others, we extend our experiment to Human Perceptual Test. We invite 40 volunteers to conduct our test

| Methods | CUB | COCO |
|---|---|---|
| AttnGAN [23] | 5.0% | 10.4% |
| DM-GAN [28] | 21.9% | 36.1% |
| TVBi-GAN | **73.1%** | **54.5%** |

**Table 3**. The results of Human Perceptual Test (Ratio of first by volunteers' ranking).

| ID | Components | | | | IS↑ | |
|---|---|---|---|---|---|---|
| | Attn | SCBN | SEAttn | SEBN | CUB | COCO |
| 1 | √ | √ | - | - | $4.66 \pm .03$ | $26.55 \pm .26$ |
| 2 | - | √ | √ | - | $4.81 \pm .05$ | $28.26 \pm .37$ |
| 3 | √ | - | - | √ | $4.73 \pm .02$ | $27.98 \pm .39$ |
| 4 | - | - | √ | √ | $\mathbf{4.95 \pm .06}$ | $\mathbf{31.33 \pm .41}$ |

**Table 4**. Ablation study of SEAttn and SEBN.

and our aim is to find out whether our TVBi-GAN could generate veritable images based on corresponding text descriptions. Each participant was presented 100 groups of images, 50 from CUB dataset and 50 from COCO dataset. Images are random generated from each dataset. In each group, volunteers are given at most 90 seconds to tell the best image according to the corresponding text description.

As shown in Table 3, we compare TVBi-GAN with AttnGAN [23] and DM-GAN [28]. TVBi-GAN shows better semantic consistency than others, and these results are similar to our improvement on IS and FID. In addition, Figure 3 shows that TVBi-GAN generates vivid parts according to text descriptions. However, previous methods suffer from varying degrees of deformation (b, d, e) and semantic inconsistency (f). These results demonstrate the superiority of TVBi-GAN on synthesizing semantic consistent images.

### 4.3. Component Analysis

**Ablation study.** We evaluate the performance of each proposed module. As shown in Table 4, we quantitatively evaluate our modules with Attn [23] and SCBN (SCBN-sent) [24]. Note that all of the ablation experiments don't include the encoder network. By comparing each proposed component, we find our SEAttn and SEBN could adaptively extract proper features and improve the performance on IS.

We also visualize the importance of every single word compared with semantic features (Figure 4). On CUB dataset, some words (e.g., colors) have higher value than others, while results on COCO dataset are different. Every word in descriptions plays far more role than that on CUB dataset. We assume that CUB is a simple bird dataset, so SEAttn doesn't have to pay attention to every word. However, COCO is a far more diverse dataset and SEAttn has to focus on most words.

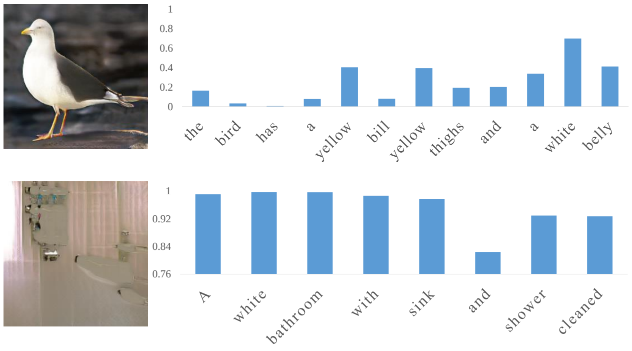


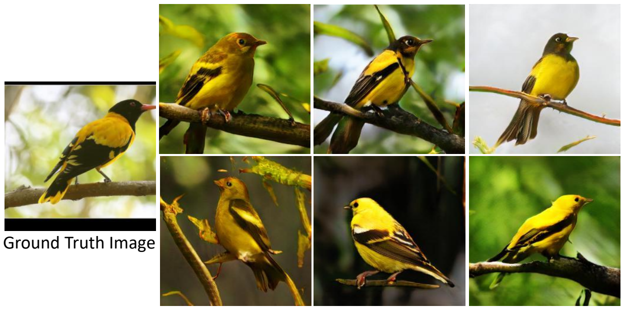
**Fig. 4**. Importance of every single word compared with semantic features in SEAttn. Left synthesized images are conditioned on the text descriptions below the histogram.

| Evaluation Metric | IS↑ | | FID↓ | |
|---|---|---|---|---|
| | CUB | COCO | CUB | COCO |
| TVBi-GAN ($64^2$) | $4.91 \pm .06$ | $29.71 \pm .82$ | 12.86 | 32.87 |
| TVBi-GAN ($128^2$) | $4.99 \pm .07$ | $30.79 \pm .70$ | 12.37 | 32.39 |
| TVBi-GAN ($256^2$) | $\mathbf{5.03 \pm .03}$ | $\mathbf{31.01 \pm .34}$ | **11.83** | **31.97** |

**Table 5**. The performance of different encoder network. Numbers in bracket represent input resolution.

**Encoder Analysis.** As shown in Table 5, we conduct an experiment to understand the performance of different grained encoder networks. IS and FID slightly improve, when encoders extract semantic features from high-quality images, because high resolution images usually contain more detailed features than lower ones. Due to the limitation of equipments, we encode $256^2$ images for our TVBi-GAN. Note that generator is same across all three encoder networks.

We could see an interesting phenomenon in Figure 5. We seperately generate images from different semantic features. Both generation ways synthesize high-quality images conforming to text descriptions. But image generated from encoder is almost similar to original image. This indicates that semantic features could encode object's orientations, shape details and postures. Significantly, descriptions toward this bird lack color information about the bird's head. As a result, generator soley based on text descriptions indicates this bird head is yellow, which is as same as the body. But when we synthesize images from semantic features of the original image, they don't lose the specific head color. This experiment tells us that abundantly semantic features are hidden behind images and text descriptions. Due to conditional loss (Section 3.3), we can't learn identically semantic features from original images entirely. But we could exploits this feature space to dig out precisely corresponding features between visual contents and natural language. Our experiment results presented in Tables 1, 2, and 3 demonstrate this process could help generator synthesize high-quality images.



