Disentangled Representations and Hierarchical Refinement of Multi-Granularity Features for Text-to-Image Synthesis

Pei Dong
Shandong University
Jinan, Shan Dong, China
dongpei1992@outlook.com

Lei Wu*
Shandong University
Jinan, Shan Dong, China
i_lily@sdu.edu.cn

Lei Meng
Shandong University
Jinan, Shan Dong, China
lmeng@sdu.edu.cn

Xiangxu Meng
Shandong University
Jinan, Shan Dong, China
mxx@sdu.edu.cn

ABSTRACT

In this paper, we focus on generating photo-realistic images from given text descriptions. Current methods first generate an initial image and then progressively refine it to a high-resolution one. These methods typically indiscriminately refine all granularity features output from the previous stage. However, the ability to express different granularity features in each stage is not consistent, and it is difficult to express precise semantics by further refining the features with poor quality generated in the previous stage. Current methods cannot refine different granularity features independently, resulting in that it is challenging to clearly express all factors of semantics in generated image, and some features even become worse. To address this issue, we propose a Hierarchical Disentangled Representations Generative Adversarial Networks (HDR-GAN) to generate photo-realistic images by explicitly disentangling and individually modeling the factors of semantics in the image. HDR-GAN introduces a novel component called multi-granularity feature disentangled encoder to represent image information comprehensively through explicitly disentangling multi-granularity features including pose, shape and texture. Moreover, we develop a novel Multi-granularity Feature Refinement (MFR) containing a Coarse-grained Feature Refinement (CFR) model and a Fine-grained Feature Refinement (FFR) model. CFR utilizes coarse-grained disentangled representations (e.g., pose and shape) to clarify category information, while FFR employs fine-grained disentangled representations (e.g., texture) to reflect instance-level details. Extensive experiments on two well-studied and publicly available datasets (i.e., CUB-200 and CLEVR-SV) demonstrate the rationality and superiority of our method.

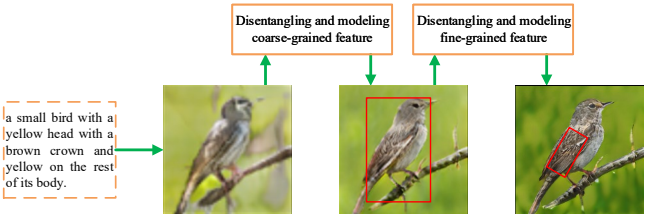

Figure 1: The illustration of our idea. Our model explicitly disentangles multi-granularity features (pose, shape and texture). The coarse-grained disentangled representations are used in CFR to clarify category information, and the fine-grained disentangled representations are used in FFR to reflect instance-level details.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; *Machine learning*.

KEYWORDS

Text-to-Image Synthesis, Hierarchical Disentangled Representations, Generative Adversarial Networks, Multi-granularity Features

ACM Reference Format:
Pei Dong, Lei Wu, Lei Meng, and Xiangxu Meng. 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 (ICMR '22), June 27–30, 2022, Newark, NJ, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3512527.3531389

1 INTRODUCTION

Text-to-image synthesis requires an agent to generate a photo-realistic image according to the given text descriptions. Due to the significant potential in many applications, such as art generation [31] and computer-aided design [1], text-to-image synthesis has become an active research area in both natural language processing and computer vision communities.

To express more explicit category information and richer instance-level details, leveraging the power of GANs [4], multi-stage methods [27–29] are designed to first generate initial low-resolution images and then refine the initial images to high-resolution ones. Based on the multi-stage generative model, recent methods are not

*Corresponding Author

ICMR '22, June 27–30, 2022, Newark, NJ, USA

ACM ISBN 978-1-4503-9238-9/22/06…$15.00
https://doi.org/10.1145/3512527.3531389

only conditioned on text embedding but incorporated with additional supervision. Common additional supervision information includes layout [6, 7, 18], scene graph [12, 14], and semantic mask [8, 10, 25, 26].

All the above methods do not explicitly disentangle and individually model the factors of semantics in the image but indiscriminately refine all granularity features at each stage. However, the ability to express different granularity features in each stage is not consistent, and the refinement result of each stage depends heavily on the output quality of the previous stage [34], which means that it is challenging to express precise semantics by further refining the features with poor quality generated in the previous stage. This results in that current text-to-image synthesis methods cannot ensure good modeling of all granularity features simultaneously at a particular stage and make the final generated results unable to express explicit category information and instance-level details.

