DAC-GAN: Dual Auxiliary Consistency Generative Adversarial Network for Text-to-Image Generation

Zhiwei Wang[1], Jing Yang[1]*, Jiajun Cui[1], Jiawei Liu[1], and Jiahao Wang[1]

East China Normal University, Shanghai, China
{wlf,liujiawei,jhwang}@stu.ecnu.edu.cn,
jyang@cs.ecnu.edu.cn, cuijj96@gmail.com

Abstract. Synthesizing an image from a given text encounters two major challenges: the integrity of images and the consistency of text-image pairs. Although many decent performances have been achieved, two crucial problems are still not considered adequately. (i) The object frame is prone to deviate or collapse, making subsequent refinement unavailable. (ii) The non-target regions of the image are affected by text which is highly conveyed through phrases, instead of words. Current methods barely employ the word-level clue, leaving coherent implication in phrases broken. To tackle the issues, we propose DAC-GAN, a Dual Auxiliary Consistency Generative Adversarial Network(DAC-GAN). Specifically, we simplify the generation by a single-stage structure with dual auxiliary modules. (1) Class-Aware skeleton Consistency(CAC) module retains the integrity of image by exploring additional supervision from prior knowledge and (2) Multi-label-Aware Consistency(MAC) module strengthens the alignment of text-image pairs at phrase-level. Comprehensive experiments on two widely-used datasets show that DAC-GAN can maintain the integrity of the target and enhance the consistency of text-image pairs.

1 Introduction

Cross-modal tasks are rapidly evolving and text-to-image generation[1] is one of the significant branches with a broad range of applications, like image-editing, computer-aided inpainting, etc. Literally, the task is defined as to generate a text-consistent image with fidelity from a given caption. Existing methods have achieved great success, but obstacles stand in the way of two principal aspects. First, deflection or collapse of the target object is a frequent occurrence, leading to slow convergence and poor quality of synthesized images. Second, the semantics of the text is embodied in the non-target objects of the image and word-level correspondence undermines semantic coherence.

As shown in Figure 1(a-ii), the main frame of the object deviates or even deforms within the generation phase, which leads to slow convergence and low

* Corresponding author

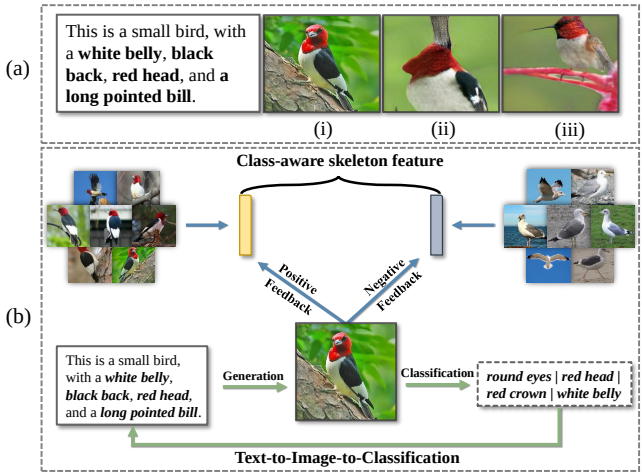


Fig. 1. Given the text description, existing methods yielded some unsatisfactory results. Compared with the ground-truth image in (a-i), the holistic frame of the target in (a-ii) has deviated during the training phase. And in (a-iii), some key attributes in the text are represented in non-target objects. To overcome the defects, the proposed DAC-GAN in (b) extracts the overall skeleton of each class to supervise the generation phase(the top half) and employs image multi-label classification to strengthen semantic consistency(the bottom half).

quality of images. Similar to the process of human painting upon a given text, we associate the class-aware frame of target objects at first and then generate high-quality, category-accurate images under this specific framework. In order to make the model perceive the discriminative attributes related to the category, we propose Class-Aware skeleton Consistency(CAC) module. The CAC leverages an image prior knowledge extractor(IPKE) to obtain a class-aware skeleton feature as additional supervision so as to retain the structural logic integrity of the image skeleton in the training process. Note that our skeleton features are unnecessarily needed to be trained with GAN, which accelerates the convergence of the model, and it can be easily transplanted to other networks.

As shown in Figure 1(a-iii), the text semantics not only changes the target but also affects non-target objects. What's more, the semantics is explicitly expressed by phrases, whereas the common practice leverage the word-level clue, i.e. splitting the sentence into discrete words. Word-level semantics breaks the coherency. For example, in the sentence "this small bird has a red crown and a white belly.", here, "small bird", "red crown" and "white belly" are the key elements that should not be split. To quantify the relevance between the generated image and the text, if such phrase-level attributes can be expressed from the image, it means that the latent manifold transformation is satisfactory. Then we introduce Multi-label-Aware Consistency(MAC), a module that embraces text prior knowledge extractor(TPKE) to tighten up the alignment of text and image

while mitigating the impact of word-level semantics on context in an intuitive and concise manner.

In addition, the majority of methods adopted multi-stage structure, sacrificing computational complexity, to generate images from coarse to fine like [2]. Inspired by [3], we adjust the primitive Conditional Batch Normalization(CBN) module with sentence-level and phrase-level clues. In the paper, we introduce Dual Conditional Batch Normalization(DCBN) as the backbone of the generator to form an end-to-end paradigm.

