Information Sciences 614 (2022) 548–562

Contents lists available at ScienceDirect

Information Sciences

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

HR-PrGAN: High-resolution story visualization with progressive generative adversarial networks

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

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

ARTICLE INFO

Article history:
Received 6 May 2021
Received in revised form 25 August 2022
Accepted 15 October 2022
Available online 21 October 2022

Keyword:
Story visualization
Generative adversarial networks
High-resolution
Unconditional distribution
Fine-grained feature

ABSTRACT

Generating a series of images to describe a story of multiple sentences is a challenging task in computer vision as it needs to consider both the image-level precision and story-level consistency. Existing methods usually focus on the consistency between images at the cost of the fine-grained object details. We propose a progressive adversarial learning algorithm, termed HR-PrGAN, to achieve high-resolution image sequences with rich details by decomposing the problem of generating into multiple stages. Specifically, HR-PrGAN has two stages, where the Coarse-grain Stage generates a series of coherent coarse-grained images from both the story and context embeddings, and an additional unconditional loss is proposed to restrict their deformation and preserve the object contours and layouts. Subsequently, the Refinement Stage further refines the series of coherent coarse-grained images by injecting the story-level text embeddings and preserving the image-level details via a Coarse-grained feature Supplementary Module (CSM). Moreover, two commonly-used datasets, i.e. the CLEVR-SV and PororoSV datasets, are applied to evaluate the proposed method. Extensive experiments demonstrate that the proposed model significantly outperforms state-of-the-art methods in terms of image anti-deformation, fine-grained feature synthesis and human perception based image quality evaluation.

1. Introduction

Story visualization refers to providing a multi-sentence paragraph description, the story is visualized by generating a meaningful and coherent sequence of images, each image for one sentence [16]. It has recently received attention and has tremendous potential practical applications, including storyboard generation, book illustration editing, graphic design, etc. A good visualization puts the color, spatial relationships, and fine-grained object details inside the story world to efficiently assist the text semantics comprehension.

Similar to the text-to-image generation task, the story visualization model learns the mapping from text distribution to image distribution. The difference is that the sequence of images generated by the story visualization model needs to consistently and coherently depict the whole story while displaying the logic of the storyline. Specifically, the appearance of objects and the layout in the background must evolve coherently as the story progresses. This means that story visualization requires understanding and reasoning on both individual images and the whole story and that both the image and story distribution must be modeled simultaneously during the generation process. Although some impressive methods [16,17] have been presented, there remain following two challenges. 1) There are problems of deformation and missing key characters in

* Corresponding author.

https://doi.org/10.1016/j.ins.2022.10.083
0020-0255/

the generated results. Due to the lack of consistency between global story information and single sentence information in text encoding, the distribution based on the whole story will inevitably interfere with the distribution of the individual image. Therefore, existing methods are difficult to generate images with clear semantics (see the first row of Fig. 1). 2) Current methods can only generate 64 × 64 low-resolution images, and these images lose a lot of detailed information. In text-to-image tasks, the multi-stage generative model [35,36] refines coarse-grained images by extracting text semantics multiple times. This type of model has more advantages over single-stage models since it is challenging for single-stage models to synthesize higher resolution images (e.g., 128× 128) without providing additional spatial annotations of objects. However, the difference between story visualization and text-to-image is that story visualization needs to consider implicit context information. Directly applying multi-stage text-to-image methods for story visualization may result in losing coarse-grained features when refining the coarse-grained image sequence used to describe the story. This will lead to more severe deformation and layout errors.

In this paper, to tackle the above challenges, we propose a novel framework, namely, High-Resolution Progressive Generative Adversarial Networks (HR-PrGAN), to achieve high-resolution image sequences with rich details by decomposing the problem of generating into multiple stages. Specifically, in HR-PrGAN, we gradually improve the generated image through the Coarse-grain Stage and the Refinement Stage network. The Coarse-grain Stage generates a sequence of low-resolution images to describe a story of multiple sentences. Each image in story provides coarse-grained information such as object contours and layout. In this stage, HR-PrGAN improves the anti-deformation ability of characters by adding additional unconditional terms to simultaneously approximate the unconditional image-only distribution and the image distribution conditioned on text descriptions and provides high-quality initialization input for the next Refinement Stage. The Refinement Stage is used to add details to the output of the Coarse-grained Stage. To better preserve coarse-grained features in the process of simultaneously approximating the distribution at image and story level, we design a Coarse-grained feature Supplement Module (CSM) which allows quick access to the underlying information through the network. This means that during the generation process of the up-sampling layer, its expression of coarse-grained features will be subject to additional constraints from the down-sampling stage. CSM can reduce the mutual interference between image distribution and story distribution, supplement coarse-grained information lost in the refinement process such as object contours and spatial layout of the character and generate high-resolution story sequences with more fine-grained features (see Fig. 1).

In summary, the main contributions of the proposed method are shown as follows:

- We propose a two-stage generative adversarial model, termed HR-PrGAN, for story visualization, which aims at handling the trade-off between image-level precision and story-level consistency.
- Two methods, including unconditional terms loss and the CSM, are proposed to alleviate the deformation and preserve the object details of the generated image sequences.
- Comprehensive experimental results manifest that HR-PrGAN significantly outperforms the state-of-the-art methods on CLEVR-SV and Pororo-SV datasets.

Fig. 1. Taking text paragraphs as input, HR-PrGAN can generate high-quality image sequences with details compared to the story visualization model StoryGAN [16] and the single image generation model StackGAN [35].

In the remainder of this paper, we first review related work and preliminaries in Section 2 and Section 3, respectively. Then, we introduce HR-PrGAN in Section 4. In Section 5, extensive experiments are conducted to evaluate the proposed method. Finally, we move to a conclusion in Section 6.

2. Related work

Retrieval based Text-to-Image Synthesis In early research, text-to-image synthesis is carried out through the process of retrieving and supervising learning the combination of scenes and objects [38,7,10,24]. By extracting keywords with rich semantic information in the text and searching for the most suitable image conditioned on the keywords, the connection between the text and the image can be established. The optimization process focuses on the layout of generated results conditioned on text and images. The major limitation of the traditional text-to-image synthesis approaches is that they lack the ability to generate new image content; they can only change the specific objects of the given images. This means that this type of work often relies on complete datasets annotations and lacks generalization capabilities.

GAN-based Text-to-Image Synthesis With the success of Generative Adversarial Networks(GANs) [8] in the image generation modeling, a large number of deep learning-based methods have been proposed to handle text-to-image synthesis task. Mansimov et al. [18] propose the alignDRAW model based on Deep Recurrent Attention Writer (DRAW) [9] to iteratively draw image patches while attending to the relevant words in the caption. Reed et al. [26] propose a parallelized PixelCNN [22] to synthesize images from text that allows more efficient inference by modeling certain pixel groups as conditionally independent. Nguyen et al. [20] extend DGN-AM [21] by introducing an additional prior on the latent code which can generate images from captions. Based on a series of improvements and expansions of GANs [2,1,19,4,28,29,34], Reed et al. [25] first apply conditional GAN to synthesize plausible images from text descriptions. Dong et al. [5] adopt the approach of Kiros et al. [15] to find joint embedding space for both image and text embeddings.