To address the above issues, we propose a Hierarchical Disentangled Representations Generative Adversarial Networks (HDR-GAN). In our model, we first add a novel component called multi-granularity feature disentangled encoder. Adding this component is inspired by pose estimation [24, 32] and semantic segmentation [13] tasks which show learning features in multi-granularity with a multi-branch deep network can effectively improve network performance. We build a multi-granularity feature disentangled encoder upon FineGAN [21] and MixNMatch [30], which learn to disentangle the factors of variation in the data using information theory. Multi-granularity feature disentangled encoder contains three branches which can represent image information comprehensively through explicitly disentangling multi-granularity features including pose, shape and texture. We divide these branches into a coarse-grained disentangled encoder related to category information and a fine-grained disentangled encoder related to instance-level details. Moreover, we develop a novel Multi-granularity Feature Refinement (MFR) to enhance the expression of semantic information at different scales by gradually refining the disentangled representations of different granularity features and serving them as additional supervision information (as shown in Figure 1). MFR applies a Coarse-grained Feature Refinement (CFR) model and a Fine-grained Feature Refinement (FFR) model for image refinement. Specifically, CFR disentangles coarse-grained representations from the initial image through a coarse-grained disentangled encoder and utilizes disentangled representations as additional supervision to clarify category information. In the following stage, FFR disentangles fine-grained representations from the generated image of CFR through a fine-grained disentangled encoder and employs fine-grained representations as additional supervision to reflect instance-level details. MFR can match the generation capabilities of the refinement process at different stages, refine different granularity features independently, strengthen the expression of specific scale semantics at different stages, and finally generate realistic images from text descriptions. Extensive experiments on CUB-200 and CLEVR-SV datasets demonstrate that our method significantly outperforms the state-of-the-art methods, with more accurate category information and richer instance-level details in the generated image. Moreover, we conduct a series of analysis experiments to evaluate the importance of each component in our approach and further validate the effectiveness of HDR-GAN.

Our main contributions are summarized as follows:

• We are the first to introduce multi-granularity feature disentangled representations into text-to-image synthesis. We introduce a novel component called multi-granularity feature disentangled encoder to represent image information comprehensively through explicitly disentangling multi-granularity features.

• We develop a novel Multi-granularity Feature Refinement (MFR) to preferentially refine features that have better initial representation in the previous stage and thus gradually enhance the expression of semantic information at different scales.

• The experimental results demonstrate that HDR-GAN outperforms the state-of-the-art approaches. The generated images contain more explicit category information and richer instance-level details.

2 RELATED WORK

According to the level of supervision, current methods can be divided into approaches that use single captions as input versus approaches that use captions and additional information.

Direct Text-to-Image Synthesis Approaches that do not use additional supervision information use only the image caption as conditional input. Reed et al. propose GAN-INT-CLS [17] which conditions the generation process on sentence embedding obtained from a pre-trained text encoder to generate images with consistent content. TAC-GAN [3] employs an additional auxiliary classification loss inspired by AC-GAN [20], where the discriminator outputs both the sources (real/fake) and the classes. In order to synthesize higher resolution images, StackGAN [28] divides the generation process into two stages which first generates a low-resolution rough initial image, and then refines the initial image to a high-resolution photo-realistic one. Compared to StackGAN [28], StackGAN++ [29] improves the architecture to a tree structure which contains multiple generators and multiple discriminators. AttnGAN [27] is built upon StackGAN++ [29] and adds attention mechanisms component into a multi-stage generator pipeline. The attention mechanisms component allows the network to synthesize fine-grained details based on relevant local words embedding. DM-GAN [34] introduces a memory writing gate to the refinement stage that is capable of dynamically selecting relevant words according to the initial image. MirrorGAN [16] learns text-to-image generation by aligning the re-description of the generated image with the given text description. In VQA-GAN [15], questions and answers (QAs) are chosen as locally-related texts, which makes it possible to use VQA accuracy as a new evaluation metric.

Text-to-Image Synthesis with Additional Supervision Approaches use not only the image caption but additional supervision information as conditional input. These methods pay attention to refining salient objects in a local region. GAWWN [18] generates images which condition on both textual descriptions and object locations to control what content to draw in which location. Hinz et al. [6] introduce an object pathway to both the generator and the discriminator, which allows to explicitly model the location of arbitrarily many objects within an image. OP-GAN [7] extends [6] by adding additional object pathways at higher layers of the generator and discriminator. Another line of research leverages masks to provide an even better location and shape signal of objects to the

