Character-Preserving Coherent Story Visualization

Yun-Zhu Song[0000−0003−4542−8505], Zhi Rui Tam[0000−0001−9968−2416], Hung-Jen Chen[0000−0003−3129−1595], Huiao-Han Lu[0000−0002−4630−2915], and Hong-Han Shuai[0000−0003−2216−077X]

National Chiao Tung University, Taiwan
{yunzhusong.eed07g,ray.eed08g,hjc.eed07g,hsiaohan.eed05g,hhshuai}@nctu.edu.tw

Abstract. Story visualization aims at generating a sequence of images to narrate each sentence in a multi-sentence story. Different from video generation that focuses on maintaining the continuity of generated images (frames), story visualization emphasizes preserving the global consistency of characters and scenes across different story pictures, which is very challenging since story sentences only provide sparse signals for generating images. Therefore, we propose a new framework named Character-Preserving Coherent Story Visualization (CP-CSV) to tackle the challenges. CP-CSV effectively learns to visualize the story by three critical modules: story and context encoder (story and sentence representation learning), figure-ground segmentation (auxiliary task to provide information for preserving character and story consistency), and figure-ground aware generation (image sequence generation by incorporating figure-ground information). Moreover, we propose a metric named Fréchet Story Distance (FSD) to evaluate the performance of story visualization. Extensive experiments demonstrate that CP-CSV maintains the details of character information and achieves high consistency among different frames, while FSD better measures the performance of story visualization.

Keywords: Story visualization, evaluation metric, foreground segmentation

1 Introduction

"Objects in pictures should so be arranged as by their very position to tell their own story."

— Johann Wolfgang von Goethe *(1749-1832)*

Story Visualization task aims to generate meaningful and coherent sequences of images according to the story text [18], which is challenging since it requires an understanding of both natural language and images. Specifically, *Story Visualization* generates a sequence of images to narrate a given story written in a multi-sentence paragraph. Figure 1 shows an illustrative example of *Story Visualization*. As the saying goes, "A picture is worth a thousand words," and

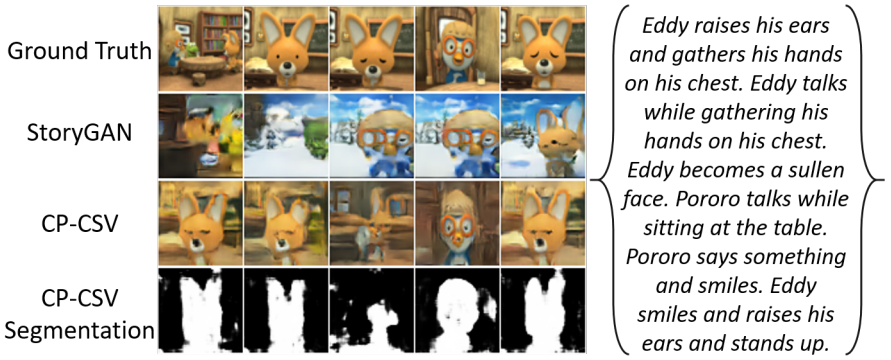

Fig. 1: Story Visualization task prompts to generate image sequences given story descriptions. Our proposed CP-CSV can generate image sequences closely related to the synthesized segmentation result.

a good visualization puts the color inside of the story world and assists the comprehension quickly.

Nevertheless, it remains a challenging task due to the following three challenges. 1) Sequence Coherence. Building transformation between sentences and images requires the ability to tackle cross-domain representation learning, which is highly similar to the text-to-image tasks. However, the major difference between *Story Visualization* and *Text-to-Image* is that *Story Visualization* generates an image sequence based on the whole story comprehension whereas *Text-to-Image* is only based on a single sentence information. In other words, directly applying *Text-to-Image* for story visualization may result in an incoherent image sequences, i.e., images with different contexts. 2) High Variety by Character Dominance. Since the story characters usually occupy a large proportion of pictures (dominance), the pictures change dramatically when different characters appear (variety). The high variety between frames makes the learning of sequence coherence difficult. 3) Implicit Objectives. The goal of the task is to generate high-quality and coherent image sequences that can depict the whole story. However, the subjective and obscure concepts are not standardized into a learning objective function.

Based on the observations, in this paper, we propose a novel framework, namely, *Character-Preserving Coherent Story Visualization (CP-CSV)*, to visualize a story with distinctive characters and a highly-coherent storyline. Specifically, to address the first challenge, two text encoders are used to process the sentence-level and story-level input text. The context encoder focuses on a single sentence and extracts the character information to enhance the character generation, while a story encoder embraces the whole paragraph and utilizes two GRU layers to filter the context information at each time step. Moreover, for the second challenge, we introduce figure-ground information to enable the model to be aware of the foreground and background. Since the foreground images of-

ten represent the point of the story, especially the character appearance, while the background images are usually related to the whole story scenario, CP-CSV generates not only the image sequences but also their segmentation results. In detail, the intermediate features of the segmentation generator assist the image sequence generation layer by layer, and the two generators share the same sentence encoder. Second, following the previous work [18], we adopt an image-level discriminator to assure the relevance between image and reference sentence and a story-level discriminator to maintain the whole paragraph consistency.