Since previous generative models cannot generate high-resolution images, StackGAN [35] and StackGAN++ [36] decompose the problem of text to image synthesis into more tractable sub-problems with two stages. Zhang et al. [37] introduce accompanying hierarchical-nested adversarial objectives inside the network hierarchies, which regularizes mid-level representations and assists generator training to capture the complex image statistics. Since conditioning on the global sentence coding may result in low-quality images, AttnGAN [32] synthesizes fine-grained details at different subregions of the image by paying attention to the relevant words in the natural language description. Based on this, MirrorGAN [23] adopts a mirror structure, which reversely learns whether generated results are indeed consistent with the input texts. HDR-GAN [39] introduce multi-granularity feature disentangled representations into text-to-image synthesis, and image information is represented comprehensively through explicitly disentangling multi-granularity features.

Story visualization Story visualization is a new branch of text-to-image synthesis. Instead of generating static images, another line of text-to-image synthesis research focuses on sequences of images. In this context, the synthesized sequence of images describes a story written in a multi-sentence paragraph. Li et al. [16] first propose this task and name it story visualization. Using multi-level text encoders and discriminators, StoryGAN learns a low-dimensional embedding vector of a story to maintain the consistency of text semantics and to generate sequences. Story visualization technology greatly expands the application range of text-to-image synthesis due to its potential for providing beneficial properties and opportunities. Maharana et al. [17] add a dual learning framework that utilizes video captioning to reinforce the semantic alignment between the story and generated images. Zeng et al. [33] use a variety of textual alignment modules and a patch-based image discriminator to improve the semantic relevance and overall quality of the images. Song et al. [30] introduce the segmentation images during training to enhance the model being aware of the figure-ground components. However, because there is no global level of similarity between story distribution and image distribution, the current story visualization methods cannot well inherit the technical solution of static image generation, the generated results are limited to low-resolution levels and are more prone to deformation. To this end, we propose a progressive story visualization network that uses the loss which contains unconditional terms and Coarse-grained feature Supplement Module(CSM) to obtain high-resolution story sequences with fine-grained features.

3. Problem formulation

Story visualization aims at generating an image sequence $\widehat{X} = [\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_t]$ based on the semantics of the text sequence $S = [s_1, s_2, \cdots, s_t]$. In the following, we assume s_t and S are both encoded feature vectors using a pre-trained sentence encoder [3].

Specifically, the story visualization task takes the story paragraph $S = [s_1, s_2, \cdots, s_t]$ as input, where s_t denotes the t-th sentence in S. The text encoder $E(\cdot)$ learns to encode the story S into a low-dimensional embedding vector φ_t while keeping the continuity of the story. Then, a two-layer recurrent neural network (RNN) based Context Encoder encodes input sentence s_t and its contextual information into a vector o_t for each time point t. Using encoded text o_t as a condition, the generation model generates images successively $\hat{x}_t \leftarrow G(z, s_t), t \in (1, 2, \ldots, n)$, finally obtains an image sequence $\widehat{X} = [\hat{x_1}, \hat{x_2}, \cdots, \hat{x_t}]$ corresponding to the content of story S. The results we expect can not only achieve local semantic consistency of the single image $\hat{x_i}$ and sentence s_i, but also maintain global consistency between the dynamic scenes and characters. In addition to the semantic consistency, image sequences should also convey content information as much as possible, such information is

often expressed in the form of fine-grained features of objects and backgrounds. Therefore, how to obtain high-resolution image sequences with rich and fine-grained semantics is the focus of our research in this paper.

4. Approaches

4.1. Overview

Our story visualization model is designed for generating image sequences with contextual semantic connections to express coherent storylines. To generate high-resolution image sequences containing fine-grained features, we propose a simple yet effective High-Resolution Progressive Generative Adversarial Network, of which the model architecture is shown in Fig. 2.

In our model, instead of directly creating a high-resolution sequence of images conditioned on the multi-sentence paragraph, we first generate a low-resolution image sequence in our Coarse-grain Stage, which focuses on the expression of coarse-grained features. And then, based on coarse-grained images, the Refinement Stage uses the Coarse-grained feature Supplement Module to ensure that the coarse-grained features are preserved in the up-sampling process and uses text description information to enhance details further and add missing content.

Specifically, in the Coarse-grain Stage, the generator G first takes s_t as input and obtains the context encoding o_t through the Text Encoder proposed by [16]. Then, o_t is concatenated with a noise vector z, the fused information is converted to an underlying common semantic space of visual features by a perception layer and used to generate a $W_0 \times H_0$ image by a series of up-sampling blocks. o_t is continuously updated during image generation, and the generated image sequences $\hat{x}_t$ form a story S with context semantics. For the image discriminator D_I, in order to address the challenges of deformation and missing key characters mentioned in the **Introduction**, HR-PrGAN not only uses the joint features of the image and text to calculate the conditional loss, but also uses the encoded real-fake image pairs $\{(x_t, \hat{x}_t)\}$ to calculate the unconditional image loss. Unconditional image loss focuses on whether the generated result is close to the real image and will not be affected by the lack of consistency between global story information and single sentence information in text encoding. Similarly, for the story discriminator D_S, the story sequences S and image sequences X are concatenated and encoded in the same dimension. The element-wise product of the image and text sequence coding is input to a fully connected layer with sigmoid nonlinearity to predict whether it is a false story pair or a true story pair.

Similar to the previous stage, in the Refinement Stage, we use the Text Encoder to obtain the context encoding o_t. Meanwhile, the Coarse-grain Stage result $\hat{x}_t$ is fed into several down-sampling blocks to form an image encoding tensor. In addition, the residual blocks are also used in the generator of the Refinement Stage to learn multi-modal representations across image and text features. The Refinement Stage uses the CSM to ensure that the coarse-grained features are preserved in the up-sampling process, and uses text description information to enhance details further and add missing content. The generator in the Refinement Stage finally generates $W_1 \times H_1$ images with fine-grained features. For the discriminator, its structure is similar to that of the Coarse-grain Stage with only extra down-sampling blocks since the image size is larger in this stage.

4.2. Text encoder

We follow the standard design in story visualization tasks [16] and divide the Text Encoder into Story Encoder and Context Encoder.

During training, each sentence in story S is first encoded into an embedded vector s_t by the encoder [3] and the encoding of complete story S is obtained by the concatenated sentence encoding. Following the StackGAN [35], we use Conditional

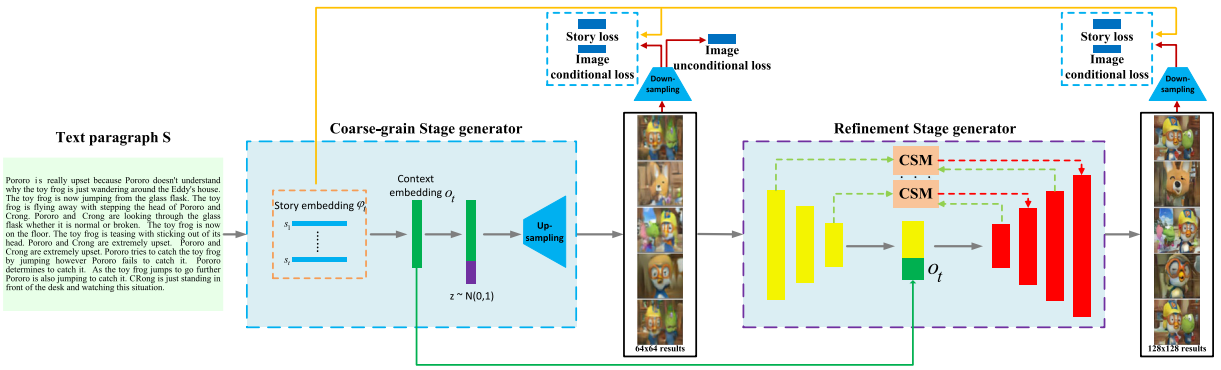

Fig. 2. The architecture of the proposed Story Visualization Network. The Coarse-grain Stage generator generates low-resolution image sequences containing coarse-grained features from context embedding o_t and noise z. Conditioned on Coarse-grain Stage results, the Refinement Stage generator uses the Coarse-grained feature Supplement Module (CSM) to preserve coarse-grained features and get realistic high-resolution image sequences.