In this article, motivated by the aforementioned observations, we propose a novel model in Figure 1(b), called Dual Auxiliary Consistency Generative Adversarial Network(DAC-GAN).

Contributions in this article are expended as follows:

- The Class-Aware skeleton Consistency(CAC) leverages IPKE to distill the class-aware skeleton feature from prior knowledge to maintain the integrity of target object.
- The Multi-label-Aware Consistency(MAC) embraces TPKE to enhance the correspondence between text and image at the phrase-level.
- We propose an integral birdy structure to generate text-related images of high quality. The DCBN is the backbone to synthesize images. The CAC and the MAC are wings of birds to coordinate in the generation procedure. The extensive experimental results demonstrate that our method can obtain more integral images and higher correspondence between the image and text.

2 Related work

Due to the successful application of GANs[4, 5] in the field of image generation[6–11], a great quantity of works have been devoted to more complex tasks, such as text to image(T2I), image inpainting, etc. T2I is an interesting branch of image synthesis and one of its major difficulties lies in how to combine text semantics with image features.

Concatenating. Reed et al.[1] first attempted to synthesize photographic images from text descriptions by simply concatenating text vectors to visual features. To decompose the difficulty of generating high-resolution problems, Zhang et al.[2] stacked cascaded GANs to refine images from low resolution to high resolution, and introduced a common technique named Conditioning Augmentation. In order to stabilize the training phase and improve sample diversity, [12] arranged the multiple generators and discriminators in a tree-like structure. Different from StackGAN which required multi-stage training, HDGAN[13] introduced the hierarchically-nested discriminators to leverage mid-level representations of CNN in an end-to-end way.

Cross-modal Attention. Xu et al.[14] took advantage of the attention mechanism to help the model obtain more fine-grained information. Observing poor correlation between text and generated image, Qiao et al.[15] constructed a symmetrical structure like a mirror, text-to-image-to-text, to maintain the consistency between image and text. In order to alleviate the dependence on the

initial generated image, Zhu et al.[16] introduced an additional module Memory Network[17] to dynamically rectify the quality of the image. Cheng et al.[18] made the most of the captions in the training dataset and enriched the given caption from prior knowledge to provide more visual details. [19] employed contrastive loss in image to sentence, image region to word, and image to image to enforce alignment between synthesized pictures and corresponding captions.

Conditional Batch Normalization(CBN). Yin et al.[20] used a Siamese scheme[21] to implicitly disentangle high-level semantics from distinct captions and first adopted CBN[22, 3] for visual feature generation. For the efficiency of training, DF-GAN[23] decomposed the affine transformation from CBN and designed a Deep text-image Fusion Block(DFBlock) to enhance semantic fusion. As the backbone of DAC-GAN, DCBN leverages the whole caption and strong-feature phrases to strengthen the fusion between text and image.

3 Dual Auxiliary Consistency Generative Adversarial Network

As Figure 2 shows, the architecture of our DAC-GAN is integrated like a bird structure. (1)CAC and (2)MAC are similar to the wings of birds which manipulate the progress from their respective perspectives. To maintain the integrity of the target object, the CAC obtains class-aware skeleton features as additional supervision by IPKE. The MAC leverages TPKE module to enhance semantic consistency at phrase-level. (3)As the bone of the generator, DCBN plays a principal role in image generation with sentence-level and phrase-level clues. Details of the model are introduced below.

3.1 CAC: Class-Aware skeleton Consistency

Within the training phase, the main framework of the target will be deviated or even distorted, resulting in poor quality of the visual image. Taking the bird dataset as an example, results such as multi-headed birds, swordfish birds, and multiple pairs of eyes will appear. In analogy to painters with different painting styles, they can draw according to their imagination or outline the overall framework of the object first and then refine under this frame. In order to keep the structure of the generated image logical and complete, we obtain a common feature from multiple images under each category and use the feature as a skeleton feature in the generation, thus constraining the image generation procedure to continuously have the underlying structure of such species. The images I of dataset is organised as: $I = \{I_c | \, c = 0, \cdots, C-1\}$ where C represents the number of total species of dataset. $I_c = \{x_i | \, i = 0, \cdots, n-1\}$ and n notes the sum of pictures under c-th species.

In detail, the class-aware feature is a low-dimensional vector. By combining contrastive learning and CNN, we not only integrate a feature that distinguishes deep semantics in inter-class images, but also contains the diversity of intra-class images. It is different from a trainable class-aware embedding obviously.