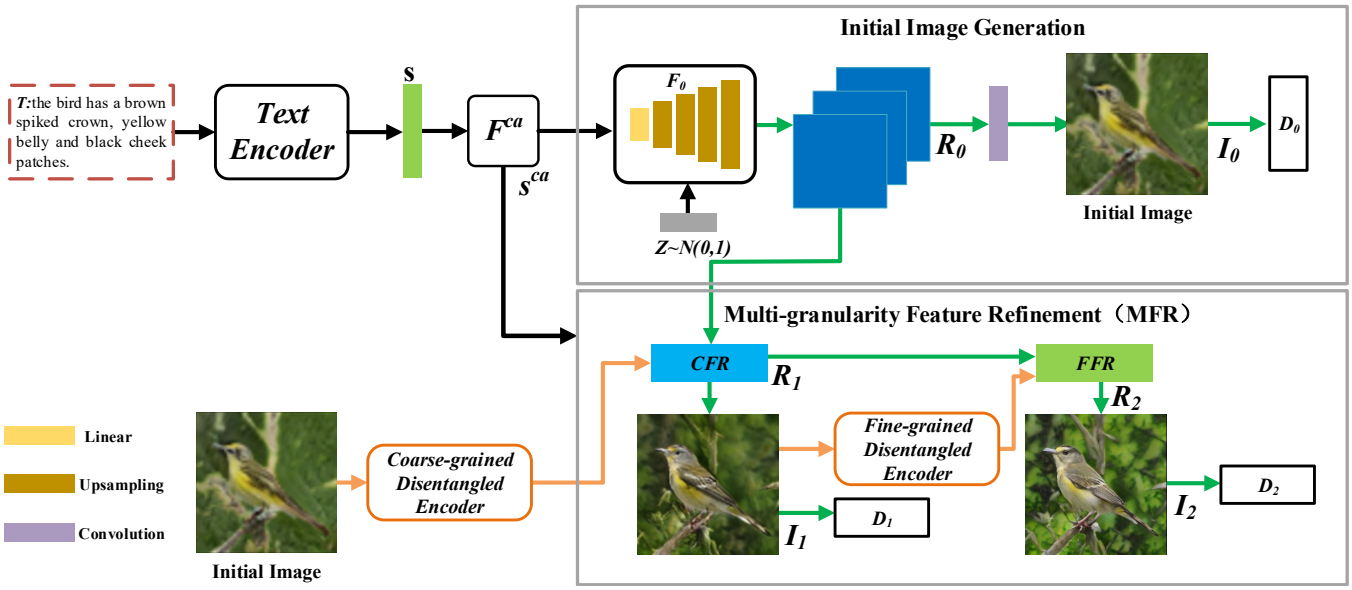

Figure 2: Overview of our model architecture called HDR-GAN. HDR-GAN first generates an initial image, and then MFR employs disentangled representations of different granularity features to enhance the expression of semantic information at different scales in each stage. The multi-granularity feature disentangled encoder contains a coarse-grained disentangled encoder and a fine-grained disentangled encoder.

network. In [8], semantic masks are obtained by first generating a layout from the input description and then using it to predict the shape. The image generator has single stage and conditions only on the generated mask and global sentence information. Obj-GAN [10] is built upon [8] and consists of an object-driven attentive generator and an object-wise discriminator. AGAN-CL [26] introduces a sub-network which is trained to predict masks, thereby providing the number of objects, location, size and shape information. The image mask is given as input to a cyclic auto-encoder, similar to [33], to produce photo-realistic images. Wang et al. [25] propose an end-to-end framework with spatial constraints using semantic layout to guide the image synthesis. At each stage, the generator produces an image and additionally a layout to be used by the corresponding discriminator. Scene graph is also common additional supervision information. Related research is often used as follow-up processing for text-to-scene graph models. PasteGAN [12] embeds the objects as well as their relationships conditioned on scene graph and encodes the appearance of each object into one map. An interactive framework in [14] updates an image obtained from a scene graph by updating the scene graph while keeping the generated content as much as possible. These methods that use additional supervision information can help the network learn salient object features in images and push the state-of-the-art performance.

However, the above methods typically model all granularity features at each stage, and the refined results are susceptible to poor quality features from the previous stage. This results in that current methods cannot ensure accurate expression of semantics at different scales.

3 HDR-GAN

As illustrated of our model architecture in Figure 2, HDR-GAN consists of three parts: 1) *Initial Image Generation* stage generates low-resolution images conditioned on text embeddings and random noise vectors; 2) *Multi-granularity Feature Disentangled Encoder* applies a coarse-grained disentangled encoder and a fine-grained disentangled encoder to represent semantic information of images comprehensively through explicitly disentangling multi-granularity features; 3) *Multi-granularity Feature Refinement* refines the initial image by gradually updating the disentangled representations and serving them as additional supervision information to enhance the expression of semantic information.

3.1 Initial Image Generation

At the initial image generation stage, first, the text encoder encodes the input text description T into a text embedding s. We enhance sentence features with Conditioning Augmentation (CA) technique [28], which resamples the input sentence vector from an independent Gaussian distribution and is represented as follows:

$$s_{ca} = F^{ca}(s), \tag{1}$$

where $F^{ca}(\cdot)$ is the CA function, and s_{ca} stands for the enhanced text embedding with CA. Then, we utilize a random noise vector $z \sim N(0,1)$ sampled from a normal distribution and the enhanced sentence embeddings s_{ca} to generate a low-resolution rough initial image I_0. $\boldsymbol{R}_0$ is the corresponding image features at the initial stage:

$$\boldsymbol{R}_0 = F_0\left(\boldsymbol{s}_{ca}, \boldsymbol{z}\right), \tag{2}$$

where $F_0(\cdot)$ is the image generator which consists of fully connected and upsampling layers.

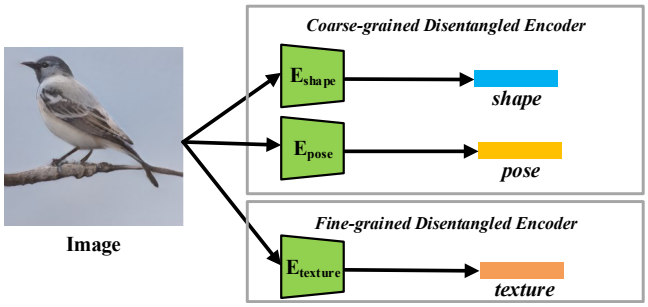

Figure 3: Overview of multi-granularity feature disentangled encoder.