Finally, we propose a new evaluation metric, namely Fréchet Story Distance (FSD), to measure the quality of the generated story sequence, which is built on the principle of Fréchet Inception Distance (FID). The FID is a commonly used evaluation metric to quantify the generated image quality by the feature distance between generated images and the reference images. However, the FID takes one image at a time, therefore, it can not capture the temporal series features. On the other hand, Fréchet Video Distance (FVD) adopts the pre-trained Inflated 3D Convnet (I3D, [30]) as a backbone to extract the temporal-aware data distribution. Despite the FVD can evaluate the quality of generated videos, it is not suitable for the Visual Story task because of the limitation of the backbone network. Specifically, the minimum length requirement in I3D is seven, however, most story visualization datasets, e.g. Pororo-SV and VIST, take five sentence-image pairs to form a story. Consequently, we replace the I3D network with a different backbone model R(2+1)D [29] to form a novel story-consistency evaluation metric FSD. The spatial-temporal feature extractions in R(2+1)D are decomposed, therefore, eliminate the input length limitation naturally. We study the FSD behaviors under different consistency perturbations.

For model performance comparison, in addition to objective evaluations, we conduct a user study for comparing the proposed CP-CSV with state-of-the-art methods. Moreover, we illustrate the connection between segmentation results and the generated images. The experimental results manifest that our model can improve the image quality as well as the image sequence coherence.

The contributions are summarized as follows.

- We introduce the segmentation images during training to enhance the model being aware of the figure-ground components and propose a feasible architecture to incorporate the information. The illustration of the synthesized visual story indicates the effectiveness of the segmentation images.
- We build a consistency-evaluation metric FSD for *Story Visualization* task and study the metric behavior under different perturbations.
- Both quantitative and qualitative experimental results manifest that CP-CSV achieves the state-of-the-art for image quality and story consistency.

2 Related Work

2.1 GAN-based Text-to-Image Synthesis

Automatically synthesizing realistic images from text by Generative Adversarial Networks (GANs) [9] has been widely-studied recently. To improve the quality

of text-to-image generation, a variety of models have been proposed and can be categorized into three types: 1) semantic relevance enhancement, 2) resolution enhancement and 3) image diversification. Specifically, semantic relevance enhancement focuses on improving the correlation between ground truth text and generated images. For example, given the base image and text attributes on a desired location, multi-conditional GAN (MC-GAN) [22] is proposed to control both the object and background information jointly for generating a realistic object with the given background. Moreover, since diverse linguistic expressions pose difficulties in extracting consistent semantics, SD-GAN [31] is proposed to implicitly disentangle semantics to attain both high- and low-level semantic knowledge for refined synthesis. Because texts usually contain unimportant words, SEGAN [27] suppresses the word-level attention weights of unimportant words to improve the semantic consistency between the synthesized image and its corresponding ground truth image. MirrorGAN [23] employs a mirror structure, which reversely learns from generated images to output texts for validating whether generated images are consistent with the input texts.

On the other hand, to enhancing the image resolution, different mechanisms are incorporated into GAN. For example, a two-stage stackGAN [34] is proposed to generate low-resolution images and refine the resolution by reading the text description again. To further improve the quality of generated images, StackGAN++ [33] is proposed to use multi-stage GANs to generate multi-scale images. AttnGAN [28] uses attention mechanism to integrate the word-level and sentence-level information into GANs, while DM-GAN [35] uses dynamic memory to refine blurry image contents generated from the GAN network. The third type of methods aims at improving the diversity of generated images and avoiding mode collapse. For example, the discriminator in AC-GAN [21] predicts not only the authenticity of the images but also the label of the images, promoting the diversity of the generated images. Building on the AC-GAN, TAC-GAN [7] synthesizes the image conditioned on corresponding text descriptions instead of on a class label, which helps disentangle the content of images from their styles and makes model generate diverse images based on different content. Text-SeGAN [2] follows the similar idea and revises the discriminator in TAC-GAN by measuring the semantic relevance between the image and text instead of class label prediction to diversify generated results.