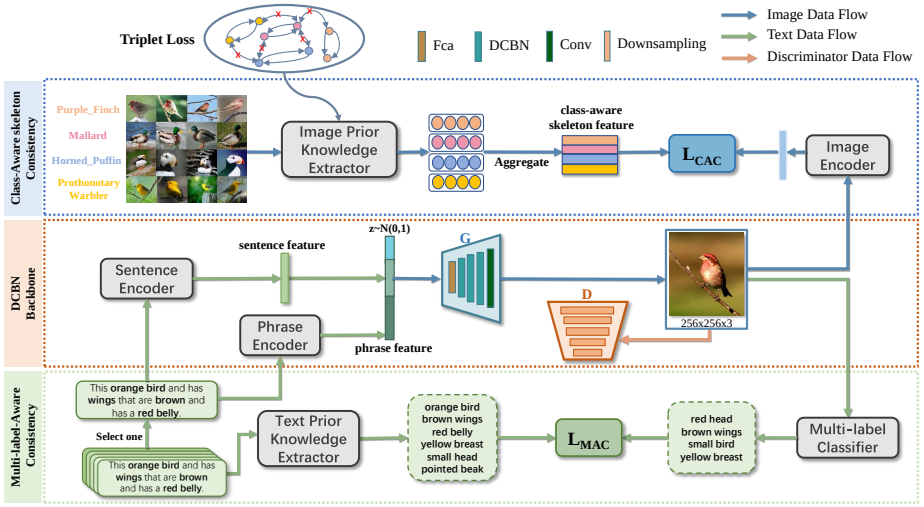


Fig. 2. The DAC-GAN architecture for text-to-image generation. We make the most of the caption by Dual Conditional Batch Normalization(DCBN). The Class-Aware skeleton Consistency(CAC) module is introduced to supervise which objects are currently being drawn and focus more on the main frame. We leverage the Multi-label-Aware-Consistency(MAC) module to strengthen the semantic consistency.

The latter can not capture such distinctness for it treats intra-class images as a single. It is also noted that, during the experiments we also apply clustering, and attention methods to integrate information. However, the improvement in effectiveness is limited, but it adds significant time complexity. We therefore go straight to the simplest averaging method. Taking categories as supervision, we customize the metric learning method[24–29] to increase the inter-class distance and decrease the intra-class distance by triplet loss[28, 30, 31]. At the same time, the triplet loss will implicitly act as a data augmentation. The triplet loss is defined as:

$$L\left(x_a, x_p, x_n\right) = \max\left(0, m + \left\|f\left(x_a\right) - f\left(x_p\right)\right\| - \left\|f\left(x_a\right) - f\left(x_n\right)\right\|\right), \tag{1}$$

where x_a, x_p, x_n indicate anchor, positive, negative samples respectively. The m is a margin constant.

Based on metric learning, an Image Prior Knowledge Extractor($IPKE$) transforms multiple pictures under each category into features and then aggregates a skeleton feature sk_c from the features. In order to improve efficiency, we have adopted the simplest method to obtain the average value of the features, which is used as additional skeleton-level supervision in the generation stage.

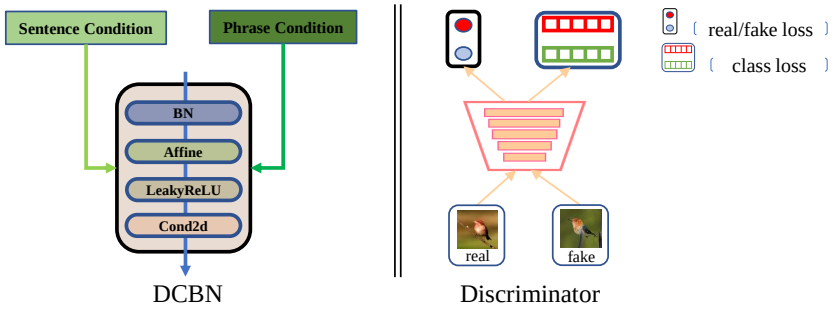


Fig. 3. The structure of DCBN and Discriminator. The generator consists of several DCBN modules, and the discriminator not only identifies the object as real or fake but also determines the category to which the object belongs.

The class-skeleton-aware feature f_{cls-sk} is plainly calculated as:

$$\begin{aligned} f_{cls-sk} &= \{sk_c \mid c = 0, \ldots, C-1\}, \\ sk_c &= \frac{1}{n} \cdot \sum_{i=1}^{n} IPKE(x_i), \end{aligned} \tag{2}$$

where $sk_c \in R^D$, D is the dimension after the extractor with shared parameters. We train the extractor and then obtain the skeleton feature of each category in advance, and the parameters of the extractor are fixed during the training stage.

3.2 MAC: Multi-label-Aware Consistency

A very challenging aspect of the T2I task is to maintain the coherency between text and image. MirrorGAN[15] employed text-to-image-to-text structure to enhance semantic consistency. The uncertainty associated with the passing of the results of two generation tasks is enormous and it inevitably introduced massive noise which degenerates models. However, it is simple to give from the perspective of human thinking. Take Figure 1 as an example, people intuitively observe salient features including red head, black wings, and white belly. If they can be matched with the real labels, it means that the generated image and text description are highly related.

By imitating the concise insight, we embrace a universal CNN-based model to classify images with multi-label. Note that, the labels are phrases instead of words. Taking a phrase as an example, "red throat" will be split into separate words "red" and "throat", altering the coherent semantics. We utilize a Text Prior Knowledge Extractor($TPKE$) to extract phrases from all n captions as target labels(n indicates the number of captions per image in the dataset), and then a ResNet model[32] is employed to encode the synthesized image x_i into visual features f_i. The MAC module is expressed as:

$$\begin{aligned} P_i &= TPKE(S_i^j), j \in \{1, \ldots, m-1\}, \\ f_i &= CNN\left(x_i\right), prob = \sigma\left(f_i\right), \end{aligned} \tag{3}$$

where m denotes the number of captions per image in datasets. The σ denotes the sigmoid function, and $prob$ is the predicted probability distribution over phrases. The $TPKE$ consists of the dependency parsing module of the NLP library Spacy and a rule-based approach.