3.2 Multi-granularity Feature Disentangled Encoder

We build multi-granularity feature disentangled encoder upon FineGAN [21] and MixNMatch [30]. FineGAN designs a multi-stage hierarchical generative model which takes as input randomly sampled latent codes to generate semantic features at different scales gradually. To disentangle multi-granularity features expressing semantics at different scales, FineGAN uses information theory [2], and imposes constraints based adversarial training [4] on the relationships between the latent codes and mutual information maximization between the latent codes and corresponding image, so that each code gains control over the respective factor. MixNMatch [30] extends this approach to disentangled representations and modelling tasks conditioned on real image, the key idea of which is to make sure the paired image-code distribution produced by the encoder ($x \sim P_{\text{data}}, \hat{y} \sim E(x)$) and the paired image-code distribution produced by the generator ($\hat{x} \sim G(y), y \sim P_{code}$) are matched via a paired adversarial loss as follows:

$$L = \min_{G,E} \max_{D} \mathbb{E}_{x \sim P_{\text{data}}} \mathbb{E}_{\hat{y} \sim E(x)} [\log D(x, \hat{y})] + \mathbb{E}_{y \sim P_{\text{code}}} \mathbb{E}_{\hat{x} \sim G(y)} [\log(1 - D(\hat{x}, y))], \tag{3}$$

where $E(\cdot)$ stands for an encoder, P_{data} is the real image distribution, P_{code} is the latent code distribution, x is the real image, $\hat{x}$ is the generated image, y is a placeholder for the latent codes and $\hat{y}$ is disentangled representation from the encoder E.

In this paper, we referring to FineGAN and MixNMatch, train a multi-granularity feature disentangled encoder with three branches for disentangling coarse-grained features related to category information such as shape, pose and fine-grained features related to instance-level details such as texture features (shown as Figure 3). In our model, we combine the branches which disentangle pose and shape features to form a coarse-grained disentangled encoder and form a fine-grained disentangled encoder based on the branch which disentangles texture features.

3.3 Multi-granularity Feature Refinement

To the best of our knowledge, we are the first to introduce multi-granularity feature disentangled representations into text-to-image synthesis. How to integrate disentangled representations into the image refinement stage is the main challenge that we should tackle. Previous research has shown that in the multi-stage generative model, the refinement result of each stage depends heavily on the output quality of the previous stage [34]. However, the ability of each stage to express different granularity features is not consistent, the shallower stages generate coarse-grained features related to categories, and the deeper stages generate fine-grained features related to instance-level details. Unlike previous methods that refine all granularity features at each stage, the Multi-granularity Feature Refinement(MFR) that we proposed matches the generation capabilities of the refinement process at different stages, and preferentially refines features that have better initial representation in the previous stage. Specifically, MFR applies a Coarse-grained Feature Refinement (CFR) model and a Fine-grained Feature Refinement (FFR) model for image refinement. CFR utilizes coarse-grained representations as additional supervision to clarify category information. FFR employs fine-grained representations as additional supervision to reflect instance-level details. In the following part, we will introduce the technical details of CFR and FFR.

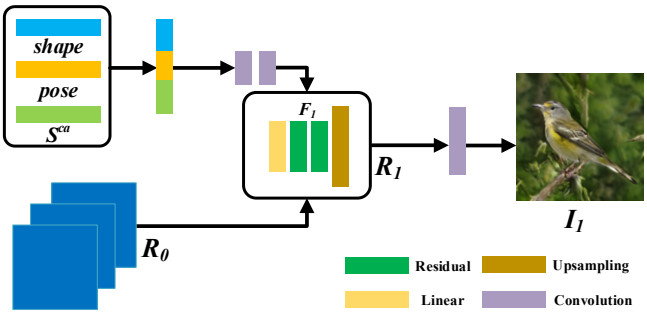

Figure 4: The architecture of CFR.

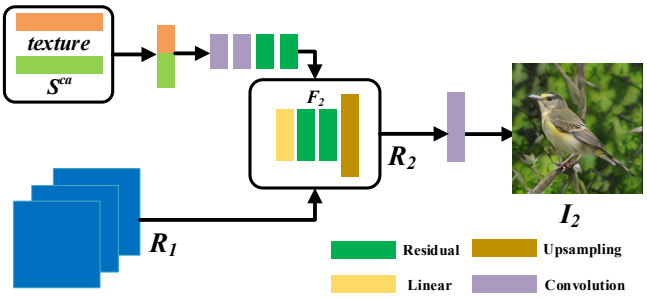

Figure 5: The architecture of FFR.

3.3.1 CFR.

For multi-stage generative models, to synthesize photo-realistic and semantically consistent images, it is necessary to first clarify the category information of the objects. Therefore, CFR is developed for refining coarse-grained features related to categories based on the initial image.