(a) The bill is orange and starts wide before narrowing.

(b) This bird is yellow and black in color with a orange beak, and red eye rings.

(c) This bird has wings that are black and yellow and has an orange bill.

**Fig. 5**. Images from top line are conditioned on semantic features from the encoder, $G(E(image), s)$. The followings are generated solely from text, $G(CA(s), s)$.

## 5. CONCLUSION

In this paper, we propose a Textual-Visual Bidirectional Generative Adversarial Network called TVBi-GAN, for the fine-grained text-to-image synthesis task. First, we design two semantic enhanced modules, denoted as SEAttn and SEBN. SEAttn embeds semantic features in word vectors and improves reality of synthesized images. SEBN balances the semantic consistency and individual diversity. Furthermore, to explore semantic features behind images and text descriptions, we propose a cross-modal network and integrate visual contents into text-to-image generation process. Our experiment results on two real-world datasets show that TVBi-GAN achieves the state-of-the-art performance.

## 6. ACKNOWLEDGMENT

This work was supported in part by the National Key R&D Program of China (2018YFB0204302), and the National Natural Science Foundation of China (No.61972425, U1811264).

## 7. REFERENCES

[1] A. Brock, J. Donahue et al., “Large scale GAN training for high fidelity natural image synthesis.” In: *ICLR*, 2019 .

[2] J. Donahue and K. Simonyan, “Large scale adversarial representation learning.” *CoRR*, 2019, abs/1907.02544.

[3] V. Dumoulin, I. Belghazi et al., “Adversarially learned inference.” In: *ICLR*, 2017 .

[4] I. J. Goodfellow, J. Pouget-Abadie et al., “Generative adversarial nets.” In: *NeurIPS*, 2014 pp. 2672–2680.

[5] K. He, X. Zhang et al., “Deep residual learning for image recognition.” In: *CVPR*, 2016 pp. 770–778.

[6] D. P. Kingma and M. Welling, “Auto-encoding variational bayes.” In: *ICLR*, 2014 .

[7] Y. Le and Z. W. et al., “Acv-tree: A new method for sentence similarity modeling.” In: *IJCAI*, 2018 pp. 4137–4143.

[8] B. Li, X. Qi et al., “Controllable text-to-image generation.” In: *NeurIPS*, 2019 .

[9] T. Lin, M. Maire et al., “Microsoft COCO: common objects in context.” In: *ECCV*, 2014 pp. 740–755.

[10] X. Lin and Z. W. et al., “Saliency detection via multi-scale global cues.” *IEEE Trans. Multimedia*, 2019, 21(7):1646–1659.

[11] W. Liu and Z. W. et al., “Geo-alm: POI recommendation by fusing geographical information and adversarial learning mechanism.” In: *IJCAI*, 2019 pp. 1807–1813.

[12] A. Nguyen, J. Clune et al., “Plug & play generative networks: Conditional iterative generation of images in latent space.” In: *CVPR*, 2017 pp. 3510–3520.

[13] T. Qiao, J. Zhang et al., “Mirrorgan: Learning text-to-image generation by redescription.” In: *CVPR*, 2019 pp. 1505–1514.

[14] S. E. Reed, Z. Akata et al., “Generative adversarial text to image synthesis.” In: *ICML*, 2016 pp. 1060–1069.

[15] S. E. Reed, Z. Akata et al., “Learning what and where to draw.” In: *NeurIPS*, 2016 pp. 217–225.

[16] H. Tan, X. Liu et al., “Semantics-enhanced adversarial nets for text-to-image synthesis.” In: *ICCV*, 2019 .

[17] A. van den Oord, N. Kalchbrenner et al., “Pixel recurrent neural networks.” In: *ICML*, 2016 pp. 1747–1756.

[18] C. Wah, S. Branson et al., “The caltech-ucsd birds-200-2011 dataset.” *Advances in Water Resources*, 2011.

[19] C. Wang and S. D. et al., “Saliencygan: Deep learning semisupervised salient object detection in the fog of iot.” *IEEE Trans. Industrial Informatics*, 2020, 16(4):2667–2676.

[20] Z. Wang and L. M. et al., “MSGC: A new bottom-up model for salient object detection.” In: *ICME*, 2018 pp. 1–6.

[21] Z. Wang and L. M. et al., “Saliency detection via multi-center convex hull prior.” In: *ICASSP*, 2018 .

[22] Q. Xu, G. Huang et al., “An empirical study on evaluation metrics of generative adversarial networks.” *CoRR*, 2018, abs/1806.07755.

[23] T. Xu, P. Zhang et al., “Attngan: Fine-grained text to image generation with attentional generative adversarial networks.” In: *CVPR*, 2018 pp. 1316–1324.

[24] G. Yin, B. Liu et al., “Semantics disentangling for text-to-image generation.” In: *CVPR*, 2019 pp. 2327–2336.

[25] H. Zhang, T. Xu et al., “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks.” In: *ICCV*, 2017 pp. 5908–5916.

[26] H. Zhang, T. Xu et al., “Stackgan++: Realistic image synthesis with stacked generative adversarial networks.” *TPAMI*, 2019, 41(8):1947–1962.

[27] Z. Zhang, Y. Xie et al., “Photographic text-to-image synthesis with a hierarchically-nested adversarial network.” In: *CVPR*, 2018 pp. 6199–6208.

[28] M. Zhu, P. Pan et al., “DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis.” In: *CVPR*, 2019 pp. 5802–5810.