When we perform multi-label classification tasks using CE-based loss, the category imbalance issue is inevitably encountered, which disturbs the model's concentration on the labels with less frequency. Inspired by[33], we introduce Circle Loss to overcome the imbalance. Modified multi-label circle loss treats multi-label classification as a pairwise comparison between the target category Ω_{pos} and the non-target category Ω_{neg}. The final Multi-label-Aware Consistency(MAC) loss is promoted as:

$$L_{MAC} = \log\left(1 + \sum_{i \in \Omega_{neg}} e^{s_i}\right) + \log\left(1 + \sum_{j \in \Omega_{pos}} e^{-s_j}\right), \tag{4}$$

where s_i denotes the score of $i-th$ target category. Meanwhile, with the help of the good property of logsumexp, the weight of each item is automatically balanced.

3.3 DCBN: Dual Conditional Batch Normalization

The majority of previous methods enhanced the fusion of text and image by multi-stage structure or attention scheme, which complicated the training phase. Inspired by several works[20, 23, 20, 34], we introduce the Dual Conditional Batch Normalization(DCBN) module to strengthen semantic fusion while simplifying the typical multi-stage structure. The global sentence contains coherent semantics, however, the noise in it affects implications while phrases contain explicit semantics and less noise. Through the analyses, we refine CBN with dual clues, sentence-level clue and phrase-level clue.

The naive CBN is a class-condition variant of Batch Normalization(BN) and the core of this change is that linguistic embeddings can be wielded to modulate the scaling up and down of the feature map. It inputs the linguistic vector $x \in R^{N \times C \times H \times W}$ into a multi-layer perceptron to obtain γ and β. Since the parameters depend on the input feature, cross-modal information interaction between text and image can be achieved. The CBN formula is defined as:

$$\begin{cases} y = \frac{x - E[x]}{\sqrt{\mathrm{Var}[x] + \varepsilon}} \cdot \gamma_{\mathrm{new}} + \beta_{\mathrm{new}} \\ \gamma_{\mathrm{new}} = \gamma + \gamma_c \\ \beta_{\mathrm{new}} = \beta + \beta_c, \end{cases} \tag{5}$$

where $E[x]$ and $Var[x]$ are the mean and variance for each channel, and γ_c, β_c are modulation parameters of condition c. Our refined DCBN function is formatted as:

$$\begin{cases} \gamma_{\mathrm{new}} = \gamma + \lambda_1 \cdot \gamma_s + \lambda_2 \cdot \gamma_p \\ \beta_{\mathrm{new}} = \beta + \lambda_1 \cdot \beta_s + \lambda_2 \cdot \beta_p, \end{cases} \tag{6}$$

where λ_1 and λ_2 are weights of condition s and p.

As shown in Figure 2, a sentence encoder S–$Encoder$[35] is used to extract global sentence embedding s from the given caption and a phrase encoder P–$Encoder$[36] to embed strong-feature phrases into a semantic vector p by

$$\begin{cases} s = S\text{–}Encoder(S) \\ p = P\text{–}Encoder(P), \end{cases} \tag{7}$$

where $P = \{P_l | \, l = 0, \cdots, L-1\}$ and L represents the number of phrases extracted from the whole given sentence S.

In previous studies[29, 1], the linguistic embedding was high dimensional which caused discontinuity in latent semantics because of lacking data. To transmute the vector into a desirable manifold, we follow the conventional technique in [2]. The condition augmentation marked by F_{ca}, yields more pairs of image-text under the small data limitation, thereby promoting the robustness to a tiny disturbance on the condition manifold. The equation is defined as:

$$\begin{cases} s_{ca} = F_{ca}(s) \\ p_{ca} = F_{ca}(p). \end{cases} \tag{8}$$

3.4 Objective functions

We note binary cross entropy and multi-label cross entropy with label smoothing as BE and CE_S respectively. The formula of CE_S is calculated by:

$$\begin{aligned} CE_S &= -\sum_{i=1}^{K} p_i \log q_i, \\ p_i &= \begin{cases} (1-\varepsilon), & \text{if } (i = y) \\ \frac{\varepsilon}{K-1}, & \text{if } (i \neq y) \end{cases}. \end{aligned} \tag{9}$$

Here, ε is a small constant and K denotes the number of labels.