As shown in Figure 4, when the initial image I_0 is obtained from the *Initial Image Generation* stage, the coarse-grained disentangled encoder encodes the initial image to obtain disentangled representations of coarse-grained features such as shape d_{shape} and pose d_{pose}. CFR first uses two convolutional layers to update the shape and pose disentangled representations based on text embedding s_{ca} and to fuse all information as supervision, and then the fused supervision information is used to update the image features $\boldsymbol{R}_0$ of

the initial image. Compared to previous methods, CFR focuses on utilizing coarse-grained representations as additional supervision to clarify category information. This process can be mathematically formulated as follows:

$$\boldsymbol{R}_1 = F_1\left(\boldsymbol{R}_0, (s_{ca}, d_{shape}, d_{pose})\boldsymbol{V}\right), \tag{4}$$

where $\boldsymbol{R}_1$ is the corresponding image features of the CFR, $F_1(\cdot)$ is the image generator which consists of fully connected, upsampling layers and residual blocks, $\boldsymbol{V}$ is perception layers to convert fused supervision information to an underlying common semantic space of visual features.

3.3.2 FFR.

After the image can clearly express the category information of the objects, the generative model needs to add fine-grained details to the image that reflect the differences between instances. As shown in Figure 5, FFR is developed for refining fine-grained features related to instance-level details based on generation results I_1 of CFR. The fine-grained disentangled encoder first obtains disentangled representations of texture features $d_{texture}$ in I_1. FFR uses two convolutional layers and two residual blocks to update and fuse the texture disentangled representations based on text embedding s_{ca}, then the fused supervision information of text embedding s_{ca} and texture features $d_{texture}$ is used to update the image features $\boldsymbol{R}_1$. FFR focuses on employing fine-grained representations as additional supervision to reflect instance-level details. This process can be mathematically formulated as follows:

$$\boldsymbol{R}_2 = F_2\left(\boldsymbol{R}_1, (s_{ca}, d_{texture})\boldsymbol{W}\right), \tag{5}$$

where $\boldsymbol{R}_2$ is the corresponding image features of the FFR, $F_2(\cdot)$ is the image generator in FFR, $\boldsymbol{W}$ is similar to $\boldsymbol{V}$ and is used for feature space transformation.

3.4 Objective Function

During training HDR-GAN, the generator G and discriminator D are trained alternately at each stage. Specially, the generator G in i^{th} stage is trained by minimizing the loss as follows:

$$L_{G_i} = -\frac{1}{2}[\underbrace{\mathbb{E}_{I_i\sim p_{G_i}} \log D_i\left(I_i\right)}_{\text{unconditional loss}} + \underbrace{\mathbb{E}_{I_i\sim p_{G_i}} \log D_i\left(I_i, T\right)}_{\text{conditional loss}}], \tag{6}$$

where I_i is the generated image from the distribution G_i at i^{th} stage. The first unconditional loss term is used to distinguish whether the image is visually real or fake. The second term is a conditional loss used to determine whether the underlying image and sentence semantics are consistent.

Following StackGAN++ [29], the *CA* loss is defined as the Kullback-Leibler divergence (KL divergence) between the standard Gaussian distribution and the conditioning Gaussian distribution for text embeddings, which is calculated by

$$L_{CA} = D_{KL}\left(\mathcal{N}\left(\mu(\mathbf{s}), \sum(\mathbf{s})\right) \| \mathcal{N}(0, I)\right), \tag{7}$$

where $\mathcal{N}\left(\mu(\mathbf{s}), \sum(\mathbf{s})\right)$ is an independent Gaussian distribution, the mean $\mu(\mathbf{s})$ and diagonal covariance matrix $\sum(\mathbf{s})$are functions of the text embedding s. Both $\mu(\mathbf{s})$ and $\sum(\mathbf{s})$ are learned jointly with the rest of the network.

The final objective function of the generator networks is

$$L_G = \sum_i L_{G_i} + \lambda L_{CA}, \tag{8}$$

where λ is a regularization parameter that balances the L_{CA} terms.

Discriminator D is optimized to distinguish real images and synthetic images generated by G. The discriminator D in i^{th} stage is trained by minimizing the loss as follows:

$$\begin{aligned} L_{D_i} = -\frac{1}{2}[&\underbrace{\mathbb{E}_{I_i^{GT}\sim p_{GT}} \log D_i\left(I_i^{GT}\right) + \mathbb{E}_{I_i\sim p_{G_i}} \log D_i\left(I_i\right)}_{\text{unconditional loss}} \\ &+ \underbrace{\mathbb{E}_{I_i^{GT}\sim p_{GT}} \log D_i\left(I_i^{GT}, T\right) + \mathbb{E}_{I_i\sim p_{G_i}} \log D_i\left(I_i, T\right)}_{\text{conditional loss}}], \end{aligned} \tag{9}$$

where the unconditional loss is responsible for distinguishing synthesized images from real images and the conditional term determines whether the image matches the input text description vector. I_i^{GT} is sampled from the real image distribution p_{GT} at i^{th} stage. The final objective function for whole discriminator network is as follows:

$$L_D = \sum_i L_{D_i}. \tag{10}$$

4 EXPERIMENT

In this section, we will first introduce the experiment setup. Next, we will compare HDR-GAN with GAWWN [18], StackGAN++ [29], AttnGAN [27], DM-GAN [34] and MirrorGAN [16] on CUB-200 and CLEVR-SV which are publicly available and well-studied datasets. Then, a visualization study will be discussed to show the effectiveness of HDR-GAN. Finally, we present ablation studies on the key components of HDR-GAN, including CFR and FFR. Meanwhile, we analyze and demonstrate the effectiveness and interpretability of the HDR-GAN design.