Augmentation (CA) technology to augment training data and avoid overfitting by resampling the input story vector φ_t from an independent Gaussian distribution $\mathcal{N}(\mu(S), \Sigma(S))$, where $\mu(S)$ is the mean and $\Sigma(S)$ is the diagonal covariance matrix. Thus, the CA loss is defined as the Kullback–Leibler(KL) divergence between the standard Gaussian distribution and the trained Gaussian distribution. To enforce the smoothness over the conditional manifold in latent semantic space, we add the regularization term, which can be expressed as

$$\mathcal{L}_{KL} = KL(\mathcal{N}(\mu(S), \Sigma(S)) || \mathcal{N}(0, I)), \tag{1}$$

with the reparameterization trick [14], parameters of Story Encoder can be updated during back propagation(BP).

Taking φ_0 as the initial input, the goal of the Context Encoder is to capture contextual information during the sequential generation of images. The Context Encoder we use is a deep Recurrent Neural Network(RNN) containing two hidden layers [16]. The first layer is implemented using Gated Recurrent Unit(GRU) cell. The GRU cell accepts the original embedding vector s_t of the sentence and equidistant Gaussian noise ϵ_t as input, outputs the encoding vector i_t and memory vectors g_t at step t. The second layer takes the output i_t from the first layer and the resampled story vector φ_t as input, uses convolution operators to generate o_t that reflects the change of potential context information. Context Encoder works as follows:

$$i_t, g_t = GRU(s_t, \epsilon_t, g_{t-1}), \tag{2}$$

$$z_t = \sigma_z(W_z i_t + U_z \varphi_{t-1} + b_z), \tag{3}$$

$$r_t = \sigma_r(W_r i_t + U_r \varphi_{t-1} + b_r), \tag{4}$$

$$\begin{aligned} \varphi_t &= (1 - z_t) \odot \varphi_{t-1} \\ &+ z_t \odot \sigma_\varphi(W_\varphi i_t + U\varphi(r_t \odot \varphi_{t-1}) + b_\varphi), \end{aligned} \tag{5}$$

$$o_t = \text{Filter}(i_t) \bullet \varphi_t, \tag{6}$$

Eq. (3) and Eq. (4) represent the output of update and reset gates. σ_z, σ_r and σ_φ are sigmoid non-linearity functions. The output of the Context Encoder contains both global information from i_t and local information from φ_t.

4.3. Coarse-grained feature supplement module

Previous story visualization works [16,17] can only generate low-resolution images. The existing refinement stage scheme in the text-to-image task obtains high-resolution images with more details by up-sampling step by step from a unified embedding encoding information of both coarse images and text. However, in the story visualization task, the text embedding encodes both global story and local sentence information, while the generator needs to model the image distribution and the story distribution simultaneously, leading to the original encoder-decoder network being disadvantageous for sharing low-level information between input and output. Therefore, when the existing high-resolution image modeling method [35,23,36] for a single image is used directly to generate an image sequence, the coarse-grained features in the initial stage cannot be effectively preserved, resulting in the further degraded generated results. Inspired by image segmentation [27] and image style translation tasks [12], we design a Coarse-grained feature Supplement Module, which directly obtains and fuses coarse-grained features as an effective supplement to the up-sampling process to effectively preserve the coarse-grained features such as object contours and spatial layout.

As shown in Fig. 3, our Coarse-grained feature Supplement Module takes two inputs: (1) Coarse-grained features obtained by down-sampling of images generated in the Coarse-grained stage, (2) Hidden features $h \in \mathbb{R}^{C \times H \times D}$ in the up-sampling process of unified embedding encoding of both coarse image and text. The coarse-grained features and the hidden features are

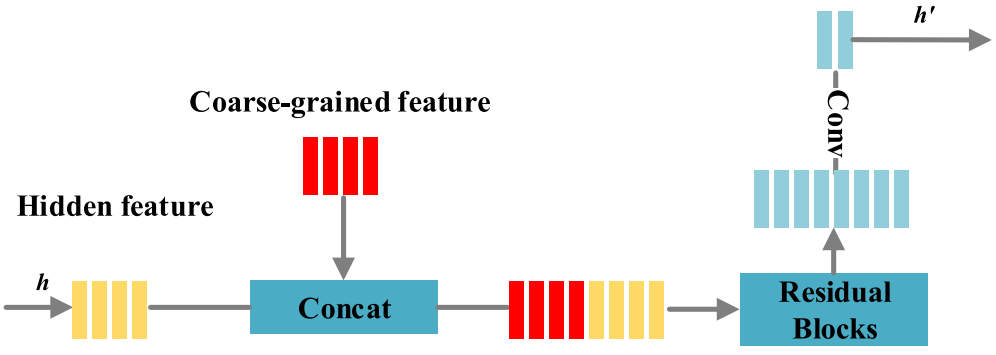

Fig. 3. The architecture of the Coarse-grained feature Supplement Module(CSM) in HR-PrGAN.

concatenated along the channel dimension. The encoded coarse-grained features coupled with hidden features are fed into several residual blocks, which are designed to learn multi-level representations across image features. Finally, the fused feature vector is up-sampled and further processed with two convolutional layers to produce the hidden features $h\prime \in \mathbb{R}^{C/2\times 2H\times 2D}$ of the next layer. By providing image constraint information beyond text-driven, the CSM can effectively preserve the coarse-grained features in the up-sampling process, and can effectively reduce the problem of output quality degradation caused by the loss of coarse-grained features in the refinement stage.

4.4. Objective function

Different from the previous work that considers only the conditional distribution of images, HR-PrGAN jointly approximates the conditional and unconditional image distribution when generating images. Let z be a noise vector randomly sampled from the Gaussian distribution, the image loss $\mathscr{L}_{Image}$ is defined as

$$\begin{aligned}\mathscr{L}_{Image} &= \sum_{t=1}^{T}\left(\mathbb{E}_{(x_t,s_t)}[\log D_{I\_C}(x_t,s_t,\varphi_t)]+\mathbb{E}_{(z_t,s_t)}[\log(1-D_{I\_C}(G_C(z_t,s_t),s_t,\varphi_t))]\right)\\ &+\sum_{t=1}^{T}(\mathbb{E}_{x_t}[\log D_{I\_C}(x_t)]+\mathbb{E}_{z_t}[\log(1-D_{I\_C}(G_C(z_t)))]),\end{aligned} \tag{7}$$

where the first two terms defined as conditional loss are used to determine whether the image matches the input sentence, and the second two terms are additional unconditional loss, which determines whether the image is real or fake. We find that applying unconditional constraints in HR-PrGAN to make the model pay more attention to the quality of the generated images, can reduce the distortion of the generated images and generate coarse-grained features with higher discrimination compared to methods [16,17,35] using only conditional losses.

