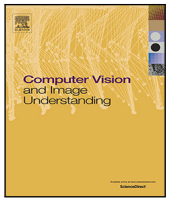
Computer Vision and Image Understanding 245 (2024) 104042

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



Text to image synthesis with multi-granularity feature aware enhancement Generative Adversarial Networks

Pei Dong, Lei Wu *, Ruichen Li, Xiangxu Meng, Lei Meng

School of Software, Shandong University, 1500 ShunHua Road, High Tech Industrial Development Zone, JiNan 250101, China

ARTICLE INFO

Communicated by Lu Jiang

Keywords:
Generative adversarial network
Multi-granularity feature aware enhancement
Text-to-image
Autoregressive
Diffusion

ABSTRACT

Synthesizing complex images from text presents challenging. Compared to autoregressive and diffusion model-based methods, Generative Adversarial Network-based methods have significant advantages in terms of computational cost and generation efficiency yet remain two limitations: first, these methods often refine all features output from the previous stage indiscriminately, without considering these features are initialized gradually during the generation process; second, the sparse semantic constraints provided by the text description are typically ineffective for refining fine-grained features. These issues complicate the balance between generation quality, computational cost and inference speed. To address these issues, we propose a Multi-granularity Feature Aware Enhancement GAN (MFAE-GAN), which allows the refinement process to match the order of different granularity features being initialized. Specifically, MFAE-GAN (1) samples category-related coarse-grained features and instance-level detail-related fine-grained features at different generation stages based on different attention mechanisms in Coarse-grained Feature Enhancement (CFE) and Fine-grained Feature Enhancement (FFE) to guide the generation process spatially, (2) provides denser semantic constraints than textual semantic information through Multi-granularity Features Adaptive Batch Normalization (MFA-BN) in the process of refining fine-grained features, and (3) adopts a Global Semantics Preservation (GSP) to avoid the loss of global semantics when sampling features continuously. Extensive experimental results demonstrate that our MFAE-GAN is competitive in terms of both image generation quality and efficiency.

1. Introduction

Text-to-image synthesis requires generating photo-realistic images based on given text guidance. Due to the significant practical value in various applications, such as computer-aided design (Chen et al., 2018; Liu et al., 2021a) and art generation (Zhi, 2017; Cheng et al., 2021), text-to-image synthesis has become an active research area in both natural language processing and computer vision communities.

The remarkable evolution in Generative Adversarial Networks (GANs) (Goodfellow et al., 2020; Mirza and Osindero, 2014) has spear-headed promising results in text-to-image synthesis. To express more explicit category information and richer instance-level details while ensuring text–image semantic consistency, multi-stage methods (Zhang et al., 2017, 2018) stack a series of generator–discriminator pairs to generate initial low-resolution images and refine the initial images to high-resolution ones. Based on them, additional DAMSM module (Xu et al., 2018), Cyclic Consistency (Qiao et al., 2019), Disentangled Encoder (Dong et al., 2022) or Dynamic Memory module (Zhu et al., 2019) is used to ensure the semantic consistency between text and images. GAN networks demonstrate efficiency in inferencing image distributions based on text embeddings, yet the quality of the generated results falls short of competitiveness.

Recent autoregressive and diffusion models, such as DALL-E (Ramesh et al., 2021) and Imagen (Saharia et al., 2022), based on extensive data collection and pre-training, demonstrate remarkable capabilities in synthesizing complex scenes. However, these models often comprise billions of parameters, resulting in substantial computational costs. This large model size presents significant challenges for academic institutions and individual researchers, limiting the feasibility of conducting further research. Additionally, the lack of intuitive smoothing latent space makes the generation process rely on the delicately designed text prompts. Finally, these models generate images through token-by-token generation or progressively denoising, a process involving numerous inference steps, significantly reducing real-time performance.

To enhance the balance between generation quality, the computational cost during training, and the model's inference speed, we focus on further optimizing GAN and have found two challenges. Firstly, current GAN-based methods typically attempt to refine all granularity features indiscriminately at each stage. However, the different granularity features are initialized gradually rather than simultaneously during generation. This means the network needs to refine poor-quality features within regions that have yet to generate initial semantics,

* Corresponding author.
E-mail address: i_lily@sdu.edu.cn (L. Wu).

https://doi.org/10.1016/j.cviu.2024.104042
Received 3 September 2023; Received in revised form 1 May 2024; Accepted 17 May 2024
Available online 20 May 2024

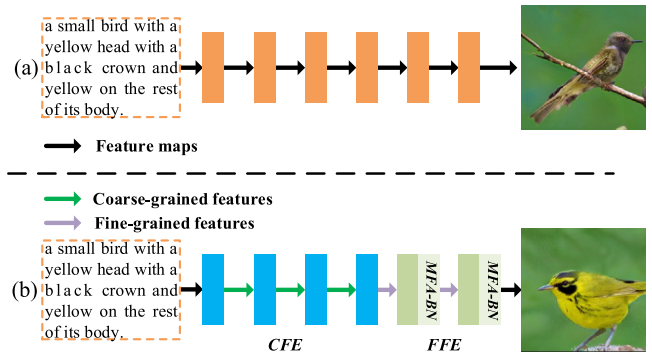


Fig. 1. (a) Existing models typically attempt to refine all feature outputs from the previous stage. (b) MFAE-GAN samples the features in different granularity respectively, which are successively fed to the Coarse-grained Feature Enhancement (CFE) or Fine-grained Feature Enhancement (FFE) modules for refinement. Multi-granularity Features Adaptive Batch Normalization (MFA-BN) module in FFE enables denser semantic constraints.

Table 1
Comparative analysis of previous text-to-image model. Considering Multiple Stage, Attention Mechanism, Multi-Granularity, Additional Control and Deep Fusion, MFAE-GAN uses only text as input in training and is performed in an end-to-end process.

Model	Multiple Stage	Attention Mechanism	Multi Granularity	Additional Control	Deep Fusion
StackGAN (Zhang et al., 2017)	✓	×	×	×	×
AttnGAN (Xu et al., 2018)	✓	✓	×	×	×
DM-GAN (Zhu et al., 2019)	×	✓	✓	×	×
MirrorGAN (Qiao et al., 2019)	×	✓	×	×	×
Obj-GAN (Li et al., 2020)	✓	✓	×	✓	×
ControlGAN (Li et al., 2019)	✓	×	×	✓	×
DF-GAN (Tao et al., 2022)	×	×	×	×	✓
SSA-GAN (Liao et al., 2022)	×	×	×	×	✓
RAT-GAN (Ye et al., 2022)	×	✓	×	×	✓
GALIP (Tao et al., 2023)	×	×	×	×	✓
StyleGAN-T (Sauer et al., 2023)	✓	✓	×	×	✓
GigaGAN (Kang et al., 2023)	✓	✓	×	×	×
MFAE-GAN (ours)	×	✓	✓	×	✓

such as the refinement of texture features of bird wings generated in the early stages of the network. This strategy compromises the overall refinement quality of the network. Consequently, the final generated images frequently suffer from deformations and fail to convey explicit category information due to a lack of instance-level details. Secondly, current methods refine image features by fusing text embeddings. However, almost text embeddings can only provide sparse semantics and lack detailed descriptions of fine-grained features; therefore, the semantic constraints provided by the text descriptions hardly work when generating fine-grained features. This inadequacy poses a significant challenge for networks in synthesizing fine-grained details, especially when sufficient data is lacking to augment the network's reasoning capabilities.

To address above issues, we propose a Multi-granularity Feature Aware Enhancement GAN (MFAE-GAN) to make the refinement process in each module match the order of different granularity features being initialized during the generation process. MFAE-GAN contains two core modules: the Coarse-grained Feature Enhancement (CFE) module and the Fine-grained Feature Enhancement (FFE) module. The two modules sample the different granularity features separately at different stages (as shown in Fig. 1) and constrain the generation process spatially by explicitly mapping the multi-granularity image sketches. Specifically, the CFE and the FFE employ word-level channel attention and additional word-level spatial attention, respectively, to capture coarse-grained representations (e.g. poses and shapes) and fine-grained details (e.g. textures and colors). These are then mapped into feature sketches to enhance category clarity in the initial stages and highlight instance-level specifics in subsequent stages. Furthermore, to provide denser semantic constraints than textual semantic information when adding fine-grained features, we propose a Multi-granularity Features Adaptive Batch Normalization (MFA-BN) module. MFA-BN adapts modulation parameters conditioned on the different granularity features sampled by word-level spatial and channel attention for text–image fusion at multi-granularity. This encourages the network to focus on refining fine-grained features while preserving the integrity of coarse-grained features. Finally, since continuous sampling of specific granularity features may result in the loss of global semantics, we propose a Global Semantics Preservation (GSP) module. This module fuses intermediate features and sentence embeddings through a streamlined way to supplement global semantics and maintain semantic integrity. We also introduce a pre-trained CLIP model (Radford et al., 2021) as an evaluation metric in calculating the loss to maximize the text–image similarity. Extensive experiments on CUB, COCO and CC series datasets illustrate that the proposed MFAE-GAN model significantly outperforms the previous methods, quantitatively and qualitatively. Moreover, we conduct a series of analysis experiments to evaluate the importance of each component in our approach and further validate the effectiveness of MFAE-GAN in balancing generation quality, computational cost and inference speed.