4.1 Experiment Setup

Datasets. We evaluate our model on CUB-200 [23] and CLEVR-SV [11] datasets which are commonly used dataset. The CUB-200 bird dataset contains 8,855 training images and 2,933 test images belonging to 200 categories, and each bird image has 10 text descriptions.

The CLEVR-SV dataset is modified from the CLEVR [9] dataset which is used for the visual question answering. The dataset contains 10,000 training images and 3,000 test images generated by dataset generation. The automatically generated text encoding mainly describes the positional relationship between multiple geometric objects and the characteristics of the object in the form of a binary dictionary. We provide some qualitative results obtained with CLEVR-SV to verify the effectiveness of HDR-GAN on multi-objective scene generation tasks.

Evaluation Metrics. Following previous works, we quantitatively evaluate the performance of our HDR-GAN in terms of Inception Score (IS) [19], Frechet Inception Distance (FID) [5], and R-precision [27].

We obtain IS by employing a pre-trained Inception-v3 network [22] to compute the KL-divergence between the marginal class distribution and the conditional class distribution. A larger IS signifies

Figure 6: Example results for text-to-image synthesis by StackGAN++ [29], AttnGAN [27], DM-GAN [34] and our proposed HDR-GAN on CUB-200.

Figure 7: Example results for text-to-image synthesis by our proposed HDR-GAN on CLEVR-SV.

that the generated images contain richer and more discriminative semantic information.

FID computes the Frechet distance between the synthetic images 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.

R-precision is utilized to assess the semantic consistency between the synthetic image and the given text description. We utilize pre-trained DAMSM [27] 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.

Implementation Details. Following [17], a pre-trained character-level ConvNet with a recurrent neural network called Char-CNN-RNN[18] is used to calculate the text embedding from text descriptions T. The generator noise z is sampled from a 100-dimensional unit normal distribution. The size of the initial generated low-resolution image I_0 is set to 64×64, the size of the CFR output I_1 is set to 128×128 and the final synthesized high-resolution image I_2 at the FFR stage has the size of 256×256. Coarse-grained disentangled encoder disentangles pose and shape features from the initial image which is upsampled to 128×128; the fine-grained disentangled encoder takes unprocessed output image from CFR as input and disentangles texture features. We use the Adam optimizer and set the learning rate of the generator to 0.0004 and that of the discriminator to 0.0002. We train HDR-GAN on a single NVIDIA 2080 GPU with a batch size of 24 on each one. HDR-GAN is trained with 300 epochs and 100 epochs on CUB-200 and CLEVR-SV respectively.

4.2 Quantitative Results

We conduct experiments on multiple evaluation metrics to compare our method with state-of-the-art methods on CUB-200, a dataset with full of semantic details. The overall results are summarized in Table 1. For meaningful and fair comparisons with previous methods, all evaluation metrics are computed in two settings: in

the first setting, the different models are compared directly in terms of their generated images, which have different resolutions; in the second setting, all generated images are resized to 128×128 before computing IS*, FID* and R-precision* score for fair comparison. This reduces the impact of computing power on network depth limitations (such as directly increasing the upsampling stage).

It is clear that our proposed HDR-GAN achieves the best performance compared to previous methods, among IS and FID score evaluation metrics which are used to measure the performance of generated image feature expression. HDR-GAN improves IS score to 5.03 and decreases FID score to 14.70. These results show that HDR-GAN can generate images with more diversity and better quality from a single given caption. This is because HDR-GAN can comprehensively represent image information by disentangling features of different granularities. Moreover, the core component MFR is developed to make each stage of refinement network focus on refining semantic information which is at similar scales and has better initial representation.

HDR-GAN also achieves competitive results on R-precision scores that measure semantic consistency. HDR-GAN achieves an R-precision score of 71.89 without relying on word-level semantic representation. The score is very close to that of the state-of-the-art method [34], which uses an attention mechanism and a memory network to optimize synthetic images based on word-level semantic features. This result shows that synthesizing images through explicitly disentangling the multi-granularity feature as supervision information can enhance text-image semantic consistency.