To enhance both semantic relevance and resolution, several works take the layout or segmentation as intermediate structures and show the improvement on image quality of text-to-image generation. For example, Hong *et al.* [14] propose to construct a semantic layout based on the text, and generate image consequently. Obj-GAN [17] applies the same two-stage structure as above while using an object-driven attention mechanism for extracting fine-grained information. However, although the above approaches improve the quality of text-to-image generation, story visualization imposes different challenges that are not well-addressed as mentioned in the introduction. Specifically, the challenge of story visualization is to ensure the consistency across the generated sequence. For example, StoryGAN [18] preserves global consistency by using a story-level

discriminator. Several topics are also related to the consistency maintenance, e.g., text-to-video generation [19, 10], dialogue-to-image [4, 25], instruction-to-image generation [8], storyboard creation [3]. CP-CSV is different from previous works since 1) the consistency of video generation emphasizes the continuity between consecutive frames, and 2) CP-CSV further utilizes the figure-ground auxiliary information to preserve the characters and disentangle the background for a better consistency.

2.2 Evaluation Metrics of Image Generation

Evaluation methods of generated images are vital for assessing model performance. Traditional evaluation metrics of image generation, including IS (Inception Score) [24] and FID (Fréchet Inception Distance) [11], focus on scoring the image quality and diversity by comparing the generated images to real images in the distribution. Due to the rise of the text-to-image generation, the semantic relationship between text and generated images should be considered. Therefore, R-precision [28] is proposed to measure the correlation between a generated image and its corresponding text. SOA (Semantic Object Accuracy) [13] measures semantic connection by checking if a generated image contains objects that are specifically mentioned in the image caption. FVD (Fréchet Video Distance) [30] extends FID to calculate the distance between videos. However, it is limited to a long image sequence that contains over seven images. Our proposed FSD (Fréchet Story Distance) eliminates the length limitation and thus is better for evaluating the quality of short story.

3 Character-Preserving Coherent Story Visualization

3.1 Overview

Story Visualization aims at generating a sequence of images from an input story $S = [s_1, \cdots, s_t, \cdots, s_T]$, where s_t denotes the t-th sentence in S and T is the number of sentences in the story. The ground truth image sequence is denoted as $X = [x_1, \cdots, x_t, \cdots, x_T]$, while the generated image sequence is denoted as $\hat{X} = [\hat{x}_1, \cdots, \hat{x}_t, \cdots, \hat{x}_T]$. To address the challenges mentioned in the introduction, we propose Character-Preserving Coherent Story Visualization (CP-CSV), of which the model architecture is shown in Figure 2. In our model, the input story S is first encoded into a vector h_s with a story encoder proposed by [18]. Afterward, to make the generated images consistent with each other, the context encoder takes h_s as the initial state and sequentially encodes each sentence in S into the sentence representations $O = [o_1, \cdots, o_t, \cdots, o_T]$. Different from the video generation task, of which two consecutive frames are similar, two consecutive pictures may change significantly in *Story Visualization*. Therefore, CP-CSV is designed to construct a sequence of distinct but coherent images. Since similar background pictures are usually shared in the same story, whereas foregrounds might change dramatically due to the different character appearances from frames to frames as shown in Figure 1. Therefore, in the training

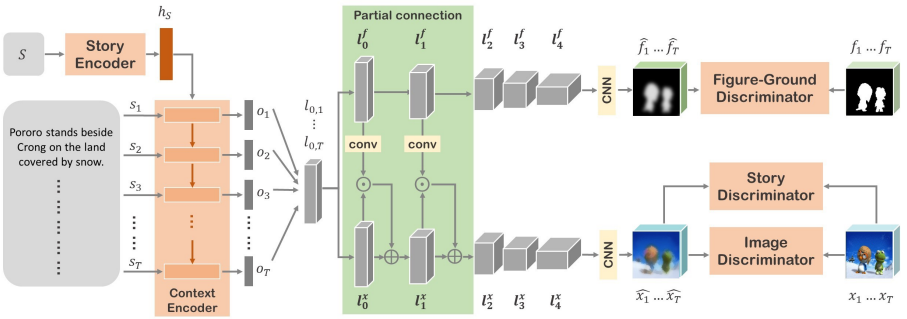

Fig. 2: System framework of CP-SCV. Story/context encoder extracts story/sentence level information. Figure-ground/story/image discriminator learns to distinguish whether the figure-ground/image sequence-story pair/image-sentence pair is true. Our proposed partial connection network is applied to the first and the second level features.

stage, we introduce an auxiliary task, i.e., figure-ground segmentation, to assist CP-CSV for recognizing the figure-ground positions. When the model is capable of locating the foreground and background-position, it is easier to preserve the character formation while maintaining the scene coherence of the backgrounds.

Meanwhile, to automatically evaluate the performance of story visualization, the evaluation metric should take both the image quality and the sequence coherency into consideration. One possible way is to exploit the metrics for video generation. For example, the Fréchet Video Distance (FVD) [30] tackles the video quality evaluation problem by calculating the 2-Wasserstein distance between the synthesized video data distribution and the reference video data distribution. However, the feature extraction model of FVD, I3D [1], requires a long image sequence to calculate the distance. Therefore, FVD may not be suitable for evaluating the quality of story visualization. The sequence length of a story is usually short and does not reach the minimum frame length requirement in I3D. Therefore, we build a novel evaluation metric called Fréchet Story Distance (FSD), which can extract the features of image sequences with arbitrary length. Moreover, the proposed FSD is consistent, even with the noise intensity.