Our main contributions can be summarized as follows:

- We propose a novel framework, MFAE-GAN, to sample and refine features of different granularity separately, which can reduce the mutual interference between different granularity features during the generation process. It can guarantee the image refinement quality of the network.
- The CFE and FFE modules are designed to enhance category-related coarse-grained and instance-level detail-related fine-grained features, respectively, and impose constraints spatially from multi-granularity feature sketches.
- We propose an MFA-BN module, which provides denser semantic constraints than textual semantic information when adding fine-grained features, and a GSP module that preserves semantic integrity while continuously sampling specific granularity features, which allows the generated images to contain richer details.

2. Related work

Reed et al. (2016a,b) first successfully synthesizes plausible images using a conditional generative model (cGANs). Based on cGANs, input conditions such as mask (Park et al., 2019), sketch (Zhang et al., 2019; Lu et al., 2017; Toda et al., 2022; Liu et al., 2021b; Koley et al., 2023), segmentation (Liu et al., 2019) and layout (Zhao et al., 2019; He et al., 2021; Xue et al., 2023) are tried for controlled image generation. Compared to the above input conditions, text becomes the main control means due to its advantages such as conforming to subjective human expressions, high flexibility, easy interaction and collection of training data. Most current text-to-image synthesis approaches can be broadly classified into GAN-based and Large Pre-training Models based on autoregressive and diffusion models.

GAN-based Models To synthesize higher-resolution images based on text descriptions, StackGAN (Zhang et al., 2017) and StackGAN++ (Zhang et al., 2018) stack multiple generators and discriminators and employ the text information to refine the rough initial image to a high-resolution photo-realistic one. AttnGAN (Xu et al., 2018) adds attention mechanism components into a multi-stage generator pipeline. The attention mechanism components can help the network synthesize more fine-grained details based on relevant local word embedding. DM-GAN (Zhu et al., 2019) employs a memory writing gate-based Memory Network (Sukhbaatar et al., 2015; Gulcehre et al., 2018) to select relevant words according to the initial image dynamically. MirrorGAN (Qiao et al., 2019) exploits the idea of learning text-to-image generation by redescription to progressively enhance the diversity and

semantic consistency of the generated images. To compensate for the lack of semantic information, especially spatial constraints, provided by a single textual input, Obj-GAN (Li et al., 2020) and ControlGAN (Li et al., 2019) use additional constraints that supply spatial information, such as layout or mask, as a supplement to the textual semantics. To fuse text and image information more effectively, DF-GAN (Tao et al., 2022) concatenates multiple Deep Fusion Blocks and operates affine transformations on the image feature maps for text–image fusion. SSA-GAN (Liao et al., 2022) effectively fuses the text and image features by predicting semantic masks separately in each affine block to guide the learned text-adaptive affine transformation. RAT-GAN (Ye et al., 2022) connects all the conditional affine transformation blocks with recurrent connections for explicitly fitting the temporal consistency. LAFITE (Zhou et al., 2022) and GALIP (Tao et al., 2023) introduce CLIP-based loss, CLIP-based discriminator and CLIP-empowered generator for text-to-image training and show significant improvements. Some studies attempt to further improve performance by scaling up the GAN model. StyleGAN-T (Sauer et al., 2023) successfully scaling to the large-scale text-to-image task in lower resolutions. GigaGAN (Kang et al., 2023) efficiently extends the capacity of the generator through retaining a set of filters and employing sample-specific linear combinations. We consider five aspects: Multiple Stage, Attention Mechanism, Multi-Granularity, Additional Control, Deep Fusion which are applied to GAN-based text-to-image generation methods and illustrate differences in characteristics of above methods in Table 1, where GigaGAN and StyleGAN-T are used for high-resource environments, the other methods are used for low-resource environments.

Large Pre-training Models Large pre-training models have shown powerful ability in text-to-image synthesis tasks. Autoregressive models, e.g., DALL-E (Ramesh et al., 2021) and CogView (Ding et al., 2021), achieve text-to-image synthesis based on a pre-trained unidirectional transformer that autoregressively models the text and image tokens together as a single stream of data. Cogview2 (Ding et al., 2022) proposes a solution based on hierarchical transformers and local parallel autoregressive generation for faster image generation. Parti-350M and Parti-20B (Yu et al., 2022) use the powerful image tokenizer to encode images as sequences of discrete tokens and take advantage of its ability to reconstruct such image token sequences of visually diverse images. Diffusion models (Sohl-Dickstein et al., 2015; Dhariwal and Nichol, 2021; Ho et al., 2020; Nichol and Dhariwal, 2021) also show impressive performance on text-to-image synthesis. VQ-Diffusion (Gu et al., 2022) is based on a vector quantized variational autoencoder (VQ-VAE) (Van Den Oord et al., 2017) whose latent space is modeled by a conditional variant of the recently developed Denoising Diffusion Probabilistic Model (DDPM) to eliminate the unidirectional bias and avoid the accumulation of errors. DALL·E 2 (Ramesh et al., 2022) adopts a CLIP decoder incorporating a diffusion model. Latent Diffusion Models (LDM) (Rombach et al., 2022a) propose a method for performing the diffusion process on the latent space, which can greatly reduce the computational complexity. Inspired by the Classifier-Free Diffusion Guidance (Ho and Salimans, 2022), GLIDE (Nichol et al., 2021) further extends the model scale to enable more powerful image generation and drive image editing. Imagen (Saharia et al., 2022) introduces dynamic thresholding and Efficient U-Net to generate more photo-realistic and detailed images.

3. MFAE-GAN

In order to obtain high-quality generation results based on limited computational cost through MFAE-GAN, we propose (i) CFE and FFE modules to sample and refine different granularity features with different strategies, (ii) an MFA-BN module, which provides denser semantic constraints than textual semantic information when adding fine-grained features, (iii) a GSP module to supplement global semantics and to maintain semantic integrity through fusing sentence embeddings. In the rest of this section, we will introduce the design of each part of MFAE-GAN in detail.

3.1. Model overview

A schematic diagram of our MFAE-GAN architecture is shown in Fig. 2. MFAE-GAN has a text encoder that is pre-trained for COCO, CUB and CC series datasets, respectively, by the same strategy as AttnGAN (Xu et al., 2018) which minimizes the Deep Attentional Multimodal Similarity Model (DAMSM) loss to extract text descriptions into word embeddings and sentence embeddings to ensure better coordination between the training processes of the text encoder and the generator, a generator consisting of 4 CFE modules, 2 FFE modules and 2 GSP modules, which can sample and refine coarse-grained features at early stages and fine-grained features at later stages respectively to avoid refinement quality's degradation, and a discriminator that is used to promote the network to synthesize images that are semantically consistent with the given text and close to an actual image. Our model takes text embeddings and a noise vector $z \in \mathbb{R}^{100}$ sampled from a Gaussian distribution as input and can finally generate 256×256 resolution images.