The story loss $\mathscr{L}_{Story}$ is used to enhance the global consistency of the generated image sequence. When calculating the story loss, the generated image sequence $\widehat{X}=[\hat{x}_1,\hat{x}_2,\cdots,\hat{x}_t]$ and the text embedding sequence $S=[s_1,s_2,\cdots,s_t]$ are concatenated and encoded into the corresponding feature vectors. The element-wise product of the encoded image sequence and story sequence is fed to the fully connected layer as input, and the sigmoid nonlinear function is used to predict whether the image is generated or real. The story loss $\mathscr{L}_{Stoty}$ is defined as

$$\begin{aligned}\mathscr{L}_{Story} &= \mathbb{E}_{\left(\widehat{X},S\right)}\left[\log D_{S\_C}\left(\widehat{X},S\right)\right]\\ &+\mathbb{E}_{(z,S)}\left[\log\left(1-D_{S\_C}\left([G_C(z_t,s_t)]_{t=1}^{T},S\right)\right)\right].\end{aligned} \tag{8}$$

The encoding result of image sequences and text embedding sequences is implemented by a convolutional network and multi-layer perception. Therefore, the objective function for Coarse-grain Stage is

$$\min_{G_C}\max_{D_{I\_C},D_{S\_C}} \lambda_1\mathscr{L}_{Image}+\lambda_2\mathscr{L}_{Story}+\mathscr{L}_{KL}, \tag{9}$$

where G_C denotes the generator, D_{I_C} and D_{S_C} stand for the image and story level discriminator in Coarse-grain Stage respectively, λ_1 and λ_2 are used to control the weight of each term, $\mathscr{L}_{KL}$ is defined by Eq. (1).

The training process in the Refinement Stage conditions on the Coarse-grain Stage results and text embedding, and obtains high-resolution realistic image sequences. The image and story loss of GAN in the Refinement Stage can be defined as

$$\begin{aligned}\mathscr{L}_{Image} &= \sum_{t=1}^{T}\left(\mathbb{E}_{(\hat{x}_t\prime,s_t)}[\log D_I(\hat{x}_t\prime,s_t,\varphi_0)]+\mathbb{E}_{(\hat{x}_t,s_t)}[\log(1-D_I(G_R(\hat{x}_t,s_t,),s_t,\varphi_0))]\right)\\ &+\sum_{t=1}^{T}(\mathbb{E}_{\hat{x}_t\prime}[\log D_I(\hat{x}_t\prime)]+\mathbb{E}_{\hat{x}_t}[\log(1-D_I(G_R(\hat{x}_t)))]),\end{aligned} \tag{10}$$

$$\begin{aligned}\mathscr{L}_{Story} &= \mathbb{E}_{\left(\widehat{X}\prime,S\right)}\left[\log D_S\left(\widehat{X}\prime,S\right)\right]\\ &+\mathbb{E}_{(\hat{x}_t,S)}\left[\log\left(1-D_S\left([G_R(\hat{x}_t,s_t)]_{t=1}^{T},S\right)\right)\right].\end{aligned} \tag{11}$$

Different from the original formulation of Coarse-grain Stage, the random noise z is not used in this stage with the assumption that the randomness has already been preserved by the Coarse-grain Stage results $\hat{x}_t$. Referring to the expression of Coarse-grained Stage, the objective function of Refinement Stage is

$$\min_{G_R}\max_{D_{I\_R},D_{S\_R}} \lambda_3\mathscr{L}_{Image}+\lambda_4\mathscr{L}_{Story}+\mathscr{L}_{KL}, \tag{12}$$

where G_R denotes the generator, D_{I_R} and D_{S_R} stand for image and story level discriminator in Refinement Stage respectively. λ_3 and λ_4 are the hyper-parameters for balancing each term.

4.5. HR-PrGAN training

Based on above model structure, we give the pseudo-codes for training as Algorithm 1. The parameters of the image and story discriminators are updated independently in two**for** loops, and the two loops also update the generator parameters together.

Algorithm 1: Progressive Story Visualization

Input: The image sequence $X = [x_1, x_2, \cdots, x_t]$ and the corresponding encoded text embedding sequence $S = [s_1, s_2, \cdots, s_t]$. **Output:** Generator parameters ϕ_C and ϕ_R for G_C, G_R and discriminator parameters $\sigma_{IC}, \sigma_{IR}, \sigma_{SC}, \sigma_{SR}$ for $D_{I\_C}, D_{I\_R}, D_{S\_C}, D_{S\_R}$.

For *iter* = 1 to *max_iter* **do**
For $iter_I$ = 1 to k_I **do**
Sample s_t, S, x_t from the training set.
Compute φ_0 as the initialization of Context Encoder and KL regularization term is expressed as Eq. (1).
Generate a single output image $\hat{x}_t$.
Update ϕ_C and σ_{IC} for Coarse-grain Stage.

Sample s_t, S, x_t from the training set.
Sample $\hat{x}_t$ from Coarse-grain Stage.
Generate a single output image $\hat{x}_t'$.
Update ϕ_R and σ_{IR} for Refinement Stage.
end for
For$iter_S$ = 1 to k_S**do**
Sample S, X from the training set.
Compute φ_0 and update at each time step t.
Generate an image sequence $\widehat{X}$.
Update ϕ_C and σ_{SC} for Coarse-grain Stage.

Sample S, X from the training set.
Sample $\widehat{X}$ from Coarse-grain Stage.
Generate an image sequence $\widehat{X}'$.
Update ϕ_R and σ_{SR} for Refinement Stage.
end for
end for

5. Experiments

We conduct extensive quantitative and qualitative experiments to evaluate the proposed method. Our network output the results of the Coarse-grained Stage and the Refinement Stage simultaneously(shown as Fig. 4), and is compared with the state-of-the-art story visualization method [16,17] and text-to-image synthesis method [35,36]. We also verify the ability of HR-PrGAN to generate long image sequences. The generated results use the network models released by their author, and the encoding of the text follows the work [16]. In addition, we also study the overall design of the proposed model and the role of important components. We first modify HR-PrGAN to investigate the role of adding unconditional terms in different stages or losses. Then, we study different refinement strategies to investigate whether the proposed coarse-grained feature supplement module(CSM) is beneficial.

Fig. 4. The results of the Coarse-grained Stage and the Refinement Stage from our model.

5.1. Implementation details

The up-sampling blocks in the generator consist of a nearest-neighbor up-sampling followed by a 3×3 stride 1 convolution. Batch normalization [11] and ReLU [6] activation are applied after every convolution. The residual blocks consist of two 3×3 stride 1 convolutions, batch normalization and ReLU. Single residual block is used in Refinement Stage. The down-sampling blocks in discriminator consist of 4×4 stride 2 convolutions, batch normalization and LeakyReLU, except that the first one does not have batch normalization. The double convolution block consists of two 3×3 stride 1 convolutions. The first convolution halves the dimension of the feature vector. The second convolution keeps the dimensions of input and output vector consistent.

By default, $W_0 = H_0 = 64$, $W_1 = H_1 = 128$ and noise z is sampled from a 100-dimensional unit normal distribution. In the Coarse-grain Stage, we set image batchsize to 120, story batchsize to 24 and iteratively train 120 epochs. In the Refinement Stage, we set image batchsize to 30, story batchsize to 6 and iteratively train 100 epochs. All networks are trained using Adam solver for parameter update with an initial learning rate of 0.0002. The learning rate is decayed to 1/2 of its previous value every 20 epochs.

5.2. Datasets and evaluation metrics

We verified the proposed method on multiple datasets using multiple evaluation metrics.

Datasets: We use CLEVR-SV [16] dataset and Pororo-SV [16] dataset to conduct experiments.

The CLEVR-SV dataset is modified from the CLEVR [13] dataset which is used for the visual question answering. By adding a three-dimensional graphic object with different sizes and colors to the image each time, we generate image sequences with contextual semantic relations. The dataset contains 10,000 training image sequences and 3,000 test image sequences with 4 images per sequence. The text mainly describes the positional relationship between different objects and the characteristics of the newly added object. We perform ablation experiments on original CLEVR-SV. Furthermore, to verify the ability of our network to generate long image sequences, we regenerate the CLEVR-SV dataset, where each sequence contains 8 images, the regenerated dataset also contains 10,000 training image sequences and 3,000 test image sequences. The CLEVR-SV dataset is applied only for analyzing coarse-grained results.