4.3 Qualitative Results

To evaluate the visual quality of generated images, we first show some subjective comparisons among StackGAN++ [29], AttnGAN [27], DM-GAN [34] and our proposed HDR-GAN in Figure 6. Example results in Figure 6 show that HDR-GAN significantly improves the anti-deformation of the text-to-image model generation results, which indicates that HDR-GAN can provide explicit category information (1^{st}, 6^{th} and 7^{th} columns). In addition, the HDR-GAN generates results containing richer instance-level details, such as the texture of wings and the blobs of the abdomen, which are often realized as uniform color patches in the generated results of previous methods (2^{nd}, 5^{th} and 8^{th} columns). Finally, we also found that HDR-GAN can more accurately express the fine-grained features of the image. For example, in the 5^{th} column, only HDR-GAN accurately generates the image features corresponding to the text description of *"pale belly and breast"*. The reason is that HDR-GAN obtains explicitly disentangled multi-granularity features to express image information comprehensively. Moreover, MFR is developed to enhance the expression of semantic information at different scales by gradually refining the disentangled representations of different granularity features.

We also experiment on the CLEVR-SV dataset to evaluate the visual quality of generated images in multiple target scenarios. Example results in Figure 7 show that the generated results of HDR-GAN can clearly reflect the spatial relationship between multiple objects and the different granularities characteristics of each object (such as shape and color).

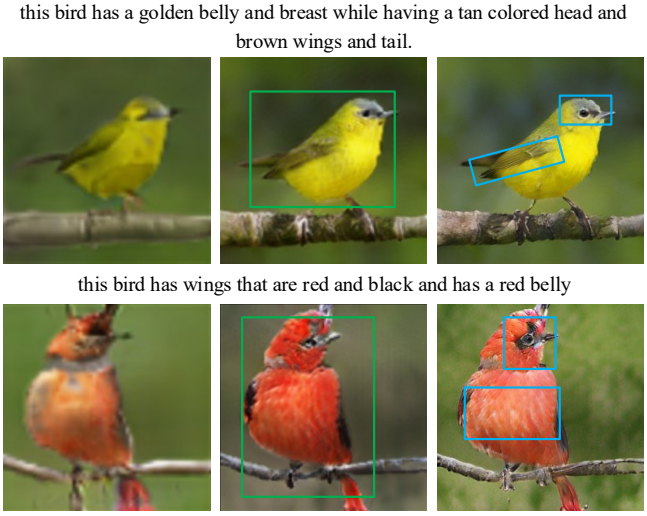

Figure 8: Intermediate results of different stages of our HDR-GAN on CUB-200.

To better understand the intermediate changes of generated images in HDR-GAN, Figure 8 shows the generation results of HDR-GAN at different stages. In each row of images, the left side is the output from the initial image generation stage, the middle is the output from CFR which expresses explicit category semantic information, and the right side is the output from FFR which has added rich instance-level details. Compared with the initial image, the output of CFR has a more accurate shape and more precise boundaries (the part of the green bounding box). Based on the explicit category information, the output of FFR contains rich instance-level details such as feather texture, colors of leg and beak (the part of the blue bounding box). These results confirm that refining different granularity features obtained from multi-granularity feature disentangled encoder independently can enhance the expression of specific scale semantics at different stages.

4.4 Ablation Study

The quantitative and qualitative experimental results have proved the superiority of our proposed HDR-GAN. In this section, we further analyze which component is significant for performance improvement. Therefore, we perform an ablation study on CUB-200 to verify the effectiveness of each part in MFR, including CFR and FFR. Corresponding results are illustrated in Table 2. According to the results, we can observe varying degrees of model performance decline when removing CFR and FFR separately from HDR-GAN. The ablation studies show that the comprehensive utilization of disentangled different granularity features is helpful for image synthesis. CFR and FFR focus on refining features at different granularities and can employ this information as additional supervision to refine the image.

4.5 Design Analysis

We believe that explicitly disentangling image features and refining independently in the order of coarse-grained to fine-grained is the key reason why HDR-GAN can effectively enhance the expression

Table 1: Quantitative results on CUB-200 with different models.

Model	IS↑	IS*↑	FID↓	FID*↓	R-precision (%)↑	R-precision* (%)↑
GAWWN	3.62 ± 0.07	3.62 ± 0.07	67.22	67.22	46.71	46.71
StackGAN++	3.83 ± 0.04	3.80 ± 0.03	15.30	**15.47**	52.40	47.80
DM-GAN	4.75 ± 0.07	4.66 ± 0.06	16.09	16.93	**72.31**	**67.44**
AttnGAN	4.37 ± 0.04	4.30 ± 0.04	22.37	24.72	64.01	52.01
MirrorGAN	4.54± 0.17	4.47± 0.13	19.71	20.13	57.67	56.13
Ours	**5.03± 0.18**	**4.91± 0.16**	**14.70**	15.70	71.89	64.33

Table 2: Ablation performance on CUB-200 about IS, FID and R-precision.

Model	IS↑	FID↓	R-precision (%)↑
HDR-GAN (w/o CFR)	4.36± 0.07	16.64	61.17
HDR-GAN (w/o FFR)	4.51± 0.23	16.33	66.31
HDR-GAN	**5.03± 0.18**	**14.70**	**71.89**

Table 3: Influence of refinement order in MFR.