3.2. Coarse-grained feature enhancement

CFE focuses on independently sampling coarse-grained global features related to category information (e.g., pose and shape) in the early generation stages, reducing the interference of fine-grained local features that have not yet been initialized to coarse-grained global features, and further optimizing the representations of coarse-grained features by deep fusion of textual supervision information. The architecture of the Coarse-grained Feature Enhancement (CFE) module is shown in Fig. 3. Since the coarse-grained features are often distributed at global scales, we achieve naturally sampling of the coarse-grained global features representations through word-level channel attention that shares weights within each channel. Specifically, the channel attention module takes word embeddings s and hidden image features $v \in \mathbb{R}^{C\times(H*W)}$ as input, where H and W define the height and width of the feature map at current CFE module. The word embedding w is first converted to an underlying common semantic space of visual features as $i\tilde{w}$. It expresses the correlation between feature channels and words by calculating the channel-wise attention matrix $m = v\tilde{w}$. Then, the channel attention module aggregates weight values in a channel-wise attention matrix α as

$$\alpha_{i,j} = \frac{\exp\left(m_{i,j}\right)}{\sum_{k=0}^{l-1}\exp\left(m_{i,k}\right)}, \tag{1}$$

where $\alpha_{i,j}$ indicates the weight that the model attends between the ith channel in the visual features v and the jth word in the sentence T. Finally, we sample the coarse-grained feature representations across all spatial locations in each channel as $f^{\alpha} = \alpha(i\tilde{w})^{T}$. Following SSA-GAN (Liao et al., 2022), we utilize two convolutional layers and a ReLU activation layer to map coarse-grained feature representations to coarse-grained feature sketches $i_c \in \mathbb{R}^{(H*W)}$ that provide explicit spatial information supervision. Unlike previous work which uses pre-labeled sketch data to pre-train the network to process or predict sketches (Zhang et al., 2019; Lu et al., 2017; Toda et al., 2022; Liu et al., 2021b; Koley et al., 2023), our strategy of multi-granularity feature mapping module is trained under the weakly supervised setting jointly with the whole network without specific loss function to guide its learning process nor additional mask annotation. This strategy reduces the requirement for computational resources while obtaining available spatial constraints.

In the early stage, the sparse semantics provided by the text description can help the network to generate category-related coarse-grained visual semantics. Therefore, we adopt two Multi-Layer Perceptrons (MLPs) conditioned on given sentence embedding to learn modulation parameters γ_{cfe} and β_{cfe}, respectively:

$$\gamma_{cfe} = MLP_{\gamma}(s_{ca}), \quad \beta_{cfe} = MLP_{\beta}(s_{ca}), \tag{2}$$

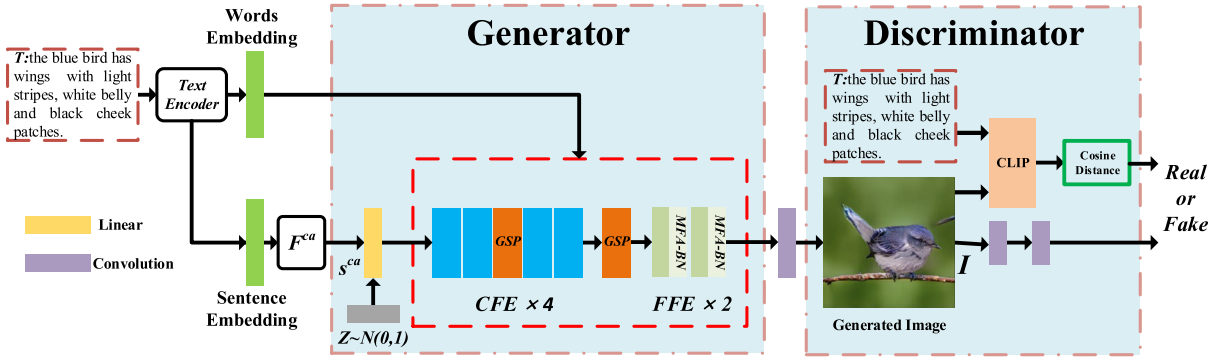


Fig. 2. The architecture of the proposed MFAE-GAN. The MFAE-GAN consists of 4 CFE, 2 FFE modules to sample and refine the different granularity features separately and complement the global semantics through MFA-BN module. Denser constraints and image features are fused using the MFA-BN module in FFE.

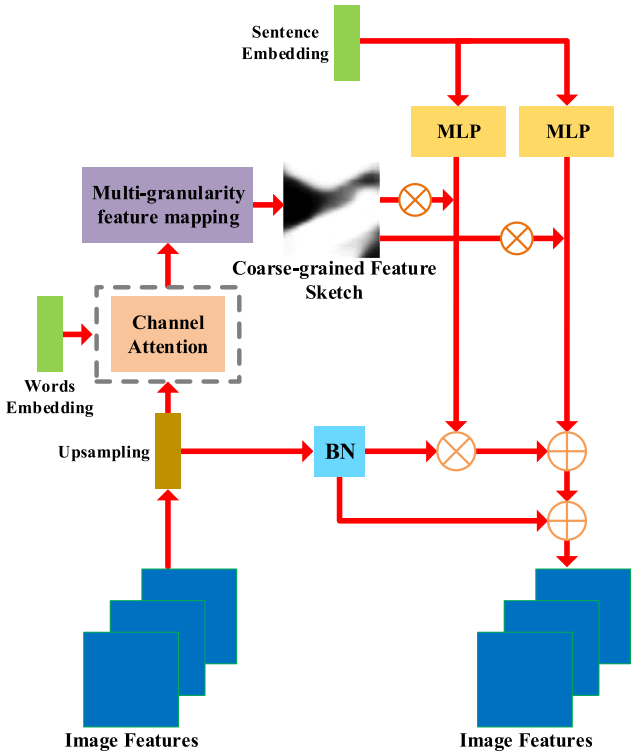


Fig. 3. The architecture of CFE Module. CFE module samples coarse-grained global features representations through word-level channel attention and maps to coarse-grained feature sketches that are used for affine transformation of the Condition Batch Normalization to provide explicit spatial information supervision.

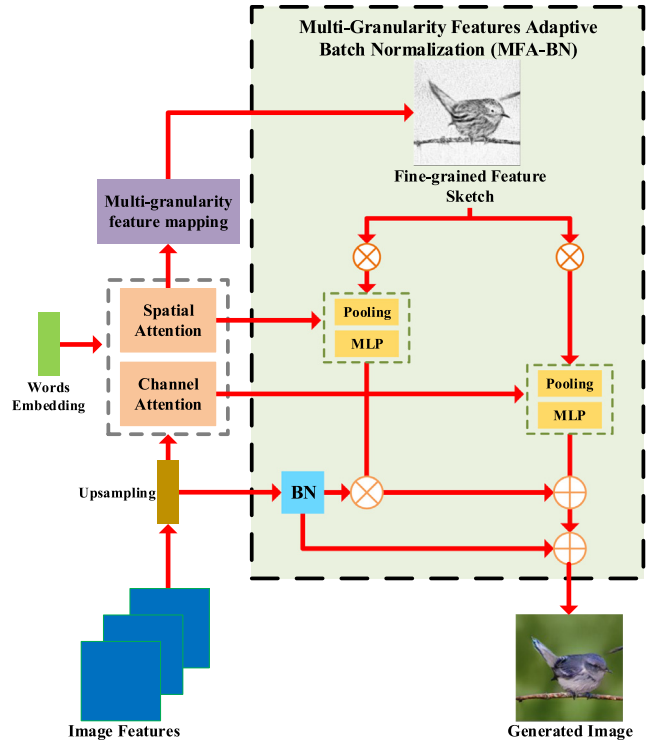


Fig. 4. The architecture of FFE Module. FFE adds additional word-level spatial attention to sample fine-grained representation and maps them to fine-grained feature sketches to reflect instance-level details. MFA-BN module in FFE fuses text–image at multi-granularity to provide denser constraint information of semantic attributes than text descriptions.

which are used for affine transformation of the Condition Batch Normalization (Dumoulin et al., 2016) as follows:

$$\tilde{v} = i_{c(h,w)}\left(\gamma_{cfe}(s_{ca})\hat{v} + \beta_{cfe}(s_{ca})\right), \tag{3}$$

where $\hat{v}$ is the batch normalized image features.

