

Clustering Generative Adversarial Networks for Story Visualization

Bowen Li
bowen.li@cs.ox.ac.uk
University of Oxford, UK

Philip H. S. Torr
philip.torr@eng.ox.ac.uk
University of Oxford, UK

Thomas Lukasiewicz
thomas.lukasiewicz@cs.ox.ac.uk
TU Wien, Austria; University of Oxford, UK

ABSTRACT

Story visualization aims to generate a series of images, semantically matching a given sequence of sentences, one for each, and different output images within a story should be consistent with each other. Current methods generate story images by using a heavy architecture with two generative adversarial networks (GANs), one for image quality, and one for story consistency, and also rely on additional segmentation masks or auxiliary captioning networks. In this paper, we aim to build a concise and single-GAN-based network, neither depending on additional semantic information nor captioning networks. To achieve this, we propose a contrastive-learning-and-clustering-learning-based approach for story visualization. Our network utilizes contrastive losses between language and visual information to maximize the mutual information between them, and further extends it with clustering learning in the training process to capture semantic similarity across modalities. So, the discriminator in our approach provides comprehensive feedback to the generator, regarding both image quality and story consistency at the same time, allowing to have a single-GAN-based network to produce high-quality synthetic results. Extensive experiments on two datasets demonstrate that our single-GAN-based network has a smaller number of total parameters in the network, but achieves a major step up from previous methods, which improves FID from 78.64 to 39.17, and FSD from 94.53 to 41.18 on Pororo-SV, and establishes a strong benchmark FID of 76.51 and FSD of 19.74 on Abstract Scenes.

CCS CONCEPTS

• Computing methodologies → Computer vision.

KEYWORDS

story visualization, GANs, clustering learning, contrastive learning

ACM Reference Format:

Bowen Li, Philip H. S. Torr, and Thomas Lukasiewicz. 2022. Clustering Generative Adversarial Networks for Story Visualization. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*, October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3503161.3548034>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '22, October 10–14, 2022, Lisboa, Portugal

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9203-7/22/10...\$15.00

<https://doi.org/10.1145/3503161.3548034>

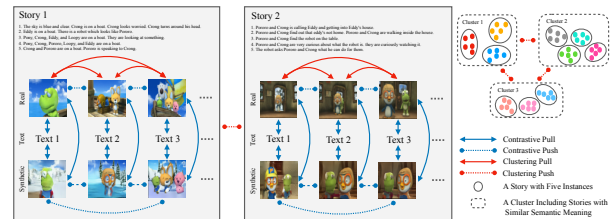


Figure 1: We extend contrastive learning with clustering learning to capture semantic similarity across modalities.

1 INTRODUCTION

There has been a great progress in text-to-image generation [16, 23, 27, 38, 48, 52, 53], due to realistic image generation using generative adversarial networks (GANs) [11]. The task of story visualization [29, 32, 42] is a more challenging variation of it, as we need to generate a sequence of story images given a multi-sentence story, and further require output images to be consistent, e.g., having a similar background or objects' appearances.

As story visualization requires synthetic story images to be realistic and also consistent, current methods [29, 31, 32, 42] implement two GANs, one for image quality, and one for story consistency. Also, to ensure the quality of results, the work [42] required segmentation masks, and [31, 32] adopted additional auxiliary captioning networks. Two GAN-based networks with additional auxiliary networks increase the requirement of computational resources, e.g., GPU memory size, which may hinder further studies of the story visualization task.

In this paper, we aim to build a concise and single-GAN-based network, neither depending on additional semantic information nor auxiliary captioning networks. Our network should have a smaller number of parameters compared to current methods, but still achieve a competitive performance. To achieve this, we seek help from contrastive learning. Contrastive learning is a powerful scheme for self-supervised representation learning [8, 14, 46, 50]. It enforces the consistency of image representations under different augmentations by contrasting positive pairs with negative ones.

However, contrastive learning does not take the samples' semantic information and similarity into account, where it simply treats two samples as a positive pair as long as they are at the same batch, and negative pair when they are at different batches, without considering their semantic information. By doing this, the learned representation can be considerably affected. For example, instances within the same story have a similar semantic information, while simply adopting contrastive learning can push these instances far away from each other, and thus break the semantic consistency between them. To consider similarity between samples, clustering algorithms [1, 2, 6, 28] can remedy the above problems, where

clustering algorithms group similar instances (i.e., instances with similar semantic meaning) into the same cluster, and push different instances away into different clusters.