Following the discriminator loss in [37], the discriminator D not only distinguishes real data distribution from synthetic distribution but also classifies generated sample to a specific category. The training loss L_{D_S} related to the source of the input (real, fake, or mismatch) is expressed as

$$L_{D_S} = BE \underbrace{\left(D_S\left(I_i, l_j\right), 1\right)}_{i \sim real, j \sim real} + BE \underbrace{\left(D_S\left(I_i, l_j\right), 0\right)}_{i \sim fake, j \sim real} + BE \underbrace{\left(D_S\left(I_i, l_j\right), 0\right)}_{i \sim mis, j \sim real}, \tag{10}$$

where l_i denotes the text captions. Similarly, L_{D_C} relates to which the input is supposed to pertain. The formulation is defined by

$$L_{D_C} = CE_S \underbrace{\left(D_C\left(I_i, l_j\right), C_r\right)}_{i \sim real, j \sim real} + CE_S \underbrace{\left(D_C\left(I_i, l_j\right), C_r\right)}_{i \sim fake, j \sim real} + CE_S \underbrace{\left(D_C\left(I_i, l_j\right), C_w\right)}_{i \sim mis, j \sim real}, \tag{11}$$

the D is trained by minimizing the loss(L_D) as follows:

$$L_D = L_{D_S} + L_{D_C}. \tag{12}$$

Besides common condition loss which is denoted as:

$$L_{G_C}=BE\underbrace{(D_S\left(I_i,l_j\right),1)}_{i\sim fake,j\sim real}+CE_S\underbrace{(D_C\left(I_i,l_j\right),C_r)}_{i\sim fake,j\sim real}, \tag{13}$$

we further introduce a cosine-based class-aware skeleton consistency loss(L_{CAC}) to maintain the frame of the objects during training as follows:

$$L_{CAC}=CS\left(sk_c,f_c^j\right),j\in\{1,\ldots,n-1\}, \tag{14}$$

where CS represents cosine similarity and f_c^j indicates the feature of j-th image which belongs to c-th species. Meanwhile, we employ L_{MAC} to align the text-image semantics. Mathematically, the generation loss is expressed as:

$$L_G=L_{G_C}+\lambda_3\cdot L_{CAC}+\lambda_4\cdot L_{MAC}, \tag{15}$$

in which λ_3,λ_4 are the modulating weights of class-aware skeleton consistency loss and multi-label-aware consistency loss.

3.5 Implementation details

Following [14, 23], a pre-trained bi-directional LSTM is employed to yield a global sentence embedding s, and we use an embedding layer in the CNN-RNN[36] framework to embed phrases into a semantic vector p in a dimension of 256. Our generator consists of 7 DCBN blocks conditioned with the sentence and phrase embeddings, then we set $\lambda_1=0.4,\lambda_2=0.6$ to weight the dual conditions. As to the CAC module, we use a metric-learning based method to extract the features from images and average the multiple features into a 256-dimensional class-aware skeleton feature. In MAC, we utilize the powerful NLP library Spacy and rule-based methods to extract the phrases and leverage the phrases with more than 200 frequencies in CUB (50 in Oxford-102) as the multi-label of the image. The target labels are all phrases extracted from the corresponding ten captions, instead of a single one. Specifically, we set hyper-parameters $\lambda_3=1.6,\lambda_4=0.8$, and the learning rates of discriminator and generator are set to 0.0001, 0.0002 respectively.

4 Experiments

In practice, we evaluate our proposed model qualitatively and quantitatively. According to different datasets, we compare our model with various state-of-the-art methods and validate key components of the model through ablation studies.

Table 1. The Inception Score(higher is better) and Fréchet Inception Distance(lower is better) of the state-of-the-art on CUB and Oxford-102. Note that, the FID on CUB of MirrorGAN and the IS and FID on Oxford-102 of DF-GAN are calculated by reproducing from the open source code. And the FID of HDGAN is calculated from released weights.

	CUB		OX-ford102	
Method	IS↑	FID↓	IS↑	FID↓
GAN-INT-CLS[1]	2.88±.04	68.79	2.66±.03	79.55
StackGAN[2]	3.70±.04	51.89	3.20±.01	55.28
TAC-GAN[37]	(-)	(-)	2.88±.04	(-)
HDGAN[13]	4.15±.05	(-)	<u>3.45±.07</u>	<u>37.19</u>
AttnGAN[14]	4.36±.03	23.98	(-)	(-)
MirrorGAN[15]	4.56±.05	23.47	(-)	(-)
DAE-GAN[38]	4.42	15.19	(-)	(-)
DF-GAN[23]	**5.10**	<u>14.81</u>	3.32±.03	39.69
DAC-GAN	<u>4.86 ±.06</u>	**14.77**	**3.59±.06**	**35.31**

4.1 Datasets and evaluation metrics

Datasets. We evaluate the proposed model on fundamental datasets, CUB bird dataset[39] and Oxford-102 flower dataset[40]. Following previous works[2, 14, 16, 23], we process these datasets into class-disjoint training and testing sets. The CUB bird dataset contains 200 species with 11788 images in which 8855 images of birds from 150 categories are employed as training data while the left-over 2933 images from 50 categories are for testing. The Oxford-102 flower contains 7034 training images and 1155 testing data belonging to 82 and 20 categories separably. Each image in both datasets has ten text descriptions.

Evaluation metrics. We leverage two evaluation metrics as same as previous works[41, 42]. The Inception Score(IS) and Fréchet Inception Distance(FID) quantitatively measure the quality and diversity of the images to a certain extent.