3.3. Fine-grained feature enhancement

The architecture of the Fine-grained Feature Enhancement (FFE) module is shown in Fig. 4. In the later stages of generation, the main task of the network is to add instance-level details (e.g., texture and color) to the local areas of generated images. Since CFE has refined coarse-grained features from the global perspective without considering independent spatial locations, we first add an extra word-level spatial attention module to sample fine-grained feature representations. The word-level spatial attention shares weights among corresponding local locations of multiple channels, which can help the generator disentangle fine-grained visual attributes through prioritized attention to independent spatial locations to enhance fine-grained local instance-level details. Specifically, for the ith image sub-region v_i represented by a column along the channel direction of image features v, we calculate the attention weight between the word embedding w and v_i denoted by matrix θ as follows:

$$\theta_{i,j} = \frac{\exp\left(e_{i,j}\right)}{\sum_{k=0}^{l-1}\exp\left(e_{i,k}\right)}, \tag{4}$$

where $e = v_i\tilde{w}$. The fine-grained feature representations can be dynamically represented as $f^{\theta} = \theta\,(\tilde{w})^T$. In FFE, we let f^{θ} instead of the coarse-grained feature representations from the channel attention module map to the sketches to provide explicit spatial information supervision related to fine-grained features. Then, we also propose a fusion module

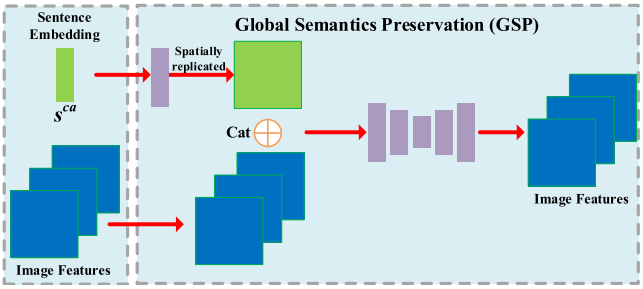


Fig. 5. The architecture of GSP Module. Global Semantics Preservation (GSP) module fuses sentence embedding and image feature maps in a unified visual space to complement the global semantics.

called Multi-granularity Features Adaptive Batch Normalization (MFA-BN) that uses the multi-granularity feature representations instead of text embedding to provide denser semantic constraints in FFE.

MFA-BN Module The modulation parameters learned from the given sentence embeddings γ_{cfe} and β_{cfe} remain two limitations when adding fine-grained details at the later stages. First, constraints conditioned on text embedding can only provide sparse semantic information, such as the category information of the object in the image (e.g., "A zebra stands on a pathway near grass".) or an incomplete description of the object-related attributes (e.g., "A small bird with a black head and wings"). Second, the modulation parameters work on the multi-granularity features equally. Ideally, we expect the modulation can encourage the network to focus on refining fine-grained features while preserving the representations of coarse-grained features. For this purpose, fine-grained and coarse-grained feature representations are processed by pooling to eliminate redundant information and then used to produce the modulation parameters γ_{ffe} and β_{ffe} respectively using MLPs.

$$\begin{aligned}\gamma_{ffe} &= MLP_{\gamma}(Pooling(f^{\theta})),\\ \beta_{ffe} &= MLP_{\beta}(Pooling(f^{\alpha})),\end{aligned} \tag{5}$$

and are multiplied and added to the normalized activation element-wise as follows:

$$\tilde{v} = i_{f(h,w)}\left(\gamma_{ffe}(f^{\theta})\hat{v} + \beta_{ffe}(f^{\alpha})\right), \tag{6}$$

where $i_{f(h,w)}$ is fine-grained feature sketches mapped from fine-grained feature representations.

The effectiveness of the MFA-BN module is essentially from a disentanglement and further enhancement of the sampled multi-granularity feature representations. The modulation parameters adaptively learned from multi-granularity features can provide denser constraint information of semantic attributes than text descriptions in the fusion process. The fine-grained feature sketches explicitly illustrate the spatial distribution information of different granularity features. Thus, the MFA-BN module enables text–image fusion at multi-granularity.

3.4. Global semantics preservation

When the CFE and FFE modules continuously sample and refine specific granularity features, with the increasing number of sampling modules, the network may ignore the initialization of other granularity features, which leads to the loss of global semantics. We address this issue by complementing the global semantics while sampling features at different granularities. The architecture of the Global Semantics Preservation (GSP) module is shown in Fig. 5. Specifically, the GSP module first converts sentence embedding s_{ca} to visual features $\tilde{s_{ca}}$ and connects intermediate image feature maps v along the channel direction. A concise and practical encoder–decoder network further fuses the concatenated features to complement the global semantics and maintain semantic integrity as follows:

$$v' = Encoder\_Decoder(v, s_{ca}). \tag{7}$$

The GSP module does not change the resolution of the image, which ensures that the CFE and FFE modules still dominate the image refinement process.

3.5. Objective functions

We adopt the one-way discriminator associated with the Matching-Aware zero-centered Gradient Penalty (MA-GP) proposed in DF-GAN (Tao et al., 2022). In addition, we add a pre-trained ViT-Base/32 CLIP model from CLIPdraw (Frans et al., 2022) as an evaluation index to maximize the similarity between the given natural language description and the generated image. We also add the CA loss to our framework to improve the performance of the network.

Generator objective The adversarial loss of generator is defined as follows:

$$\mathcal{L}_{adv}^{G} = -\mathbb{E}_{\hat{x}\sim p_G}\log D_i(\hat{x}, s), \tag{8}$$

where s is the sentence embedding, $\hat{x}$ is the generated image from the model distribution p_G, and $D(\cdot)$ calculates the matching degrees between the image and the sentence. Following StackGAN (Zhang et al., 2017), the CA loss is defined as the Kullback–Leibler (KL) divergence between the conditional Gaussian distribution and the standard Gaussian distribution of the given sentence embedding, which is calculated as follows:

$$\mathcal{L}_{CA} = D_{KL}\left(\mathcal{N}\left(\mu(s), \sum(s)\right) \parallel \mathcal{N}(0, I)\right), \tag{9}$$

where $\mathcal{N}\left(\mu(s), \sum(s)\right)$ is an independent Gaussian distribution, the mean $\mu(s)$ and diagonal covariance matrix $\sum(s)$ are learned jointly with other parameters of the network. Furthermore, the text–image similarity is measured via the cosine distance between the CLIP-encoded image and text representation. The $\mathcal{L}_{CLIP}$ is calculated as follows:

$$\mathcal{L}_{CLIP} = -CosineSim(CLIP\_T, CLIP\_Img), \tag{10}$$

where $CLIP\_T$, $CLIP\_Img$ are the CLIP-encoded text and the image representation. The final objective function of the generator networks is

$$\mathcal{L}_G = \mathcal{L}_{adv}^{G} + \lambda\mathcal{L}_{CA} + \lambda_{clip}\mathcal{L}_{CLIP}, \tag{11}$$

where λ and λ_{clip} are regularization parameters to balance different terms.

Discriminator objective The objective function of the discriminator $\mathcal{L}_D$ which is associated with the MA-GP loss can be defined as follows:

$$\begin{aligned}\mathcal{L}_D = {} & E_{x\sim p_{\text{data}}}\left[\max(0, 1 - D(x, s))\right]\\ & + \frac{1}{2}E_{\hat{x}\sim p_G}[\max(0, 1 + D(\hat{x}, s))]\\ & + \frac{1}{2}E_{x\sim p_{\text{data}}}\left[\max(0, 1 + D(x, \hat{s}))\right]\\ & + \lambda_{MA}E_{x\sim p_{\text{data}}}\left[\left(\left\|\nabla_x D(x, s)\right\|_2\right.\right.\\ & \left.\left. + \left\|\nabla_s D(x, s)\right\|_2\right)^{p}\right],\end{aligned} \tag{12}$$

where x is the real image from the real image distribution p_{data}, $\hat{s}$ is a mismatched sentence embedding to paired images. λ_{MA} is regularization parameter for MA-GP loss.

4. Experiment

We conduct extensive experiments to evaluate the MFAE-GAN. In this section, we first introduce the datasets, training details, and evaluation metrics. Then, we compare the performance with the state-of-the-art GAN-based methods on COCO and CUB datasets. We also

compare our model, which is pre-trained on a union of CC3M and CC12M. with large pre-trained models such as autoregressive and diffusion models under Zero-shot setting through a Zero-shot FID_{30k} evaluation on MS COCO. Finally, we present a series of ablation studies on the critical components of MFAE-GAN to validate their effectiveness.

Datasets The CUB dataset has 200 categories with 11788 images (8,855 images for training and 2,933 images for testing) of birds. Each image in the CUB dataset has a single object but contains rich shapes, colors, and posture details, which are always employed to evaluate the ability of a network to synthesize fine-grained features. 10 text descriptions are used per image. The COCO dataset consists of 123287 images (82783 images for training and 40504 images for testing) with 5 sentence annotations. Compared to CUB, the images in COCO demonstrate more complex visual scenes containing multiple objects, making it more challenging for text-to-image generation tasks. The CC series datasets (CC3M and CC12M) consist of text–image pairs automatically collected from the Internet and processed by algorithms. Due to data decay over time, the two datasets currently contain approximately 13 million valid pairs. We randomly exclude 10,000 image data pairs from each dataset for qualitative testing and the rest are used to train the model for quantitative tests under the zero-shot setting.