3.2 Story and Context Encoder

For the story encoder and the context encoder, we follow the design of StoryGAN [18]. Specifically, the story encoder aims at learning a mapping function for encoding the whole story S to an embedding h_s representing the whole story. The embedding h_s is assumed to be a normal distribution, and the corresponding mean and variance are defined by two neural networks with story content as input, $\mu(S)$ and $\sigma(S)$. After that, the story embedding h_s serves as the initial state of the Context Encoder and is gradually updated with sentence input at each time step. The Context Encoder is based on a deep RNN to capture

contextual information. To ensure the sentence embedding consistent with the story line, the story embedding should be considered when generating each sentence embedding. Rather than traditional RNNs, StoryGAN [18] introduces the Text2Gist which combines all the global and local context information. In contrast to standard GRU cells, the output is the convolution between the local hidden vectors and a filter with global information. As such, the convolution operation strikes to filter the crucial part from the whole story at each time step and forms a story-aware sentence embedding o_t. For the whole paragraph, the embedding is denoted as $O = [o_1, \cdots, o_t \cdots, o_T]$.

3.3 Figure-Ground Aware Generation

In story visualization, foreground items usually narrate the characters' actions and foreground visual scenes, while the clarity and spatial location of foreground items play an important role on the visualization quality. Equipped with figure-ground information, it is easier to preserve the characters and maintain the background consistency. Therefore, in order to learn such information, we introduce a foreground segmentation generation module to improve the quality of the generated images. Since the ground truth of the foreground segmentation is unavailable in story visualization datasets, we exploit a pre-trained salient object detection model [26] and manually label the foreground segmentation on 1,600 images from the Pororo-SV dataset for finetuning the model.[1]

To simultaneously incorporate the foreground segmentation into the image generation process and exploit the foreground segmentation as the auxiliary task, we design two generators: an image generator visualizing the story and a foreground map generator predicting the foreground regions. By sharing the representations between these two related tasks, we can enable CP-CSV to generalize the original generator better. Specifically, base on the previous StoryGAN generator, we add another generator parallel to the existing image generator to synthesize foreground segmentation maps conditioned on the same foreground area sentences latent vectors O as the image generator. The purpose of this segment generator is to generate the low-level foreground latent feature, which can enhance the quality of generated images. In other words, the image generator could follow the location information of the foreground area (e.g., characters) and synthesize the corresponding characters with much better quality, especially at the boundary of characters.

One possible approach to incorporate foreground features into image features is to exploit the Gated network, which has been proven to aggregate different information effectively. However, the Gated network is usually applied to the deep layers to fuse the high-level information, which may not be suitable for incorporating the figure-ground information. Therefore, we propose to exploit the Partial Connection Network for integrating the features of foreground segmentation, which can be viewed as an affine transformation based on segment

[1] The details of the foreground segmentation model will be presented in the implementation details in Section 4.1. The results are shown in Figure 3. The segmentation images we used are released at https://github.com/yunzhusong/ECCV2020_CPCSV.

features. Specifically, the segment features from the k-th layer denoted as l_k^f are first projected to the latent space of image features l_k^x through a convolution layer F_f, then multiplying the image features to highlight the scene.

$$p_k^f = \mathcal{F}_f(l_k^f) \tag{1}$$

$$l_k^x = p_k^f * l_k^x + l_k^x \tag{2}$$

For learning segment features, a figure-ground discriminator D_{fg} is needed to ensure the learning of foreground segment generation. Similar to the image discriminator, the figure-ground discriminator learns whether the generated segmentation $\hat{f}_t$ matches the given sentence and story latent feature by discriminating between ground truth segmentation f_t.

3.4 Loss function

Let W_g, W_i, W_s and W_f be the parameters representing generator, image, story and figure ground discriminator, respectively. The final objective function for CP-CSV is similar to loss function proposed in GANs:

$$\min_{W_g} \max_{W_i, W_s, W_f} \lambda_1 \mathcal{L}_{image} + \lambda_2 \mathcal{L}_{fg} + \lambda_3 \mathcal{L}_{story}, \tag{3}$$

where λ_1, λ_2, λ_3 are weighting terms to balance the model learning. The conditional loss function for foreground learning $\mathcal{L}_{fg}$ is defined as

$$\begin{aligned} \mathcal{L}_{fg,D} &= \textstyle\sum_{t=1}^{T} -\mathbb{E}[\log(D_{fg}(f_t, s_t, h_s))] - \mathbb{E}_{\hat{f}_t \sim p_g}[\log(1 - D_{fg}(\hat{f}_t, s_t, h_s)], \\ \mathcal{L}_{fg,G} &= \mathbb{E}_{z \sim p_z}[\log(1 - D_{fg}(G_{fg}(s_t, z, h_s), s_t, h_s)] + \lambda_4 \mathcal{L}_{KL}, \end{aligned} \tag{4}$$