The Pororo-SV dataset is obtained by modifying the Pororo dataset using equidistant sampling (sampling rate is 30 Hz). It contains 15336 description-story pairs, where 13000 pairs are used for training, the remaining 2336 pairs for testing. The text description provides information about objects contained in the image and what is happening. Compared with the CLEVR-SV dataset, the Pororo-SV dataset contains richer scenes and objects. In Pororo-SV, establishing semantic connections is through complex scene switching and object interaction, which is a more difficult task for story visualization.

Evaluation Metrics: Following previous works [16], we choose Structural Similarity Index Measure(SSIM) score [31] between ground truth and the generated images for quantitative evaluation. Here, SSIM is used to determine whether the generated images are aligned with the input description. SSIM is designed by modeling any image distortion as a combination of three factors: loss of correlation, luminance distortion and contrast distortion. Note that though this is a generative task, using SSIM to measure the structural similarity is reasonable because the description of the ground truth and generated images should be almost no variation. We calculate and compare the SSIM scores of the results from the different coarse-grain stage methods and the different refinement stage methods.

Specifically, for meaningful and fair comparisons with previous methods, the SSIM score is computed in two settings. In the first setting, 128×128 images produced by StackGAN [35], StackGAN++ [36], HR-PrGAN and 64×64images yielded by StoryGAN [16], DUCO-StoryGAN [17] are used directly to compute SSIM score. The different models are compared directly using their generated images, which have different resolutions. In the second setting, all generated images are resized to 64×64 before computing SSIM* score.

SSIM score is sensitive to local structural changes, slight changes in details will cause a significant drop in SSIM scores, but these changes in details often do not affect the overall semantic expression of the story sequences. Considering that the purpose of story visualization is to show the semantics of the text to humans rather than machines, we also use human evaluation methods to supplement quantitative evaluation experiments to more intuitively reflect the ability of transmitting semantic information to readers. We perform global indicators and overall quality ranking based human evaluation studies on Pororo-SV dataset. For all tasks, we use 150 generated image sequences sampled from the test set, with each assigned to 10 workers to reduce human variance. The order of the options within each assignment is shuffled to make a fair comparison.

Specifically, we first evaluate the generated results from the three global indicators: visual quality, consistency, and relevance. High visual quality means the generated images look visually appealing, rather than blurry and difficult to be understood. High consistency means the generated images are consistent with each other, have a common topic hidden behind, and naturally form a story, rather than look like independent images. Relevance means that the generated image sequences can accurately reflect the global semantics of the story. Among them, the visual quality indicator is for comparison between single images, consistency and relevance indicators are for comparison between image sequences.

5.3. Comparison with existing methods

To verify the effectiveness of HR-PrGAN, we present both quantitative and qualitative experiments. First, we compare the generation results of HR-PrGAN and their corresponding Structural Similarity Index Measure (SSIM) score with the state-of-the-art story visualization methods [16,17] and the state-of-the-art text-to-image synthesis methods [35,36] on Pororo-SV dataset containing complex scenes and character details. Then, we present the results of a human perceptual test to evaluate the global indicators and overall quality ranking. We also verify the ability of our network to generate long image sequences on the regenerated CLEVR-SV dataset.

5.3.1. Qualitative analysis

Representative examples of StoryGAN, DUCO-StoryGAN, StackGAN, StackGAN++ and HR-PrGAN are compared in Fig. 5. As shown in Fig. 5, the results generated by the original StoryGAN can maintain the consistency of the image sequence, but these results are in low resolution, and the characters are deformed and lack sufficient details. DUCO-StoryGAN improves the semantic alignment of the generated images with the input story by introducing dual learning via video redescription. However, from the results, DUCO-StoryGAN does not significantly improve the quality of the generated image sequence. StackGAN and StackGAN++ stack multi-stage networks which can generate high-resolution images. However, since the story visualization task needs to model the image and the story distribution simultaneously, it is difficult for the original encoder-decoder network to preserve coarse-grained features such as object contours and spatial layout in the refinement stage, which leads to serious deterioration of the generated results. In comparison, our HR-PrGAN improves the output of the Coarse-grained stage and adds additional image constraint information beyond text through the Coarse-grained feature Supplement Module(CSM), which can generate high-resolution image sequences that effectively preserve coarse-grained features and contain rich details.

Previous methods use datasets containing only a small number of images per image sequence (5 images in Pororo-SV and 4 in CLEVR-SV). To verify the ability of our network to generate long image sequences, we conduct experiments based on the

Fig. 5. Example results by StoryGAN[16], DUCO-StoryGAN[17], StackGAN[35], StackGAN++[36] and our HR-PrGAN conditioned on multi-sentence paragraph from Pororo-SV dataset.

regenerated CLEVR-SV dataset with image sequences twice as long as the original. 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. As shown in Fig. 6, our model still achieves semantically explicit results when generating semantically coherent long sequences of images.

5.3.2. Quantitative analysis

We first calculate the Structural Similarity Index Measure(SSIM) score of the compared methods. Calculation results are reported in Table 1. When comparing the original outputs of different models, HR-PrGAN achieves a 60.5% improvement (from 0.071 to 0.114) compared to StoryGAN and 37% improvement (from 0.083 to 0.114) compared to DUCO-StoryGAN, which indicates that the progressive network can rectify defects in coarse-grained images to make it closer to the ground-truth. Compared with StackGAN and StackGAN++, HR-PrGAN achieves a 225% (from 0.034 to 0.114) and 165% (from 0.043 to 0.114) improvement respectively. The huge gap in SSIM scores indicates that the preservation of coarse-grained features in refinement stage is the basis for the definite expression of image semantics. When we compare image sequences of different models at the same resolution of 64×64, the SSIM score of our HR-PrGAN still achieves an approximate improvement compared to other models.

We also use human evaluation methods to supplement quantitative evaluation experiments to fully evaluate the generated results of story visualization. We believe that ground truth has the best global index evaluation results, thus we give the corresponding ground truth of each group of texts as references for workers during the human evaluation process. Some examples of satisfactory and unsatisfactory HR-GAN generation results among the human evaluation are shown in Fig. 7. Among the unsatisfactory HR-GAN generation results, the examples in the first row have poor consistency and relevance, and the examples in the second have poor visual quality. This shows that human evaluation can objectively reflect the generation quality of image sequences.

In human evaluation experiments, we ask the workers to decide the best results of different methods based on global indicators. The percentage of the generated results of various methods that are selected as the best in each indicator is summarized in Table 2. The performance of HR-PrGAN in all evaluation indicators is much better than other methods and the standard error on these estimates is also small.

Next, we ask workers to rank the results of different compared methods on their overall quality. This test is closer to the application scenario of story visualization technology, reflecting the worker's perception of the generated result based on daily intuition. Results are summarized in Table 3. HR-PrGAN achieves the highest average rank, the standard error of this experiment is still very small, so we are confident the results of HR-PrGAN are more acceptable to humans. .