Evaluation Metrics We quantitatively measure the performance of our MFAE-GAN in terms of Inception Score (IS) (Salimans et al., 2016) and Fréchet Inception Distance (FID) (Heusel et al., 2017) to evaluate whether the generated results are close to the realistic image when pre-trained on COCO or CUB datasets. Specifically, we obtain IS by employing a pre-trained Inception-v3 network (Szegedy et al., 2016) to predict the class label probabilities and compute the KL-divergence between the marginal class distribution and the conditional class distribution. A larger IS signifies that the generated images contain richer and more discriminative semantic information. FID computes the Fréchet distance between the synthetic and real images based on the feature map output from the pre-prepared Inception-v3 network. A lower FID score implies a closer distance between the generated image distribution and real image distribution and therefore means the model performs better when synthesizing photo-realistic images. We follow previous work and set the input batch to 64 during testing.

We also adopt the R-precision and CLIPSIM (CS) (Wu et al., 2022) to evaluate the text–image semantic consistency. Specifically, R-precision assesses the semantic consistency between the synthetic image and the given text description. We utilize pre-trained DAMSM (Xu et al., 2018) to calculate the cosine similarities between the global image vector and 100 competitor global sentence vectors which consist of one ground truth (i.e., R = 1) and 99 randomly selected mismatching descriptions to quantify the image–text semantic similarity. CLIPSIM uses the CLIP model to encode the given image and text descriptions as embedding vectors and calculate the embedding similarity matrix between input text and the generated image. We set the input batch to 64 when calculating the CLIPSIM and R-precision.

Additionally, we adopt Zero-shot FID_{30k} based on 30,000 images from COCO test set to evaluate visual quality under Zero-shot setting when pre-trained on CC series datasets. The calculation process and settings are consistent with FID scores.

Finally, we assess the generation efficiency of different models by counting the single 256 × 256 image generation speed on a single 3090 GPU, where the diffusion-based model obtains the images by sampling over 40 time-steps.

Implementation Details During training MFAE-GAN, the batch size is set to 64 on an Nvidia RTX 3090 GPU. The generator and discriminator are trained alternately by minimizing both the generator loss $\mathcal{L}_G$ and discriminator loss $\mathcal{L}_D$. Adam (Kingma and Ba, 2014) with $\beta_1 = 0.1$ and $\beta_2 = 0.9$ is used for network optimization. The learning rate is set to 0.0001 for the generator and 0.0004 for the discriminator respectively according to TTUR (Heusel et al., 2017). The model is trained for 600 epochs on the CUB dataset, 1000 on the COCO dataset and 15 epochs on the union of CC3M and CC12M datasets.

Table 2
Quantitative comparison of state-of-the-art methods and MFAE-GAN on the test set of CUB and COCO datasets.

Methods	CUB		COCO	
	IS↑	FID↓	IS↑	FID↓
DM-GAN (Zhu et al., 2019)	4.75 ± .007	–	30.49 ± .56	32.64
AttnGAN (Xu et al., 2018)	4.36 ± .07	22.37	25.87 ±.47	35.42
DAE-GAN (Ruan et al., 2021)	4.42 ± .04	15.19	35.08 ± 1.16	28.12
MirrorGAN (Qiao et al., 2019)	4.56 ± 0.05	–	26.47 ± 0.41	–
DF-GAN (Tao et al., 2022)	5.04 ± .04	14.81	–	19.32
SSA-GAN (Liao et al., 2022)	5.17 ± .08	15.61	–	19.37
RAT-GAN (Ye et al., 2022)	5.36 ± .20	13.91	–	14.60
LAFITE (Zhou et al., 2022)	–	14.58	–	8.21
GALIP (Tao et al., 2023)	–	10.08	–	**5.85**
MFAE-GAN	**6.27 ± .31**	**8.34**	**40.03 ± .62**	6.34

Table 3
Text–Image consistency comparison of state-of-the-art methods and MFAE-GAN on the test set of CUB and COCO datasets.

Methods	CUB		COCO	
	R-precision↑	CS↑	R-precision↑	CS↑
DM-GAN (Zhu et al., 2019)	72.37	–	88.56	–
AttnGAN (Xu et al., 2018)	67.82	–	72.31	–
DAE-GAN (Ruan et al., 2021)	85.45	–	92.61	–
MirrorGAN (Qiao et al., 2019)	56.67	–	74.52	–
DF-GAN (Tao et al., 2022)	–	0.2920	–	0.2972
SSA-GAN (Liao et al., 2022)	75.9	–	90.6	–
RAT-GAN (Ye et al., 2022)	81.6	–	87.4	–
LAFITE (Zhou et al., 2022)	–	0.3125	–	0.3335
GALIP (Tao et al., 2023)	–	0.3164	–	0.3338
MFAE-GAN	**86.43**	**0.3309**	**93.93**	**0.3379**

4.1. Quantitative results

We first compare the image fidelity of our proposed MFAE-GAN with several state-of-the-art GAN-based methods. For a fair comparison, all results are taken from the data provided in paper or obtained by testing the full source code when generating 256 × 256 resolution images. Data that not provided in the original paper or without the complete test code are marked as "-". The overall results are summarized in Table 2. Compared with other leading models, our model improves IS score to 6.27 and decreases FID score to 8.34 on the CUB dataset. On the COCO dataset, MFAE-GAN improves IS score to 40.03. We also achieve competitive results compared to GALIP (Tao et al., 2023) in terms of FID scores (6.34 *v.s.* 5.85) while significantly outperforming the other recent methods. The quantitative evaluation results demonstrate the effectiveness of sampling and refining the different granularity features separately in generating more realistic images, both for a single object with rich detailed attributes and complex scenes with multiple objects.

To evaluate text–image semantic consistency, we compare the R-precision and CLIPSIM metrics of MFAE-GAN on CUB and COCO datasets with state-of-the-art methods. For a fair comparison, we use the raw data given in the papers of these methods. The results are shown in Table 3. MFAE-GAN improves the R-precision to 86.43 and the CLIPSIM to 0.3309 on the CUB dataset. Furthermore, MFAE-GAN improves the R-precision to 93.93 and the CLIPSIM to 0.3379 on the COCO dataset. The quantitative comparison shows that MFAE-GAN can effectively ensure higher semantic consistency between the synthesized results and the given text.

Moreover, we pre-train our proposed MFAE-GAN on CC series datasets and conduct a Zero-shot FID_{30k} evaluation on MS COCO to quantitatively compare the generalization ability of our model and show the results in Table 4. Although these models are often trained even with hundreds of times the parameters of our models, MFAE achieves competitive performance with much smaller model parameters (0.21B trainable parameters) and data (13 m), such as *v.s.* GigaGAN (10.14 *v.s.* 9.09) or *v.s.* SD-v1.5 (10.14 *v.s.* 9.62) . Our model also

Table 4
Quantitative comparison of large pre-trained models and MFAE-GAN pre-trained on CC series datasets with the parameter sizes, amount of training data, performance and inference speeds under the Zero-shot setting on COCO test set at a native 256 × 256 resolution using a 3090 GPU.