where $\mathcal{L}_{KL}$ is a regularization term for smoothing the semantic latent space and increasing the input variety for the sake of relieving the issue of mode collapse. Specifically, the KL regularization term is obtained as follows:

$$\mathcal{L}_{KL} = KL(N(\mu(S), diag(\sigma^2(S)))||N(0, \mathbf{I})), \tag{5}$$

where $diag(\cdot)$ is used to restrict $\sigma^2(S)$ as a diagonal matrix for computational tractability, $\mu(.)$ and $\sigma(.)$ are two neural networks and take story S as input to predict a mean and variance for a normal distribution respectively.

The loss functions for image L_{image} and story L_{story} remain unchanged from the original StoryGAN.

$$\begin{aligned} &\mathcal{L}_{image,D} = \textstyle\sum_{t=1}^{T} -\mathbb{E}[log(D_{image}(x_t, s_t, h_s))] - \mathbb{E}_{\hat{x}_t \sim p_g}[log(1 - D_{image}(\hat{x}_t, s_t, h_s))], \\ &\mathcal{L}_{image,G} = \mathbb{E}_{z \sim p_z} log(1 - D_{image}(G(s_t, z, h_s), s_t, h_s) + \lambda_4 \mathcal{L}_{KL} \end{aligned} \tag{6}$$

$$\begin{aligned} \mathcal{L}_{story,D} &= -\mathbb{E}[log(D_{story}(X, S))] - \mathbb{E}_{\hat{X} \sim p_g}[log(1 - D_{story}(\hat{X}, S))], \\ \mathcal{L}_{story,G} &= \mathbb{E}_{z \sim p_z} log(1 - D_{story}([G_{image}(s_t, z, h_s)]_{t=1}^{T}), S) + \lambda_4 \mathcal{L}_{KL} \end{aligned} \tag{7}$$

Our objective function is updated using Adam with a learning rate of 0.0001 and 0.0004 for generators and discriminators. We find out that reducing the learning rate by half at epoch 20, 40, 80 helps stabilize the learning process. The values for λ_1, λ_2, λ_3, λ_4 are 5, 1, 1, 1 respectively.

3.5 Fréchet Story Distance

Previous work usually exploits the metric of image quality evaluation for story visualization. In this case, FID is the commonly-used metric to measure the image quality by calculating the 2-Wasserstein distance between the generated images and the reference images. However, for story visualization, the evaluation metric should take not only the image quality but also the consistency between frames into consideration,not included in the FID. In the field of *Video Generation*, FVD [30] is commonly used to evaluate a sequence of generated images [6, 5], which adopts Inflated 3D ConvNet (I3D) to extract the video latent representation.

However, the inherent limitation of the I3D prevents FVD from operating directly to our task, since the minimum required frame number is seven. Simultaneously, the length of the image sequence in *Story Visualization Task* is usually smaller than the requirement. Indeed, we could modify the task description to achieve the frame length requirement, e.g., by considering more sentences as a story or generating multi-images from a sentence. One obvious shortcoming of expending story length is losing the comparability since the image generation is based on the whole story, differnet sentence length may alter the story line. On the other hand, generating multi-images from a sentence may confuse the model and even weaken the relevance between text and image.

To tackle the third challenge of *Story Visualization Task*, i.e., the lack of standard evaluation metric, we propose Fréchet Story Distance (FSD) as a new evaluation metric for *Story Visualization Task*, which is built on the principle of FID and FVD but with different backbone model, R(2+1)D [29]. The R(2+1)D network factorizes the 3D convolution filters into 2D spatial convolution and 1D temporal convolution, and the details are omitted here. The considerations of adopting R(2+1)D are the flexibility of sequence length and the strong ability to capture temporal consistency.

Given image sequence with arbitrary length, the last average pooling layer's output is taken as the sequence representation. With the representations of generated data P_G and reference data P_R, the distance between the two data representations is defined by 2-Wasserstein distance and calculated by:

$$d(P_R, P_G) = |\mu_R - \mu_G|^2 + Tr(\Sigma_R + \Sigma_G - 2(\Sigma_R \Sigma_G)^{1/2}),$$

where μ_R and μ_G are the means, and Σ_R and Σ_G are the covariance matrices. Finally, we observe the behaviors of FSD under different noise attacks. The detail experimental setting and results are discussed in Sec 4.7.

Fig. 3: Illustrate the images and segmentation images. The upper rows indicate the ground truth (GT), while the lower rows are from our model. Results show a high correlation between generated images and generated segmentations.