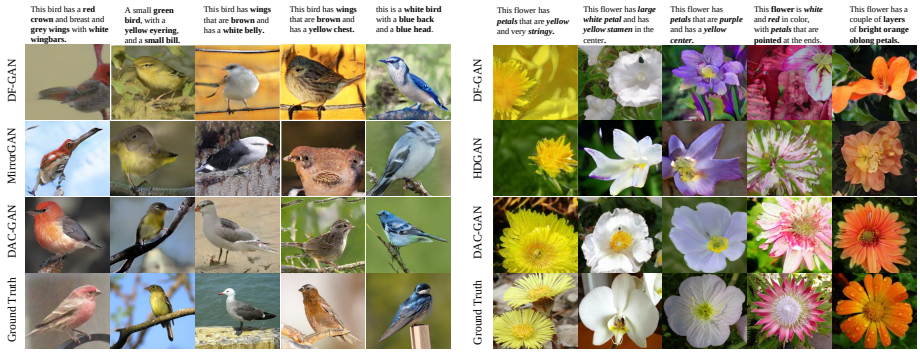
In order to measure the correspondence of text description and visual image, Xu et al.[14] first introduced the evaluation of image retrieval, named R-precision, into the text-to-image generation task. Given an image, the R-precision is calculated by retrieving correlated text in a text description set. Then we calculate the cosine distance between the global image feature and 100 texts consisting of 99 random samples and 1 ground truth. For each retrieval, if r relevant items are in top R ranked results and the $R-precision = r/R$. Following previous works, we set $R = 1$.

4.2 Experiment results

We compare our method with the state-of-the-art methods on CUB and Oxford-102 datasets from both quantitative and qualitative perspectives. For each dataset, we compare different methods. The detailed results are shown in Table 1 and Table 2. Then we use a subjective visual comparison in Figure 4 and an elaborated

Table 2. The performances of R-precision on the CUB and Oxford-102 datasets. We compare different modules of our DAC-GAN with Mirror-GAN and DF-GAN. The Baseline denotes that we barely utilize the naive CBN for image generation.

Methods	CUB	Oxford-102
MirrorGAN	19.84	(-)
HDGAN	(-)	16.95
DF-GAN	19.47	18.94
Baseline+DCBN	20.89	**20.51**
Baseline+CAC	20.30	18.85
Baseline+MAC	<u>21.17</u>	19.46
DAC-GAN	**21.62**	<u>19.64</u>
Ground Truth	27.35	21.14



(a) CUB (b) Oxford-102

Fig. 4. Visual comparison with distinct state-of-the-art methods on CUB and Oxford-102. We compare the DAC-GAN with MirrorGAN[15] and DF-GAN[23] on CUB(on the left). As to Oxford-102, we choose HDGAN[13] and DF-GAN[23] as a comparison(on the right).

human evaluation in Figure 5 to validate the integrity of images and text-related degree.

4.3 Quantitative results

As shown in Table 1, higher inception score and lower fréchet inception distance mean better quality and diversity. Our DAC-GAN achieves 4.86 IS and 14.77 FID on CUB. Compared with DF-GAN, DAC-GAN decreases the FID from 14.81 to 14.77 which outperforms other methods by a large margin. As shown in Table 1, we conduct the performance on the Oxford-102 dataset. Compared with HDGAN, DAC-GAN improves IS from 3.45 to 3.59. We measure the IS and FID of DF-GAN by reproducing and DAC-GAN decreases FID from 39.69 to 35.31(11.04% reduction). The results demonstrate that DAC-GAN achieves better quality and diversity of images.

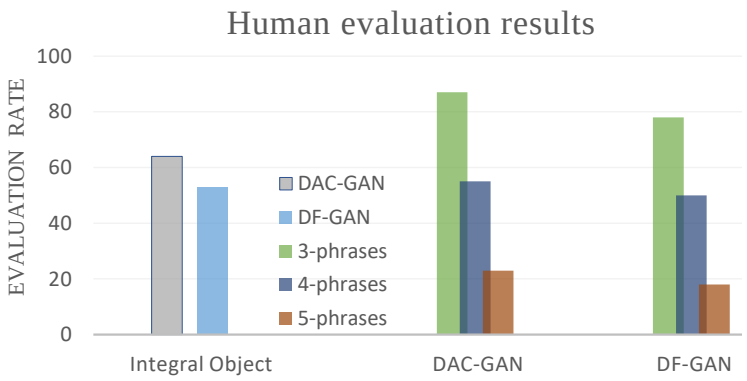


Fig. 5. The results of human evaluation on the integrity of objects and alignment of text-to-image. The higher score of the integral object means the better quality of synthesized images. n-phrases(n=3,4,5) indicates that the number of matched phrase-labels is n.

As Table 2 shows, the DAC-GAN improves the R-precision by 8.97% compared with MirrorGAN and 11.04% compared with DF-GAN on CUB. For Oxford-102, the improvements are 21.00% and 8.29% compared to HDGAN and DF-GAN respectively. Higher R-precision indicates that semantic consistency between text description and synthesized images are better.