Methods	Type	Param [B]	Data size [B]	Zero-shot FID_{30k}	Speed [s]
DALL-E (Ramesh et al., 2021)	AR	12	1.54	27.5	–
Parti-350M (Yu et al., 2022)	AR	0.35	0.8	14.1	5.46
Parti-20B (Yu et al., 2022)	AR	20	0.8	7.23	–
Cogview (Ding et al., 2021)	AR	4	0.03	27.1	–
Cogview2 (Ding et al., 2022)	AR	6	0.03	24	2.4
LDM (Rombach et al., 2022a)	DF	1.45	0.27	12.63	10.4
SD-v1.5	DF	0.9	3.16	9.12	4.7
DALL·E 2 (Ramesh et al., 2022)	DF	5.5	5.63	10.39	4.1
Imagen (Saharia et al., 2022)	DF	7.9	15.36	7.24	9.10
StyleGAN-T (Sauer et al., 2023)	GAN	1.1	0.25	13.90	0.18
GigaGAN (Kang et al., 2023)	GAN	1	0.98	9.09	0.33
GALIP (Tao et al., 2023)	GAN	0.31	0.012	12.64	0.08
LAFITE (Zhou et al., 2022)	GAN	0.23	0.012	26.94	0.11
MFAE-GAN	GAN	0.21	0.013	10.14	0.02

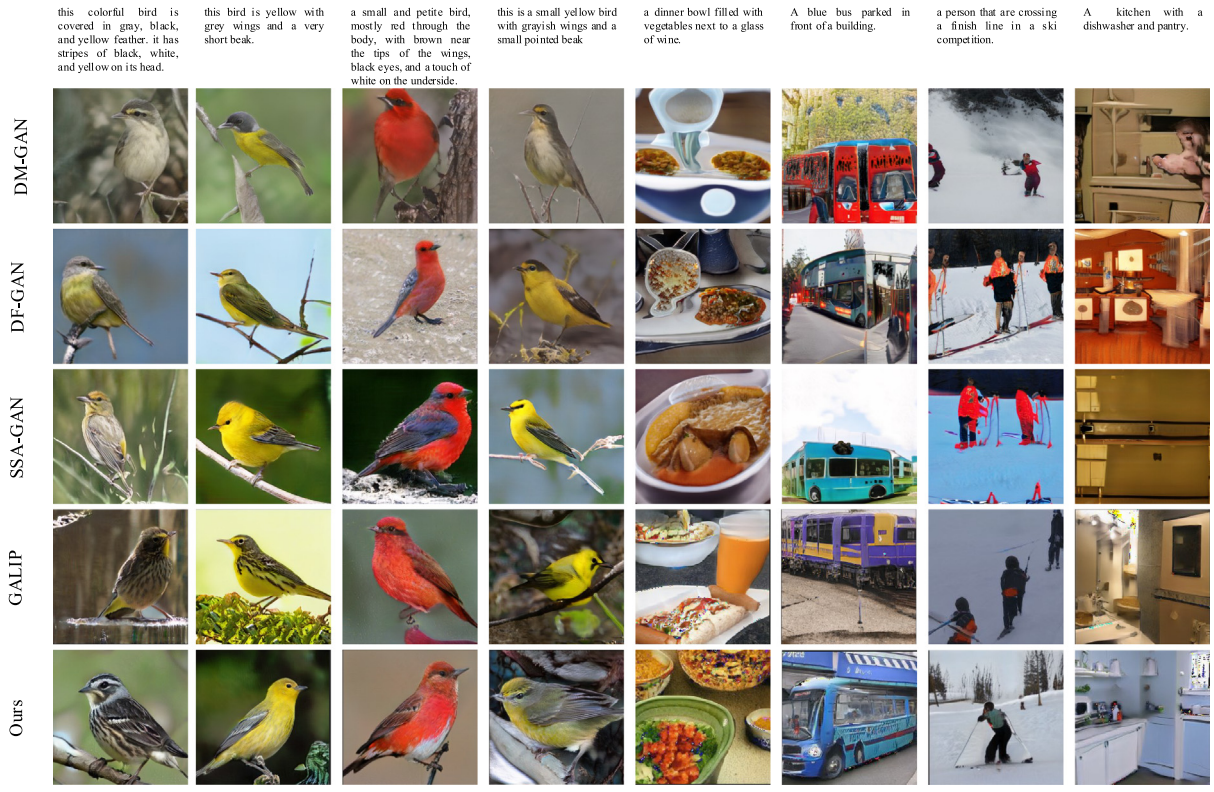


Fig. 6. Qualitative comparison of state-of-the-art GAN-based methods (AttnGAN, DM-GAN, DF-GAN, SSA-GAN) and our proposed MFAE-GAN on the CUB (1st–4th columns) and COCO datasets (5th–8th columns). The input text descriptions are given in the first row.

performs best compared to methods with similar model parameters and data sizes, such as GALIP and LAFIFE (10.14 *v.s.* 12.64 *v.s.* 26.94). Furthermore, MFAE-GAN significantly improves training and generation efficiency with limited computational cost requirements. It takes only 0.02 s to generate a 256 × 256 resolution image in real-time on a single 3090 GPU, which has an advantage over the GAN-based StyleGAN-T (0.02*s* *v.s.* 0.18*s*) and GigaGAN (0.02*s* *v.s.* 0.23*s*) due to the smaller model size, while the most important open-source large-scale pre-training model, SD-1.5 (Rombach et al., 2022b), takes more than 4.7 s.

4.2. Qualitative results

Visual Quality To evaluate the visual quality of generated images, we first show subjective comparisons between some state-of-the-art models and our proposed MFAE-GAN. For the results on the CUB dataset, as shown in the first 4 columns in Fig. 6, our MFAE-GAN significantly improves the anti-deformation compared to AttnGAN and DM-GAN, which indicates that CFE can enhance the expression of category-related coarse-grained features. In addition, when expressing fine-grained semantics, such as the texture on a bird's wing, our method can express complex and natural variations of color and texture, while SSA-GAN and DF-GAN can only synthesize fine-grained features into the regions with monotonous colors, and the textures within the regions are bent and deformed (1st, 2nd and 3rd columns). Finally, we also find that the generated results of MFAE-GAN have better consistency with the text semantics. For example, in the 3rd column, only MFAE-GAN accurately generates the image features corresponding to the text description of *"a touch of white on the underside"*. This indicates that the FFE module performs better in capturing details. For the results on

Fig. 7. The generated results when changing the color descriptions of the input text on the CUB dataset. MFAE-GAN can ensure the consistency of text–image fine-grained semantics while maintaining the diversity of coarse-grained features.

the COCO dataset, as shown in the last 4 columns in Fig. 6, the images generated by MFAE-GAN contain almost all key semantic objects in the text description, such as the text descriptions *"a dinner bowl", "a glass of wine"* in the 5th column, *"a bus", "a building"* in the 6th column and *"dishwasher", "pantry"* in the 8th column. In addition, these objects generated by MFAE-GAN can be recognized effectively and contain rich detailed features, such as the decorative patterns on the bus and the folds on the skier's clothes. In contrast, the results generated by other methods generally suffer from the problem of missing key objects and lacking semantic details.

Text–Image Consistency Our method can explicitly establish connections between text semantics and multi-granularity visual features through the CFE and FFE modules, allowing us to precisely control image generation from different text descriptions. We replace the color-related words in the input text description and show the generated results in Fig. 7. The results accurately reflect the fine-grained visual semantics consistent with the corresponding description. In contrast, the coarse-grained visual semantics (e.g., pose and backgrounds) of generated images maintaining the diversity, which is a natural requirement for the generation models; the bird in the three images of Fig. 7 looks like subspecies with different colors of the same category. This qualitative result shows that MFAE-GAN effectively disentangles the visual attributes of different granularities.

Intermediate of Multi-granularity Features To better understand the intermediate changes and roles of sampled multi-granularity features in MFAE-GAN, we show the feature sketches mapped by our method (2nd row) at different stages in Fig. 8 and compare with the method which directly predicts the sketches without disentangling multi-granularity visual attributes (Liao et al., 2022) (1st row). As the generation stage progresses, mapped sketches from MFAE-GAN express richer semantics. Specifically, compared with the prediction results in 1st row, the mapped sketches of the CFE module have more precise boundaries to represent the category-related coarse-grained features (in red bounding box). For example, in the 2nd and 3rd columns, the CFE samples the bird's body position and the demarcation of the background, and in the 4th column, the CFE samples the distribution of the bird's wings and beak. These coarse-grained features sampled by CFE show a better correspondence with the final generated results. The mapped sketches from FFE contain richer and more detailed textures (in blue bounding box) that can guide the distribution of fine-grained features spatially. For example, in the 5th columns, the mapped sketches sampled by FFE show complex textures on the wings with a natural and rich variation, while the result sampled by the method that directly predicts the sketches only shows rough lines. The above analysis shows that MFAE-GAN can sample and refine effectively for specific granularity features at different stages of the refinement process to guide image generation.

Table 5
Ablation performance of each component on CUB and COCO about IS, FID and CS.

Methods	CUB			COCO		
	IS↑	FID↓	CS↑	IS↑	FID↓	CS↑
w/o CFE	5.32	10.14	0.2986	37.18	9.31	0.3194
w/o FFE	5.02	11.51	0.3142	35.09	10.05	0.3289
w/o MFA-BN	5.39	10.01	0.3091	36.21	9.24	0.3134
w/o GSP	5.67	9.39	0.3217	37.13	8.24	0.3313
MFAE-GAN	**6.27**	**8.34**	**0.3309**	**40.03**	**6.34**	**0.3379**

Table 6
Ablation studies of CFE and FFE step on CUB and COCO about IS, FID and CS.