Model	IS↑	FID↓	R-precision (%)↑
FFR to CFR	4.31± 0.21	16.61	62.17
CFR to FFR (HDR-GAN)	**5.03± 0.18**	**14.70**	**71.89**

of image semantics. To verify this, we swap the order of refinement model in MFR to compare the IS, FID and R-precision of the generated results. Corresponding results are illustrated in Table 3, the experimental results do confirm the importance of matching the generation capabilities in the refinement process at different stages and of preferential refinement in features that have better initial representation in the previous stage. HDR-GAN can make full use of multi-granularity features disentangled representations to gradually enhance the expression of semantic information at different scales.

5 CONCLUSION

In this paper, a Hierarchical Disentangled Representations Generative Adversarial Networks for text-to-image synthesis, named HDR-GAN, is proposed to express more explicit category information and richer instance-level details. HDR-GAN successfully exploits the idea of refining different granularity features independently. In HDR-GAN, the multi-granularity feature disentangled encoder first represents image information comprehensively through explicitly disentangling multi-granularity features. Moreover, MFR, which contains a Coarse-grained Feature Refinement (CFR) model and a Fine-grained Feature Refinement (FFR) model, is developed for image refinement by gradually refining the disentangled representations of different granularity features. Extensive experimental results clearly demonstrate the importance of explicit disentanglement and individual modeling for different granularity features to enhance the expression of semantic information at different scales in text-to-image synthesis tasks.

In the future, we will extend our method to COCO, a dataset of complex scenes with multiple objects. This requires a novel disentangled representations method to deal with object occlusion and data imbalance. We will also particularly focus on exploring to achieve self-supervised image feature disentangled representations.

6 ACKNOWLEDGE

This work is supported in part by the National Key R&D Program of China (Grant no.2021YFC3300203) and the Oversea Innovation Team Project of the "20 Regulations for New Universities" funding program of Jinan (Grant no. 2021GXRC073)

REFERENCES

[1] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. 2018. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Asian Conference on Computer Vision*. Springer, 100–116.

[2] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2180–2188.

[3] Ayushman Dash, John Cristian Borges Gamboa, Sheraz Ahmed, Marcus Liwicki, and Muhammad Zeshan Afzal. 2017. Tac-gan-text conditioned auxiliary classifier generative adversarial network. *arXiv preprint arXiv:1703.06412* (2017).

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[6] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. 2018. Generating Multiple Objects at Spatially Distinct Locations. In *International Conference on Learning Representations*.

[7] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. 2020. Semantic Object Accuracy for Generative Text-to-Image Synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[8] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. 2018. Inferring semantic layout for hierarchical text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7986–7994.

[9] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2901–2910.

[10] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. 2019. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12174–12182.

[11] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6329–6338.

[12] Yikang Li, Tao Ma, Yeqi Bai, Nan Duan, Sining Wei, and Xiaogang Wang. 2019. Pastegan: A semi-parametric method to generate image from scene graph. *Advances in Neural Information Processing Systems* 32 (2019), 3948–3958.

[13] Xiankai Lu, Wenguan Wang, Jianbing Shen, Yu-Wing Tai, David J Crandall, and Steven CH Hoi. 2020. Learning video object segmentation from unlabeled videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8960–8970.

[14] Gaurav Mittal, Shubham Agrawal, Anuva Agarwal, Sushant Mehta, and Tanya Marwah. 2019. Interactive image generation using scene graphs. *arXiv preprint*

arXiv:1905.03743 (2019).

[15] Tianrui Niu, Fangxiang Feng, Lingxuan Li, and Xiaojie Wang. 2020. Image synthesis from locally related texts. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*. 145–153.

[16] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1505–1514.

[17] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*. PMLR, 1060–1069.

[18] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. 2016. Learning what and where to draw. *Advances in neural information processing systems* 29 (2016), 217–225.

[19] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems* 29 (2016), 2234–2242.

[20] Rui Shu, Hung Bui, and Stefano Ermon. 2017. Ac-gan learns a biased distribution. In *NIPS Workshop on Bayesian Deep Learning*, Vol. 8.

[21] Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. 2019. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6490–6499.

[22] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[23] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).

[24] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*. 274–282.

[25] Min Wang, Congyan Lang, Liqian Liang, Songhe Feng, Tao Wang, and Yutong Gao. 2020. End-to-End Text-to-Image Synthesis with Spatial Constrains. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 4 (2020), 1–19.

[26] Min Wang, Congyan Lang, Liqian Liang, Gengyu Lyu, Songhe Feng, and Tao Wang. 2020. Attentive generative adversarial network to bridge multi-domain gap for image synthesis. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[27] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 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*. 1316–1324.

[28] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 5907–5915.

[29] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2018. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence* 41, 8 (2018), 1947–1962.

[30] Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. 2020. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International Conference on Machine Learning*. PMLR, 11117–11128.

[31] Jiale Zhi. 2017. PixelBrush: Art Generation from text with GANs. In *Cl. Proj. Stanford CS231N Convolutional Neural Networks Vis. Recognition, Sprint 2017*. 256.

[32] Tianfei Zhou, Wenguan Wang, Si Liu, Yi Yang, and Luc Van Gool. 2021. Differentiable Multi-Granularity Human Representation Learning for Instance-Aware Human Semantic Parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1622–1631.

[33] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.

[34] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 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*. 5802–5810.