4 Experimental Results

4.1 Implementation Details

For CP-CSV implementation, two techniques are applied to stabilize the training of GAN. First, Spectral Normalization [20] is performed to discriminators. Second, Two Time-scale Update Rule (TTUR) is applied for selecting different learning rates for the generator and discriminators [12]. Using these techniques, CP-CSV can produce better images and lower variance scores between different training sessions.

We conduct extensive experiments on the Pororo-SV dataset. Since the Pororo-SV dataset does not include the figure-ground information, we train a model in a semi-supervised manner. Specifically, we utilize a pre-trained state-of-the-art salient object detection model [26] to obtain the segmentation images. The detection model is fine-tuned on the Pororo-SV dataset by 1,600 manually-labeled samples. The second row of Figure 3 demonstrates the examples of the segmentation results generated from the pre-trained model. We also release the segmentation images for the Pororo-SV dataset and Clever dataset.

4.2 Dataset

The following datasets are for training CP-CSV or for analyzing the FSD.

Pororo-SV: The Pororo-SV dataset [18] introduced by StoryGAN is modified from the Pororo dataset [16] to fit the story visualization task. It contains 13,000 training pairs and 2,334 testing pairs. Following the task formulated in StoryGAN, we also consider every five consecutive images as a story. There are several descriptions for one image, and one description is randomly selected during the training and testing phases.

VIST: The VIST dataset [15] originally used for sequential vision-to-language tasks contains 50,136 stories. Each story consists of five images and five matched captions. Different from the visual storytelling task, we take captions as input and generate the corresponding images to form a story. In this paper, the VIST dataset is only applied to analyze FSD behavior.

Fig. 4: Illustrate the synthesized image sequences. GT refers to Ground Truth.

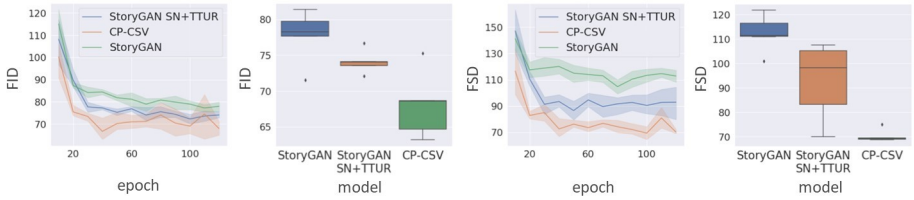

Fig. 5: FID and FSD average results using five different random seeds. CP-CSV drops faster in terms of FID and FSD scores compared to StoryGAN and StoryGAN with Spectral Normalization and TTUR.

4.3 Baselines

To the best of our knowledge, in story visualization task, there is only one state-of-the-art, i.e., StoryGAN [18]. To better know our model performance, we also compare CP-CSV with two other text-to-image models.

SAGAN: Based on the self-attention generative network proposed by Zhang *et al.* [32], we reimplement the SAGAN by taking the encoded sentence in the story S, character labels, and a noise term as input. Each image within the same story is generated independently.

StoryGAN: StoryGAN follows the Li *et al.* [18].

4.4 Qualitative Comparison

To get a more evident concept about the performance, Figure 4 shows the comparison between synthesized image sequences of baseline models and CP-CSV. In contrast to SAGAN, which does not take the whole story as input, StoryGAN and CP-CSV exhibit better ability to maintain scene consistency. Besides, with the figure-ground information, CP-CSV can preserve the character details and

Table 1: Quantitative evaluation results. ↓ denotes the lower is better.

	FID↓	FVD↓	FSD↓
SAGAN [32]	84.70	324.86	101.11
StoryGAN [18]	77.67	274.59	111.09
CP-CSV	**67.76**	**190.59**	**71.51**

Table 2: Human evaluation results of visual image quality, sequence consistency and text-image relevance. "Ours/SAGAN" represents the A/B test of CP-CSV and SAGAN. "Ours/StoryGAN" represents the A/B test of CP-CSV and StoryGAN. The values are shown in percentage.

	Quality		Consistency		Relevance	
	our	baseline	our	baseline	our	baseline
Ours/SAGAN	**0.72**	0.28	**0.82**	0.18	**0.63**	0.37
Ours/StoryGAN	**0.71**	0.29	**0.54**	0.46	**0.53**	0.47

shapes better, since it is easier to locate the position of characters. From the lower rows of Figure 3, they show the two output results of our model. We can find that the connection between segmentation images and generated images is evident and apparent, suggesting that our architecture can effectively utilize the figure-ground information. More examples are shown in Supplementary.

4.5 Quantitative Comparison

Table 1 shows the story evaluation results, measured by FID, FVD, and the proposed FSD. The FID takes one image at a time to perform the evaluation. In contrast, FVD takes the whole story, noting that the input images are self duplicated before feeding to FVD due to the limitation describing in Sec 3.5. Besides, we conduct a human survey for the synthesized visual story, and there are three modalities. Given three sequences of images and the corresponding paragraphs, users rank the sequences according to 1) the visual image quality, 2) consistency between images, and 3) the relevance between image and sentence. The pairwise comparison results in Table 2 are extracted from the ranking results. Our model is ranked higher than all baselines on three modalities. The performance on image quality is especially disparity, demonstrating the effectiveness of the proposed figure-ground aware generation.