ID	Step		CUB			COCO		
	CFE	FFE	IS↑	FID↓	CS↑	IS↑	FID↓	CS↑
0	1	2	3.77	19.44	0.2817	25.48	22.31	0.2941
1	2	2	4.44	13.28	0.2972	30.23	13.45	0.2991
2	3	2	5.19	10.43	0.3031	33.47	10.21	0.3124
3	4	1	5.77	8.94	0.3102	36.77	8.81	0.3193
4 (Ours)	4	2	**6.27**	**8.34**	**0.3309**	**40.03**	**6.34**	**0.3379**

Generalization Ability in Different Styles The COCO and CUB datasets contain only images of a single style. To verify the generalization ability of MFAE-GAN between different styles, we pre-train the network using the CC3M and CC12M datasets containing image data of multiple styles and compare it with SD-1.5, which is pre-trained based on 3.16 billion text–image pairs, on a randomly selected test set of CC3M and CC12M. As the results in Fig. 9, the larger model parameters and training data ensure that SD-1.5 has the advantages of image details and style diversity. However, thanks to the sampling of multi-granularity features and text–image semantic alignment under spatial constraints, MFAE-GAN achieves competitive image quality and has an advantage in semantic consistency, e.g., in the 4th and 6th columns of Fig. 9, MFAE-GAN correctly generates the image semantics corresponding to "starry night" and "white background".

4.3. Ablation studies

The overall experiment results have proved the superiority of our proposed MFAE-GAN. In this section, we further verify the effectiveness of each component on CUB and COCO datasets. We first design a baseline module referring to previous text-to-image methods (Zhang et al., 2017, 2018), which connects text embedding and image features directly along the channel direction and fuses the features with 2-layer convolutions. We replace the CFE and FFE modules separately with the same number of baseline modules to verify the effectiveness. Moreover, we verify the effectiveness of the MFA-BN module in FFE by replacing it with the same fusion strategy as in CFE. Finally, we eliminate the input of additional text embeddings as a complement to the global

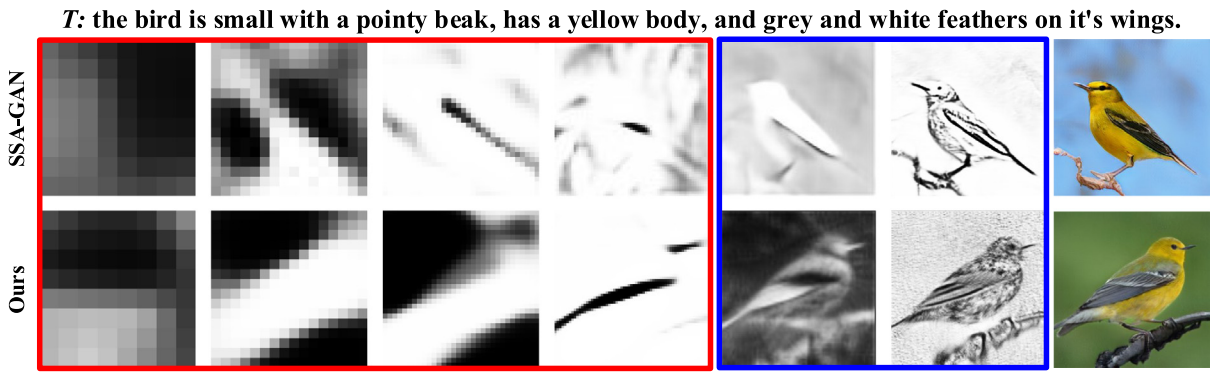


Fig. 8. Examples of feature sketches mapped by MFAE-GAN and the method that directly predicts the sketches without disentangling multi-granularity visual attributes at early stages (in red bounding box) and later stages (in blue bounding box) with the same text description as input.

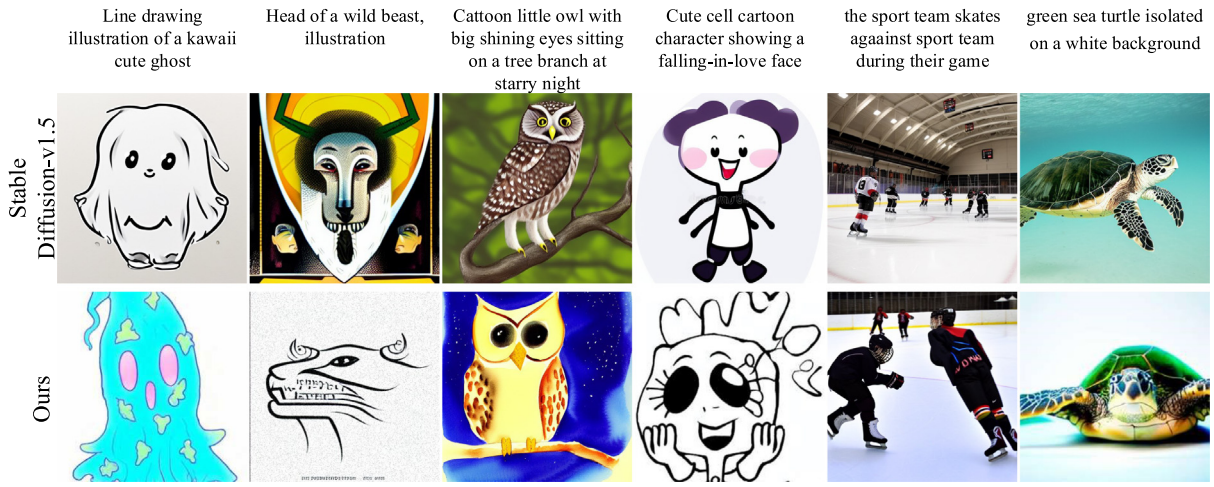


Fig. 9. Examples of different styles images synthesized by Stable Diffusion-v1.5 which is based on LDM (Rombach et al., 2022a) and our proposed MFAE-GAN conditioned on text descriptions from the test set of CC3M and CC12M datasets. MFAE-GAN achieves better semantic alignment while having competitive image quality.

semantics in the GSP module to verify the effectiveness of this strategy. Corresponding results are illustrated in Table 5. According to the results, we can observe model performance decline to varying degrees when removing CFE, FFE, MFA-BN and GSP separately from MFAE-GAN. The ablation study shows that as the generation process advances, the CFE and FFE modules achieve sampling and refinement of different granularity features via incorporating the spatial and channel attention modules, which helps to avoid the drop of the network's refinement quality. MFA-BN can further achieve multi-granularity feature fusion to generate photo-realistic images with more explicit category information and richer instance-level details. GSP can avoid losing global semantics caused by continuously sampling specific granularity features.

By adjusting the number of CFE and FFE modules, we also show the effect of different steps for sampling multi-granularity feature on the generated results. The final generated results of the different combination strategies are resized to 256 × 256 when calculating the quantitative metrics. Specifically, in ID 0~4 of Table 6, the IS, FID and CS scores of the generated images indicate better performance as the CFE and FFE modules were stacked. In particular, when more FFE modules are stacked in ID 3 and ID 4, MFAE-GAN acquires better text–image semantics alignment because the FFE module samples and refines the expression of fine-grained features.

4.4. Limitations

While MFAE-GAN shows effectiveness improvement in text-to-image synthesis, several limitations remain that should be improved in future work. Firstly, the dataset we use for pre-training is much smaller than other large models, which restricts the ability of the model to transfer between generating tasks of different styles. Secondly, subject to the computational cost, we only use a small number of CFE, FFE and GSP modules in the generation process. Stacking more modules may benefit in expressing semantic information at a specific granularity. Thirdly, when the model size is expanded and the training data is adequate, replacing the LSTM-based text encoder with CLIM or T5, which enables powerful cross-modal learning and understanding capabilities, may improve the performance.

5. Conclusion

This paper proposes a novel framework, Multi-granularity Feature-Aware Enhancement GAN (MFAE-GAN), for text-to-image tasks to generate images containing explicit category information and rich instance-level details. The framework has three core modules, where Coarse-grained Feature Enhancement (CFE) module and Fine-grained Feature Enhancement (FFE) module can sample and refine the different granularity features separately and impose constraints spatially based on multi-granularity feature sketches, Global Semantics Preservation (GSP) module is used to preserve semantic integrity during continuous sampling for specific granularity features. We also propose a Multi-granularity Features Adaptive Batch Normalization (MFA-BN) module, which provides denser semantic constraints than textual semantic information when adding fine-grained features and achieves text–image fusion at multi-granularity. Extensive experimental results demonstrate the effectiveness of our model and a significant improvement in generating quality and efficiency over state-of-the-art methods.