4.4 Qualitative results

Visual Evaluation: We visualize the results to get a better sense of different models. We compare with DF-GAN on both datasets. In addition, we replenish MirrorGAN on CUB and HDGAN on Oxford-102. First of all, it can be seen that an interesting phenomenon in the first line of Figure 4. The attributes of the text expression are not only reflected in the target object but also reflected in the background or non-target objects. It means the semantics has shifted after the sentence is divided into words. For example, attributes like color are confused about whether to decorate the background or the target. And then, let's note the second row in Figure 4. Leaving aside the details, the overall structure of the target object deviates greatly from the real. This is not to mention the detail essence, and such a phenomenon will seriously affect the convergence and the quality of images. DAC-GAN can obtain images with more integral structure and more accurate text correspondence.
Human test: For the CUB dataset, we designed a manual evaluation method. In the first stage, we selected ten random sets from the generated data and five real data sets, about 300 images in all. We hired 30 employees of different professions to perform a simple sensory evaluation of the fifteen sets(the employees do not know which are the real groups). In the second stage, we selected employees who successfully distinguished real from generated data in the first phase. These

Table 3. The Inception Score and Fréchet Inception Distance on two benchmarks with the different modules of DAC-GAN, including DCBN, CAC, MAC. Different from DAC-GAN, DAC-GAN-word denotes that we split the sentence into words.

Metric	IS ↑		FID ↓	
	CUB	Oxford-102	CUB	Oxford-102
Baseline	4.45±.06	3.22±.03	19.93	40.58
Baseline +DCBN	4.62±.02	3.48±.04	18.61	39.23
Baseline +CAC	4.68±.03	3.46±.05	18.46	37.23
Baseline + MAC	4.65±.04	3.47±.02	17.10	36.73
DAC-GAN-word	4.57±.04	3.46±.03	19.12	37.42
DAC-GAN	**4.86±.06**	**3.59±.06**	**14.77**	**35.31**

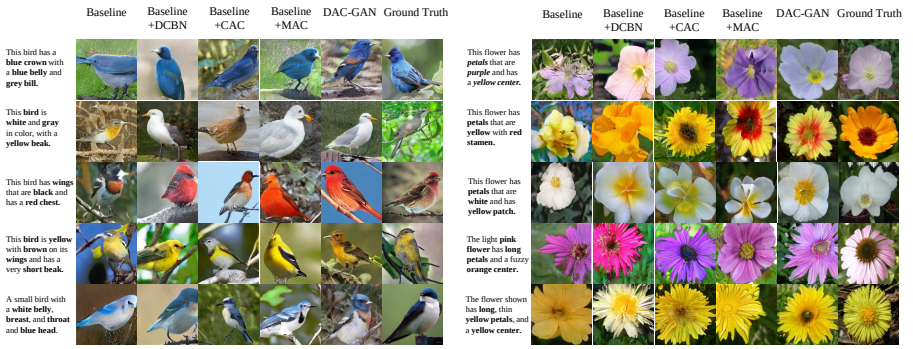
employees observe 10 groups of DF-GAN and 10 groups of DAC-GAN generated images while marking whether each image contains a complete object or not. As shown in Figure 5, DAC-GAN improves the integrity ratio from 0.53 to 0.64 (20.75% improvement) compared with DF-GAN. In the third stage, to verify the correlation between text and image intuitively and concisely, we extracted the corresponding key phrases from the text as the labels of the image. We provided this prior knowledge to the employees who only need to observe whether the corresponding labels can be found in the image. We counted the number of matched labels for 3,4 and 5 respectively. As shown in the Figure 5, DAC-GAN improves the n-phrases($n = 3, 4, 5$) by a large margin(11.54%, 10.00%, 27.78%). The results verify that the DAC-GAN generates more pictures of the integral object and correlates them more closely with corresponding text.

4.5 Ablation study

In this section, we conduct ablation studies on DAC-GAN and its variants. We define DAC-GAN with a naive CBN module as our baseline, and verify the performance of our DCBN, CAC, and MAC by including or excluding related modules.

As Table 3 shows, by including different components, IS and FID are consistently improved, which indicates that the distinct modules are effective. In addition, we compare word-level DAC-GAN to phrase-level DAC-GAN. The results of word-level DAC-GAN demonstrate that phrases have more explicit semantic information than words. As Table 2 shows, the R-precision of Baseline+DCBN and Baseline+MAC are higher than Baseline+MAC, for the former explicitly leverages phrase-level semantics.

We visualize samples generated by different modules in Figure 6. Baseline+DCBN and Baseline+MAC focus more on textual correspondence while Baseline+CAC focuses more on the integrity of the target. It indicates that phrase-level knowledge can strengthen the cross-modal correspondence and the class-aware skeleton feature maintains integrity of the target image. By integrating all modules, DAC-GAN adjusts text relevance and target integrity to obtain better quality images with more details.



(a) CUB (b) Oxford-102

Fig. 6. The images are synthesized by different modules of DAC-GAN including DCBN, CAC, and MAC on CUB and Oxford-102. The Baseline indicates that we barely employ naive CBN to generate images.