4.6 Architecture Search

To better understand the effectiveness of the proposed model and its variants, we conducted several comparative experiments by calculating FID, FVD, and FSD values. The generation flow is firstly discussed, i.e., the **cascade** generation and the **parallel** generation. In our experimental settings, the **cascade**

Table 3: The evaluation results of different architectures for combining the figure-ground information.

	FID↓	FVD↓	FSD↓
baseline	74.60	189.46	98.61
baseline + SEG (Cascade)	73.46	182.52	86.80
baseline + SEG (Parallel, k=1,2,3,4)	84.41	194.9	81.46
baseline + SEG (Parallel, k=3,4)	80.54	179.42	99.66
Ours (baseline + SEG (Parallel, k=1,2))	**69.55**	**177.27**	**72.60**

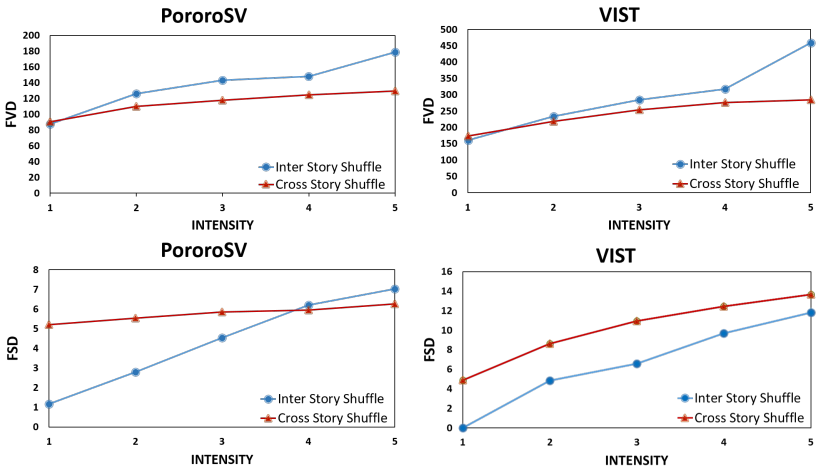

Fig. 6: FVD and FSD analysis of two kinds of perturbations.

generates the segmentation before the image. Once obtaining the segmentation, it takes the down-sample to extract the figure-ground features. The features are then combine into the image generator to form the final result. On the other hand, the **parallel** generates the segmentation and image simultaneously, and takes the latent features to combine into the image generator. The variants of **parallel** are introducing the figure-ground information at different layers. Our experimental results shown in Table 3 suggest that the figure-ground information should be incorporated at early stage, the possible reason is that the information of segmentation is close to high-level concept, i.e., the character position, however, the last few layers tackle more detail formation. As for the inferior of **cascade** may from the process of down-sampling the generated segmentation, which highly relies on the segmentation quality.

4.7 FSD Analysis

The FSD is proposed to evaluate the sequence coherency of the visual story. To identify whether FSD is sensitive to the sequence consistency, we apply two

Table 4: Pearson correlation coefficient of metric measurements and perturbation intensity.

	Pororo-SV		VIST	
	Inter-S	Cross-S	Inter-S	Cross-S
FVD	0.9671	0.9583	0.9697	0.9685
FSD	0.9984	0.9957	0.9876	0.9724

different types of perturbations to the real image sequences: 1) Inter-Story Shuffle: Swapping the sequence order within a story, and the intensity is increasing with the needed steps to reorder the sequence. 2) Cross-Story Shuffle: Exchange the images with the images from other stories, and the intensity is increasing with the number of outside images. We test these noises under two datasets, i.e., Pororo-SV and VIST. To compare with the behavior of FVD, we also analyze the deviation of the FVD under these settings. Note that to maintain consistency with the FVD evaluation for the models, the input sequence length would be duplicated to reach the minimum length requirement of the I3D network instead of considering more sentences as a story. The line charts in Figure 6 show how the metrics react to the inter-story shuffle and the cross-story shuffle perturbations. The generated images with different FSD scores are shown in Supplementary.

5 Conclusions

In this paper, we introduce the figure-ground segmentation images to address the Story Visualization task, based on the observation of different changing rate between foreground and background, and propose a novel framework called Character-Preserving Coherent Story Visualization (CP-CSV) to incorporate the segmentation information layer by layer. Qualitative and quantitative experiments suggest CP-CSV outperforms the state-of-the-art story visualization model. Moreover, to give an automatic evaluation metric of Story Visualization for consistency, Fréchet Story Distance (FSD) is built on the principle of FID and FVD. The perturbation studies show that FSD is highly sensitive to the story consistency. We provide more examples in the supplementary.