5.4. Design analysis

We found that the effects of unconditional terms on image sequences and single images are not consistent in our experiments, thus we analyze the impact of different loss function designs. In Fig. 8, (a) shows the generation results obtained by adding unconditional terms only to the image loss $\mathscr{L}_{Image}$, (b) shows the generation results obtained by adding unconditional terms to both image loss $\mathscr{L}_{Image}$ and story loss $\mathscr{L}_{Story}$. The features in (b) are stacked together in complex scenes (e.g., the facial features of people in the red bounding box results). This is because when calculating the story loss, the global coding

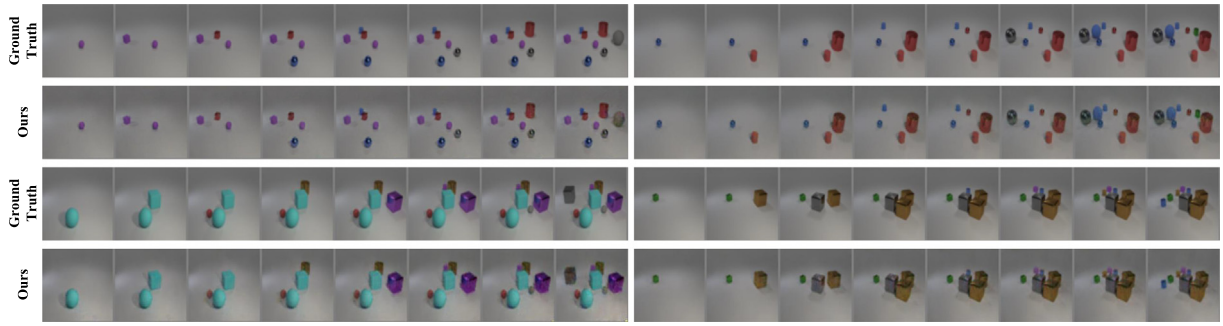

Fig. 6. Example results for long image sequence generation by our proposed HR-PrGAN on CLEVR-SV dataset.

Table 1
Results of SSIM and SSIM* comparison on Pororo-SV dataset.

	StoryGAN	DOCU-StoryGAN	StackGAN	StackGAN++	**HR-PrGAN**
SSIM	0.071	0.083	0.034	0.043	0.114
SSIM*	0.071	0.083	0.030	0.038	0.097

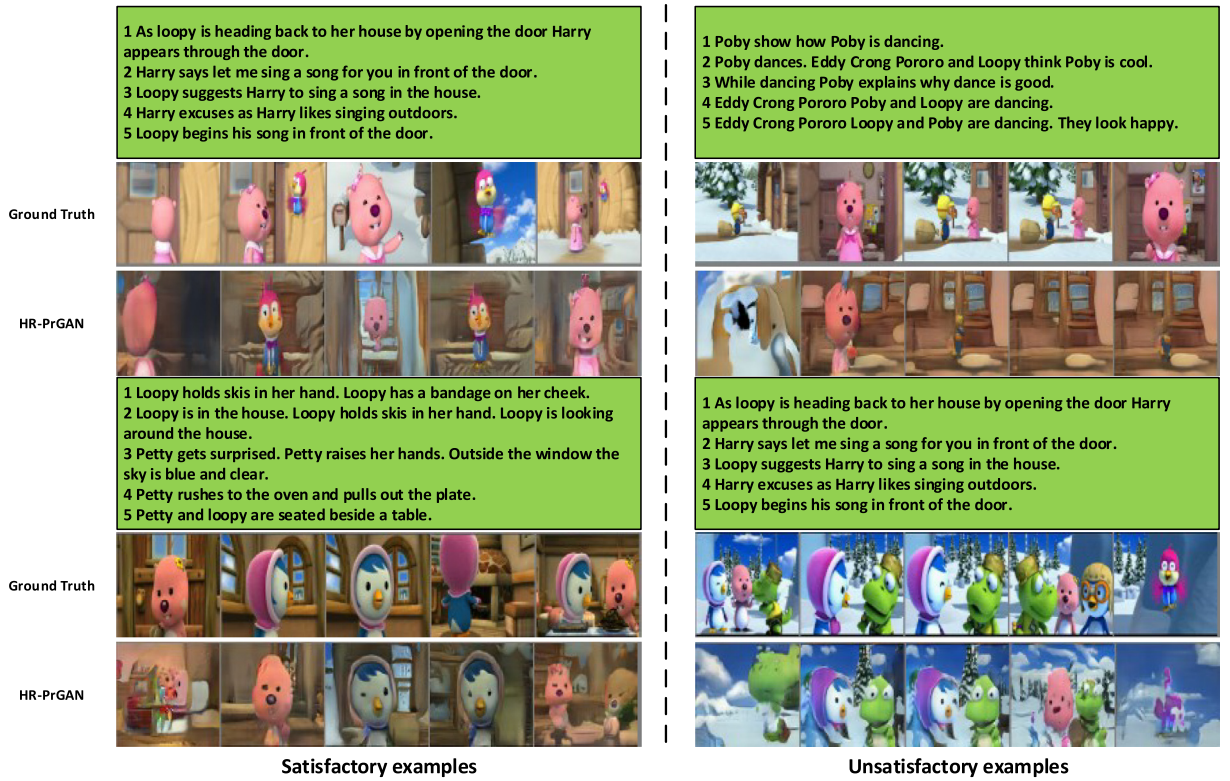

Fig. 7. Examples of satisfactory and unsatisfactory HR-GAN generation results among the human evaluation. On the left is the satisfactory examples and on the right is the unsatisfactory examples.

Table 2
Results of global indicators from human evaluation. The ± denotes standard error on the metrics.

Choice (%)	Visual quality	Consistency	Relevance
StoryGAN	6.2±1.9	11.6±1.1	10.1±2.2
DUCO-StoryGAN	4.1±1.4	3.5±2.6	8.2±1.6
StackGAN	1.1±1.0	3.4±1.4	2.0±1.3
StackGAN++	2.7±1.4	4.3±2.1	4.0±2.9
HR-PrGAN	86.0±2.3	77.1±3.9	75.7±3.2

Table 3
Results of ranking-based human evaluation. The ± denotes standard error on the metrics.

Method	StoryGAN	DUCO-StoryGAN	StackGAN	StackGAN++	**HR-PrGAN**
Rank	2.43±0.14	2.87±0.13	4.05±0.05	4.32±0.07	1.29±0.05

is generated from image sequences through the element-wise product, and this cannot constrain a single image. Table 4 shows the results of SSIM scores among different loss function designs on Pororo-SV dataset, the design which only adds unconditional terms to the image loss $\mathscr{L}_{Image}$ achieves better performance. Therefore, in HR-PrGAN, we only add unconditional terms to the image loss $\mathscr{L}_{Image}$ as shown in Eq. (7) and Eq. (10).

5.5. Ablation study

5.5.1. The effect of unconditional terms

The result of the Coarse-grained stage as the initial input of the refinement stage can provide the contours of objects and the spatial layout of the image, which has an important influence on the final refinement result. HR-PrGAN obtains more stable output by adding additional unconditional terms to the loss. In order to test the effect of unconditional terms in