CRediT authorship contribution statement

Pei Dong: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Writing – original draft. **Lei Wu:** Funding acquisition, Writing – original draft, Writing – review & editing. **Ruichen Li:** Formal analysis. **Xiangxu Meng:** Funding acquisition, Writing – review & editing. **Lei Meng:** Funding acquisition, Writing – review & editing.

Declaration of competing interest

All author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability

Data will be made available on request.

Acknowledgments

This work is supported in part by the Key R&D Program of Shandong Porvince, China (Grant no. 2023CXGC010801) and the Oversea Innovation Team Project of the "20 Regulations for New Universities" funding program of Jinan (Grant no. 2021GXRC073).

References

Chen, K., Choy, C.B., Savva, M., Chang, A.X., Funkhouser, T., Savarese, S., 2018. Text2shape: Generating shapes from natural language by learning joint embeddings. In: Asian Conference on Computer Vision. Springer, pp. 100–116.

Cheng, W.-H., Song, S., Chen, C.-Y., Hidayati, S.C., Liu, J., 2021. Fashion meets computer vision: A survey. ACM Comput. Surv. 54 (4), 1–41.

Dhariwal, P., Nichol, A., 2021. Diffusion models beat GANs on image synthesis. Adv. Neural Inf. Process. Syst. 34, 8780–8794.

Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al., 2021. Cogview: Mastering text-to-image generation via transformers. Adv. Neural Inf. Process. Syst. 34, 19822–19835.

Ding, M., Zheng, W., Hong, W., Tang, J., 2022. Cogview2: Faster and better text-to-image generation via hierarchical transformers. arXiv preprint arXiv:2204.14217.

Dong, P., Wu, L., Meng, L., Meng, X., 2022. Disentangled representations and hierarchical refinement of multi-granularity features for text-to-image synthesis. In: Proceedings of the 2022 International Conference on Multimedia Retrieval. pp. 268–276.

Dumoulin, V., Shlens, J., Kudlur, M., 2016. A learned representation for artistic style. arXiv preprint arXiv:1610.07629.

Frans, K., Soros, L., Witkowski, O., 2022. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. Adv. Neural Inf. Process. Syst. 35, 5207–5218.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Commun. ACM 63 (11), 139–144.

Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B., 2022. Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10696–10706.

Gulcehre, C., Chandar, S., Cho, K., Bengio, Y., 2018. Dynamic neural turing machine with continuous and discrete addressing schemes. Neural Comput. 30 (4), 857–884.

He, S., Liao, W., Yang, M.Y., Yang, Y., Song, Y.-Z., Rosenhahn, B., Xiang, T., 2021. Context-aware layout to image generation with enhanced object appearance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 15049–15058.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems, vol. 30.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. Adv. Neural Inf. Process. Syst. 33, 6840–6851.

Ho, J., Salimans, T., 2022. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598.

Kang, M., Zhu, J.-Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T., 2023. Scaling up GANs for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10124–10134.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Koley, S., Bhunia, A.K., Sain, A., Chowdhury, P.N., Xiang, T., Song, Y.-Z., 2023. Picture that sketch: Photorealistic image generation from abstract sketches. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 6850–6861.

Li, B., Qi, X., Lukasiewicz, T., Torr, P.H.S., 2019. Controllable text-to-image generation. ArXiv abs/1909.07083, URL: https://api.semanticscholar.org/CorpusID:202577442.

Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S., Gao, J., 2020. Object-driven text-to-image synthesis via adversarial training. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Liao, W., Hu, K., Yang, M.Y., Rosenhahn, B., 2022. Text to image generation with semantic-spatial aware GAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18187–18196.

Liu, M.-Y., Huang, X., Yu, J., Wang, T.-C., Mallya, A., 2021a. Generative adversarial networks for image and video synthesis: Algorithms and applications. Proc. IEEE 109 (5), 839–862.

Liu, X., Yin, G., Shao, J., Wang, X., et al., 2019. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. Adv. Neural Inf. Process. Syst. 32.

Liu, B., Zhu, Y., Song, K., Elgammal, A., 2021b. Self-supervised sketch-to-image synthesis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, (no. 3), pp. 2073–2081.

Lu, Y., Wu, S., Tai, Y.W., Tang, C.K., 2017. Image generation from sketch constraint using contextual GAN.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Nichol, A.Q., Dhariwal, P., 2021. Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. PMLR, pp. 8162–8171.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M., 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741.

Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y., 2019. Gaugan: Semantic image synthesis with spatially adaptive normalization. In: ACM SIGGRAPH 2019 Real-Time Live! pp. 1–1.

Qiao, T., Zhang, J., Xu, D., Tao, D., 2019. Mirrorgan: Learning text-to-image generation by redescription. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1505–1514.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M., 2022. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I., 2021. Zero-shot text-to-image generation. In: International Conference on Machine Learning. PMLR, pp. 8821–8831.

Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H., 2016a. Learning what and where to draw. In: Advances in Neural Information Processing Systems, vol. 29.

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

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022a. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022b. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 10684–10695.

Ruan, S., Zhang, Y., Zhang, K., Fan, Y., Tang, F., Liu, Q., Chen, E., 2021. Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13960–13969.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al., 2022. Photorealistic text-to-image diffusion models with deep language understanding. Adv. Neural Inf. Process. Syst. 35, 36479–36494.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training GANs. In: Advances in Neural Information Processing Systems, vol. 29.

Sauer, A., Karras, T., Laine, S., Geiger, A., Aila, T., 2023. Stylegan-t: Unlocking the power of GANs for fast large-scale text-to-image synthesis. arXiv preprint arXiv:2301.09515.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S., 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. PMLR, pp. 2256–2265.

Sukhbaatar, S., Weston, J., Fergus, R., et al., 2015. End-to-end memory networks. In: Advances in Neural Information Processing Systems, vol. 28.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the Inception Architecture for Computer Vision. IEEE, pp. 2818–2826.

Tao, M., Bao, B.-K., Tang, H., Xu, C., 2023. GALIP: Generative adversarial CLIPs for text-to-image synthesis. arXiv preprint arXiv:2301.12959.

Tao, M., Tang, H., Wu, F., Jing, X.-Y., Bao, B.-K., Xu, C., 2022. DF-GAN: A simple and effective baseline for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16515–16525.

Toda, R., Teramoto, A., Kondo, M., Imaizumi, K., Saito, K., Fujita, H., 2022. Lung cancer CT image generation from a free-form sketch using style-based pix2pix for data augmentation. Sci. Rep. 12 (1), 12867.

Van Den Oord, A., Vinyals, O., et al., 2017. Neural discrete representation learning. In: Advances in Neural Information Processing Systems, vol. 30.

Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., Duan, N., 2022. Nüwa: Visual synthesis pre-training for neural visual world creation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI. Springer, pp. 720–736.

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X., 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1316–1324.

Xue, H., Huang, Z., Sun, Q., Song, L., Zhang, W., 2023. Freestyle layout-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 14256–14266.

Ye, S., Liu, F., Tan, M., 2022. Recurrent affine transformation for text-to-image synthesis. arXiv preprint arXiv:2204.10482.

Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al., 2022. Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789.

Zhang, T., Fu, H., Zhao, Y., Cheng, J., Guo, M., Gu, Z., Yang, B., Xiao, Y., Gao, S., Liu, J., 2019. SkrGAN: Sketching-rendering unconditional generative adversarial networks for medical image synthesis.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N., 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5907–5915.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N., 2018. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE Trans. Pattern Anal. Mach. Intell. 41 (8), 1947–1962.

Zhao, B., Meng, L., Yin, W., Sigal, L., 2019. Image generation from layout. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Zhi, J., 2017. Pixelbrush: Art generation from text with GANs. In: Cl. Proj. Stanford CS231N Convolutional Neural Networks Vis. Recognition, Sprint 2017. p. 256.

Zhou, Y., Zhang, R., Chen, C., Li, C., Tensmeyer, C., Yu, T., Gu, J., Xu, J., Sun, T., 2022. Towards language-free training for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17907–17917.

Zhu, M., Pan, P., Chen, W., Yang, Y., 2019. 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. pp. 5802–5810.