Acknowledgements

We are grateful to the National Center for High-performance Computing for computer time and facilities. This work was supported in part by the Ministry of Science and Technology of Taiwan under Grants MOST-108-2221-E-009-088, MOST-109-2221-E-009-114-MY3, MOST-109-2634-F-009-018, MOST-109-2218-E-009-016 and MOST-108-2218-E-009-056.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
2. Cha, M., Gwon, Y.L., Kung, H.: Adversarial learning of semantic relevance in text to image synthesis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 3272–3279 (2019)
3. Chen, S., Liu, B., Fu, J., Song, R., Jin, Q., Lin, P., Qi, X., Wang, C., Zhou, J.: Neural storyboard artist: Visualizing stories with coherent image sequences. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 2236–2244 (2019)
4. Cheng, Y., Gan, Z., Li, Y., Liu, J., Gao, J.: Sequential attention gan for interactive image editing via dialogue. arXiv preprint arXiv:1812.08352 (2018)
5. Clark, A., Donahue, J., Simonyan, K.: Adversarial video generation on complex datasets. (2019)
6. Dandi, Y., Das, A., Singhal, S., Namboodiri, V., Rai, P.: Jointly trained image and video generation using residual vectors. In: The IEEE Winter Conference on Applications of Computer Vision. pp. 3028–3042 (2020)
7. Dash, A., Gamboa, J.C.B., Ahmed, S., Liwicki, M., Afzal, M.Z.: Tac-gan-text conditioned auxiliary classifier generative adversarial network. arXiv preprint arXiv:1703.06412 (2017)
8. El-Nouby, A., Sharma, S., Schulz, H., Hjelm, D., Asri, L.E., Kahou, S.E., Bengio, Y., Taylor, G.W.: Tell, draw, and repeat: Generating and modifying images based on continual linguistic instruction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10304–10312 (2019)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
10. He, J., Lehrmann, A., Marino, J., Mori, G., Sigal, L.: Probabilistic video generation using holistic attribute control. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 452–467 (2018)
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems. pp. 6626–6637 (2017)
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NIPS (2017)
13. Hinz, T., Heinrich, S., Wermter, S.: Semantic object accuracy for generative text-to-image synthesis. arXiv preprint arXiv:1910.13321 (2019)
14. Hong, S., Yang, D., Choi, J., Lee, H.: Inferring semantic layout for hierarchical text-to-image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7986–7994 (2018)
15. Huang, T.H., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R.B., He, X., Kohli, P., Batra, D., Zitnick, C.L., Parikh, D., Vanderwende, L., Galley, M., Mitchell, M.: Visual storytelling. In: HLT-NAACL (2016)
16. Kim, K.M., Heo, M.O., Choi, S.H., Zhang, B.T.: Deepstory: Video story qa by deep embedded memory networks. arXiv preprint arXiv:1707.00836 (2017)
17. Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S., Gao, J.: Object-driven text-to-image synthesis via adversarial training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12174–12182 (2019)

18. Li, Y., Gan, Z., Shen, Y., Liu, J., Cheng, Y., Wu, Y., Carin, L., Carlson, D., Gao, J.: Storygan: A sequential conditional gan for story visualization. CVPR (2019)
19. Li, Y., Min, M.R., Shen, D., Carlson, D., Carin, L.: Video generation from text. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
20. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. ArXiv **abs/1802.05957** (2018)
21. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 2642–2651. JMLR. org (2017)
22. Park, H., Yoo, Y., Kwak, N.: Mc-gan: Multi-conditional generative adversarial network for image synthesis. In: The British MachineVision Conference (BMVC) (2018)
23. Qiao, T., Zhang, J., Xu, D., Tao, D.: Mirrorgan: Learning text-to-image generation by redescription. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
24. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Advances in neural information processing systems. pp. 2234–2242 (2016)
25. Sharma, S., Suhubdy, D., Michalski, V., Kahou, S.E., Bengio, Y.: Chatpainter: Improving text to image generation using dialogue. arXiv preprint arXiv:1802.08216 (2018)
26. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR (2019)
27. Tan, H., Liu, X., Li, X., Zhang, Y., Yin, B.: Semantics-enhanced adversarial nets for text-to-image synthesis. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10501–10510 (2019)
28. Tao Xu, Pengchuan Zhang, Q.H.H.Z.Z.G.X.H.X.H.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks (2018)
29. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
30. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
31. Yin, G., Liu, B., Sheng, L., Yu, N., Wang, X., Shao, J.: Semantics disentangling for text-to-image generation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2322–2331 (2019)
32. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318 (2018)
33. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.: Stackgan++: Realistic image synthesis with stacked generative adversarial networks. arXiv: 1710.10916 (2017)
34. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV (2017)
35. Zhu, M., Pan, P., Chen, W., Yang, Y.: Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5802–5810 (2019)