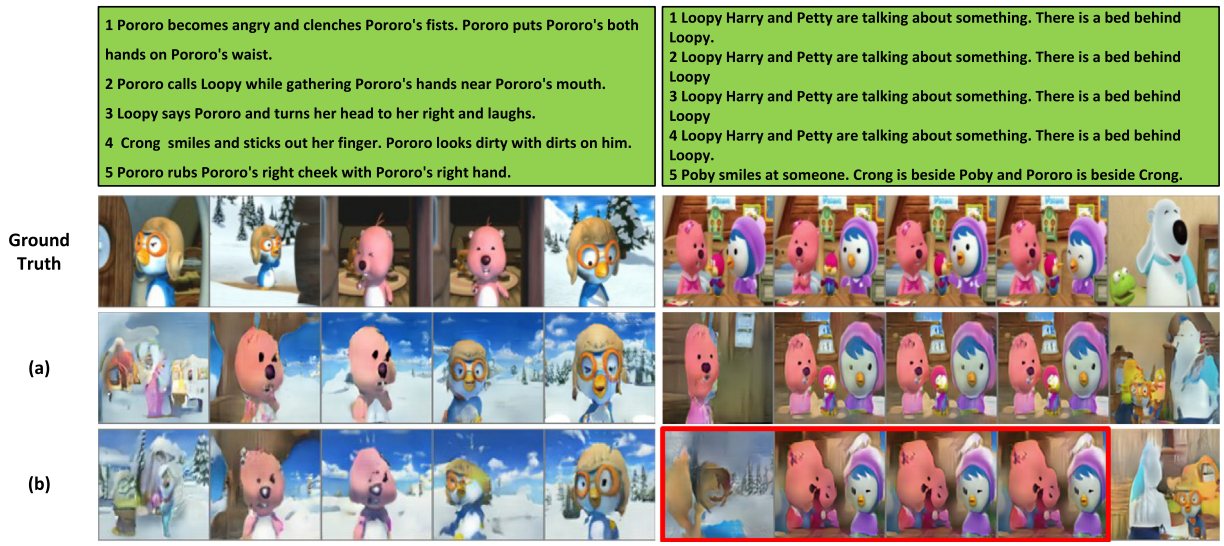

Fig. 8. Comparison among different loss function designs on Pororo-SV dataset. (a) The generation results are obtained by only adding unconditional terms to the image loss $\mathscr{L}_{Image}$. (b) The generation results are obtained by adding unconditional terms to both image loss $\mathscr{L}_{Image}$ and story loss $\mathscr{L}_{Story}$.

Table 4
Results of SSIM scores among different loss function designs on Pororo-SV dataset.

Design	Both image loss $\mathscr{L}_{Image}$ and story loss $\mathscr{L}_{Story}$	Only image loss $\mathscr{L}_{Image}$
SSIM	0.079	0.114

the story visualization task, we compare different methods for generating 64 × 64coarse-grained images. The SSIM scores of different methods with and without unconditional terms in the loss are reported in Table 5. Representative examples on Pororo-SV dataset and CLEVR-SV dataset are compared in Fig. 9 and Fig. 10.

As shown in Table 5, our HR-PrGAN achieves the best SSIM score on all datasets. Compared with the method without adding unconditional terms to the loss, HR-PrGAN achieves a 32.4% improvement (from 0.071 to 0.094) on Pororo-SV dataset, and a 6.8% improvement (from 0.672 to 0.718) on CLEVR-SV dataset. This indicates the coarse-grained results generated by HR-PrGAN are much closer to the semantics of the text. This is achieved through the better expression of coarse-grained features such as contours and spatial layout of the characters.

Fig. 9 gives the results comparison on Pororo-SV dataset. The method without adding unconditional terms to the loss uses story discriminator and Context Encoder, which can effectively improve the consistency of image sequences. However, due to the lack of effective constraints, the generated results often have serious deformations, which limitation is particularly obvious in the second example in Fig. 9. In contrast, Coarse-grain Stage of HR-PrGAN has a much higher quality than other methods. This shows the advantage of using additional unconditional loss to constrain the single image during the generation process.

The experimental results on the CLEVR-SV dataset are similar to those on the Pororo-SV dataset. As shown in Fig. 10, the method without adding unconditional terms to the loss still has errors in the expression of object attributes. In contrast, the Coarse-grain Stage of HR-PrGAN can generate more feasible images than the competitors.

5.5.2. The effect of CSM

In order to test the effect of Coarse-grained feature Supplement Module (CSM) in the refinement stage of the story visualization task, we compare the generation results of different coarse-grained image refinement strategies with and without CSM respectively on the Pororo-SV dataset.

Table 5
Results of SSIM score of different methods with and without unconditional terms in the loss on Pororo-SV and CLEVR-SV datasets.

SSIM	Pororo-SV	CLEVR-SV
w/o unconditional terms	0.071	0.672
HR-PrGAN(w/ unconditional terms)	0.094	0.718

Fig. 9. The performance of unconditional terms on Pororo-SV dataset.

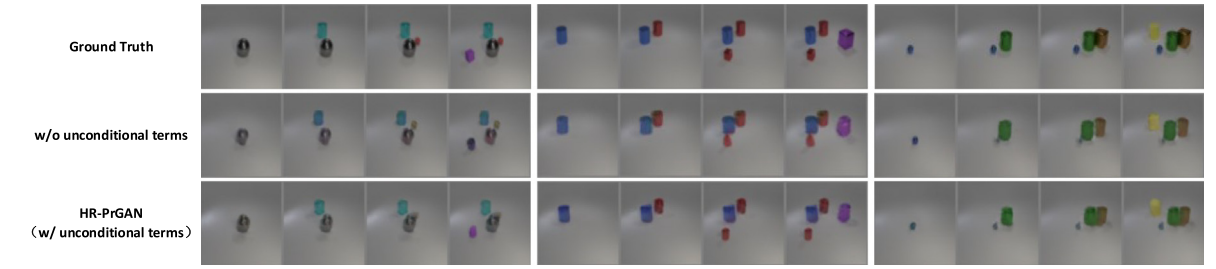

Fig. 10. The performance of unconditional terms on CLEVR-SV dataset.

Fig. 11 illustrates some examples with and without CSM on Pororo-SV dataset, where all refinement strategies take the same coarse-grained image sequences as input. As shown in Fig. 11, the multi-stage without CSM model refines the coarse-grained image by receiving the text encoding again, but in the story visualization task, simultaneously modeling global and local information makes it difficult to preserve coarse-grained features, and the expression of details is hardly improved. For

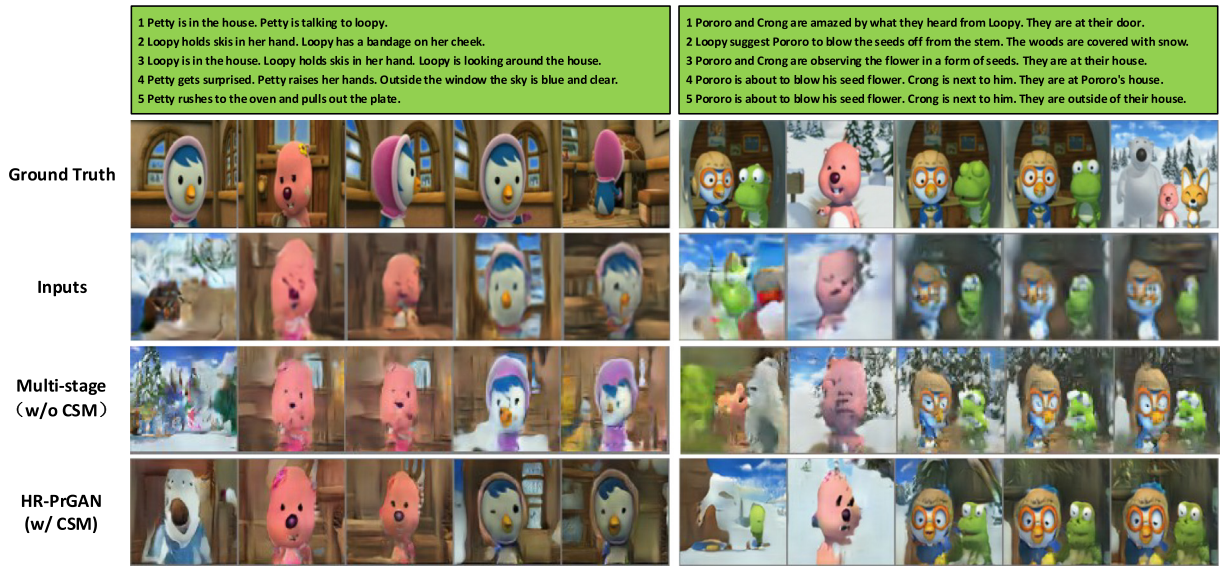

Fig. 11. The performance of CSM on Pororo-SV dataset.