4.6 Limitation and discussion

Although our model has achieved good results, there are still some shortcomings worth discussing. First, the process we extract phrases does not take the multi-hop phrases into consideration, which could be further improved. Moreover, two benchmark datasets we experiment on only have a single object and a simple scene which far from the real world. In view of the large gap between the target objects in the coco dataset[43], the CAC and MAC modules can have a greater effect on the results. The problem is how to obtain the class-aware skeleton features of different targets, which is also the direction of our future research.

5 Conclusion

In this paper, we design a novel Dual Auxiliary Consistency Generative Adversarial Network(DAC-GAN) for text-to-image generation task. To maintain the integrity of target object, the CAC module leverages an IPKE module to distill the class-aware skeleton features as additional supervision. The MAC employs a TPKE module to enhance the alignment of text-to-image. Compared with other methods, DAC-GAN can maintain the integrity of target objects and the correspondence between image and text. Moreover, The DCBN employs sentence-level and phrase-level to strengthen the fusion between language and visual.

Acknowledgement This research is funded by the Science and Technology Commission of Shanghai Municipality 19511120200.

References

1. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International Conference on Machine Learning, PMLR (2016) 1060–1069

2. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. (2017) 5907–5915
3. De Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., Courville, A.: Modulating early visual processing by language. arXiv preprint arXiv:1707.00683 (2017)
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
5. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., Chen, X.: Improved techniques for training gans. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R., eds.: Advances in Neural Information Processing Systems. Volume 29., Curran Associates, Inc. (2016)
6. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
7. Tang, H., Xu, D., Sebe, N., Wang, Y., Corso, J.J., Yan, Y.: Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 2417–2426
8. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: International conference on machine learning, PMLR (2019) 7354–7363
9. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 4401–4410
10. Tang, H., Xu, D., Liu, G., Wang, W., Sebe, N., Yan, Y.: Cycle in cycle generative adversarial networks for keypoint-guided image generation. In: Proceedings of the 27th ACM International Conference on Multimedia. (2019) 2052–2060
11. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 8110–8119
12. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE transactions on pattern analysis and machine intelligence **41** (2018) 1947–1962
13. Zhang, Z., Xie, Y., Yang, L.: Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 6199–6208
14. 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 networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 1316–1324
15. Qiao, T., Zhang, J., Xu, D., Tao, D.: Mirrorgan: Learning text-to-image generation by redescription. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 1505–1514
16. Zhu, M., Pan, P., Chen, W., Yang, Y.: Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 5802–5810
17. Weston, J., Chopra, S., Bordes, A.: Memory networks. arXiv preprint arXiv:1410.3916 (2014)

18. Cheng, J., Wu, F., Tian, Y., Wang, L., Tao, D.: Rifegan: Rich feature generation for text-to-image synthesis from prior knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 10911–10920
19. Zhang, H., Koh, J.Y., Baldridge, J., Lee, H., Yang, Y.: Cross-modal contrastive learning for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 833–842
20. Yin, G., Liu, B., Sheng, L., Yu, N., Wang, X., Shao, J.: Semantics disentangling for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 2327–2336
21. Chung, D., Tahboub, K., Delp, E.J.: A two stream siamese convolutional neural network for person re-identification. In: Proceedings of the IEEE international conference on computer vision. (2017) 1983–1991
22. Dumoulin, V., Shlens, J., Kudlur, M.: A learned representation for artistic style. arXiv preprint arXiv:1610.07629 (2016)
23. Tao, M., Tang, H., Wu, F., Jing, X.Y., Bao, B.K., Xu, C.: Df-gan: A simple and effective baseline for text-to-image synthesis. arXiv e-prints (2020)
24. Kulis, B., et al.: Metric learning: A survey. Foundations and Trends® in Machine Learning **5** (2013) 287–364
25. Li, J., Lin, X., Rui, X., Rui, Y., Tao, D.: A distributed approach toward discriminative distance metric learning. IEEE transactions on neural networks and learning systems **26** (2014) 2111–2122
26. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International workshop on similarity-based pattern recognition, Springer (2015) 84–92
27. Ding, Z., Fu, Y.: Robust transfer metric learning for image classification. IEEE Transactions on Image Processing **26** (2016) 660–670
28. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
29. Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. Advances in neural information processing systems **29** (2016) 217–225
30. Dong, X., Shen, J.: Triplet loss in siamese network for object tracking. In: Proceedings of the European conference on computer vision (ECCV). (2018) 459–474
31. Ge, W.: Deep metric learning with hierarchical triplet loss. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 269–285
32. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
33. Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., Wei, Y.: Circle loss: A unified perspective of pair similarity optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 6398–6407
34. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612 (2017)
35. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9** (1997) 1735–1780
36. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: Cnn-rnn: A unified framework for multi-label image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 2285–2294
37. Dash, A., Gamboa, J.C.B., Ahmed, S., Liwicki, M., Afzal, M.Z.: Tac-gan-text conditioned auxiliary classifier generative adversarial network. arXiv preprint arXiv:1703.06412 (2017)

38. Ruan, S., Zhang, Y., Zhang, K., Fan, Y., Chen, E.: Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. (2021)
39. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. (2011)
40. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, IEEE (2008) 722–729
41. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016) 2234–2242
42. 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. Advances in neural information processing systems **30** (2017)
43. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, Springer (2014) 740–755