Table 6
Results of SSIM score of refinement strategies with or without CSM on the Pororo-SV dataset.

Strategy	Multi-stage (w/o CSM)	HR-PrGAN (w/ CSM)
SSIM	0.43	0.114

instance, in the second, third, and fourth columns of Fig. 11, the facial details generated by the refinement strategy without CSM are misplaced. And in the complex scene generation task that contains multiple objects, deformation appears in the overall image and this makes the generated result difficult to express clear and continuous semantics (e.g., the right example in the third row of Fig. 11). In contrast, adding the CSM to the multi-stage generation model can generate high-resolution images with more convincing details to better reflect the corresponding text description and contextual semantic connections. Specifically, the objects in the results of the refinement stage have clearer boundaries and facial features, and the background of the image also shows more details. In addition, these fine-grained features maintain good consistency in the image sequences. Quantitative experiments in Table 6 show similar conclusions, the SSIM score of the results obtained by the strategy with CSM outperforms without CSM, which validates the effectiveness of CSM in refining coarse-grained image sequences.

6. Conclusion

We propose a story visualization model, HR-PrGAN, which uses text paragraphs containing multiple sentences to synthesize high-resolution image sequences with rich details. An unconditional loss is incorporated into HR-PrGAN to reduce image deformation. Besides, the Coarse-grained feature Supplementary Module is proposed to preserve coarse-grained features in the refinement process. Experiments on widely used datasets, i.e. the CLEVR-SV and PororoSV datasets, show that our model can synthesize high-resolution image sequences with more details and significantly improve SSIM scores. When sufficient computing power is available, higher resolution image sequences generated by stacking refinement stages of HR-PrGAN could be achieved. In the future, the model of story visualization task needs to be improved from the aspect of the expression of text's fine-grained semantics.

Acknowledgment

This work is supported in part by the Excellent Youth Scholars Program of Shandong Province (Grant no. 2022HWYQ-048).

CRediT authorship contribution statement

Pei Dong: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing - original draft. **Lei Wu:** Writing - review & editing, Methodology, Supervision. **Lei Meng:** Writing - review & editing, Supervision. **Xiangxu Meng:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Arjovsky, M., Bottou, L., 2017. Towards principled methods for training generative adversarial networks. arXiv preprint arXiv:1701.04862.
[2] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, International Conference on Machine Learning (2017) 214–223.
[3] Cer, D., Yang, Y., Kong, S.Y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., 2018. Universal sentence encoder. arXiv preprint arXiv:1803.11175.
[4] E.L. Denton, S. Chintala, R. Fergus, et al, Deep generative image models using a laplacian pyramid of adversarial networks, Adv. Neural Inform. Process. Syst. (2015) 1486–1494.
[5] Dong, H., Yu, S., Wu, C., Guo, Y., 2017. Semantic image synthesis via adversarial learning, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 5706–5714.
[6] Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks, in: Proceedings of the fourteenth international conference on artificial intelligence and statistics, pp. 315–323.
[7] Goldberg, A.B., Zhu, X., Dyer, C.R., Eldawy, M., Heng, L., 2008. Easy as abc? facilitating pictorial communication via semantically enhanced layout, in: CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning, pp. 119–126.
[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inform. Process. Syst. (2014) 2672–2680.
[9] K. Gregor, I. Danihelka, A. Graves, D. Rezende, D. Wierstra, Draw: A recurrent neural network for image generation, International Conference on Machine Learning, PMLR (2015) 1462–1471.
[10] Huang, Q., Gan, Z., Celikyilmaz, A., Wu, D., Wang, J., He, X., 2019. Hierarchically structured reinforcement learning for topically coherent visual story generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8465–8472.

[11] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, International Conference on Machine Learning (2015) 448–456.
[12] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134.
[13] Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R., 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2901–2910.
[14] Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.
[15] Kiros, R., Salakhutdinov, R., Zemel, R.S., 2014. Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539.
[16] Li, Y., Gan, Z., Shen, Y., Liu, J., Cheng, Y., Wu, Y., Carin, L., Carlson, D., Gao, J., 2019. Storygan: A sequential conditional gan for story visualization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6329–6338.
[17] Maharana, A., Hannan, D., Bansal, M., 2021. Improving generation and evaluation of visual stories via semantic consistency.
[18] Mansimov, E., Parisotto, E., Ba, J.L., Salakhutdinov, R., 2015. Generating images from captions with attention. arXiv preprint arXiv:1511.02793.
[19] Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.
[20] Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., Yosinski, J., 2017. Plug & play generative networks: Conditional iterative generation of images in latent space, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4467–4477.
[21] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, J. Clune, Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, Adv. Neural Inform. Process. Syst. (2016) 3387–3395.
[22] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al, Conditional image generation with pixelcnn decoders, Adv. Neural Inform. Process. Syst. (2016) 4790–4798.
[23] Qiao, T., Zhang, J., Xu, D., Tao, D., 2019. Mirrorgan: Learning text-to-image generation by redescription, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1505–1514.
[24] Ravi, H., Wang, L., Muniz, C.M., Sigal, L., Kapadia, M., Show me a story: Towards coherent neural story illustration, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7613–7621.
[25] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, International Conference on Machine Learning (2016) 1060–1069.
[26] Reed, S., Oord, A.v.d., Kalchbrenner, N., Colmenarejo, S.G., Wang, Z., Belov, D., De Freitas, N., 2017. Parallel multiscale autoregressive density estimation. arXiv preprint arXiv:1703.03664.
[27] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, International Conference on Medical image computing and computer-assisted intervention, Springer (2015) 234–241.
[28] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, Adv. Neural Inform. Process. Syst. (2016) 2234–2242.
[29] Salimans, T., Zhang, H., Radford, A., Metaxas, D., 2018. Improving gans using optimal transport.
[30] Y.Z. Song, Z.R. Tam, H.J. Chen, H.H. Lu, H.H. Shuai, Character-preserving coherent story visualization, European Conference on Computer Vision, Springer (2020) 18–33.
[31] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (2004) 600–612.
[32] 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.
[33] Zeng, G., Li, Z., Zhang, Y., 2019. Pororogan: An improved story visualization model on pororo-sv dataset, in: CSAI2019: 2019 3rd International Conference on Computer Science and Artificial Intelligence.
[34] Zhang, H., Patel, V.M., 2018. Densely connected pyramid dehazing network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3194–3203.
[35] 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.
[36] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D.N. Metaxas, Stackgan++: Realistic image synthesis with stacked generative adversarial networks, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2018) 1947–1962.
[37] Zhang, Z., Xie, Y., Yang, L., 2018b. Photographic text-to-image synthesis with a hierarchically-nested adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6199–6208.
[38] X. Zhu, A.B. Goldberg, M. Eldawy, C.R. Dyer, B. Strock, A text-to-picture synthesis system for augmenting communication, AAAI (2007) 1590–1595.
[39] Dong, P., Wu, L., Meng, L., & Meng, X. (2022). Disentangled Representations and Hierarchical Refinement of Multi-Granularity Features for Text-to-Image Synthesis. Proceedings of International Conference on Multimedia Retrieval (2022).