One problem arising from clustering learning is that story visualization involves different-modal features (i.e., language and visual features), if we simply cluster these features into different clusters, these different-modal features may be only grouped into their own modalities separately. That is, visual instances are only clustered with visual instances, and language instances with language instances. So, it is important to have an approach that can pull different-modal instances together in a joint space. To address this, we suggest to feed fusion features into clustering algorithms, which are created from fine-grained image regional features and word embeddings. Moreover, we suggest to feed fusion features into the discriminator as well, where the discriminator becomes to compare the quality of fusion features created from synthetic images and given sentences to fusion features created from ground-truth images and the sentences. We think that feeding fusion features into the discriminator can help regularize training and stabilize clustering learning with fusion features as inputs. Also, as our discriminator only critic the quality of fusion features, we do not need to separate the discriminator network into two pathways to evaluate image quality and text-image alignment separately, which satisfies our goals to build a concise network.

As the discriminator adopts word-level information to provide fine-grained training feedback, we propose to utilize it in the generator to improve its capabilities of capturing relations between language and visual features. So, we adopt word-level spatial attention [48] (WSA), and further extend it to comprehensively capture both positive and negative relations between image regions and different words, called bi-directional word-level spatial attention (BWSA), where positive effects of BWSA work similarly as WSA to highlight image regions corresponding to semantic words, and negative effects work as a complementary component to capture relations that cannot be observed by the positive part.

In summary, by adopting both contrastive learning and clustering learning, the discriminator in our approach can provide a comprehensive training feedback to the generator, in terms of both image quality and story consistency at the same time, allowing us to have a single-GAN-based network to produce high-quality synthetic results, instead of having two separate networks neither relying on additional semantic information nor auxiliary networks. Overall, our contributions are summarized as follows:

1. A concise single-GAN-based network is proposed for story visualization task, neither using additional semantic information nor captioning networks, which has a smaller number of trainable parameters, but improves FID from 78.64 to 39.17, and FSD from 94.53 to 41.18 on Pororo-SV, and establishes a strong benchmark FID of 76.51 and FSD of 19.74 on Abstract Scenes.
2. We extend contrastive learning with clustering learning in the training process to capture semantic similarity across modalities. We explore online and offline learning approaches for implementing clustering algorithms into the multimodal generation process.
3. We conduct a thorough analysis of combining the benefits of contrastive learning and clustering learning to provide general modeling insights in conditional GANs. The code is available at <https://github.com/mrlbw/Clustering-Story-Visualization>.

2 RELATED WORK

Story visualization aims to generate a series of images from a given multi-sentence story, and different output images within a story should be consistent with each other. StoryGAN [29] first introduced this task and proposed a sequential conditional generative adversarial network. Built on top of StoryGAN, CP-CSV [42] added another segmentation mask pipeline to keep character consistency, and DUCO [32] and VLC [31] utilized video captioning networks to improve text-image alignment. As all these methods have a similar architecture, which contains two sets of generative adversarial networks, one for image quality, and one for story consistency, computational requirements to hold and train their networks are large. However, our proposed network only has a generative adversarial network with a small number of parameters, but can still achieve the same competitive performance as current methods.

Video generation from text is related to our work, which generates a series of video frames matching a given sentence [4, 13, 30, 33, 36]. Differently, output results of story visualization allow more diversity between them, instead of strictly creating a continuous flow of frames. Story visualization enables a more interesting interaction between images and sentences, while this interaction can only be achieved in longer videos in the video generation task.

Text-guided image generation and manipulation is related to our work as well, which has achieved a great progress because of the success of deep generative models [11, 12, 21, 45] in realistic image generation and manipulation [10, 24–26, 34]. GAN-INTCLS [40] first introduced a GAN-based method to tackle text-to-image generation task. StackGAN [51] introduced a multi-stage architecture. Then, the attention-based methods AttnGAN [48] and ControlGAN [23] were proposed to further improve the results. DMGAN [53] introduced a dynamic memory module to refine image contents. XMC-GAN [50] introduced a contrastive-learning-based method to generate images from text using a single stage. DF-GAN [43] proposed a deep fusion module to better fuse text and image features. DALL-E [39] introduced an autoregressive-transformer-based method for zero-shot text-to-image generation. GLIDE [35] proposed a diffusion-model-based method to achieve image generation and editing. The work [27] proposed a semi-parametric method by constructing a memory bank, providing image features to the generation pipeline.

Contrastive learning in GANs and clustering learning. Contrastive learning is a powerful scheme for self-supervised representation learning [8, 9, 14, 22, 46, 50]. It enforces consistency between samples from the same instance under different augmentations (e.g., transforms or crops of an image), while pushing samples from different instances further apart. XMC-GAN [50] used intra-modality and inter-modality contrastive learning in text-to-image synthesis. ContraGAN [18] adopted contrastive learning in class-conditional image generation. DiscoFaceGAN [9] incorporated contrastive learning in face generation to enable a better disentanglement. CUT [37] used positive pairs from the same image patches in both input and output images for image-to-image translation. Differently, we further extend contrastive learning with clustering learning in conditional GANs to consider samples' semantic information and similarity between different samples. Clustering learning can discriminate between groups of similar instances during

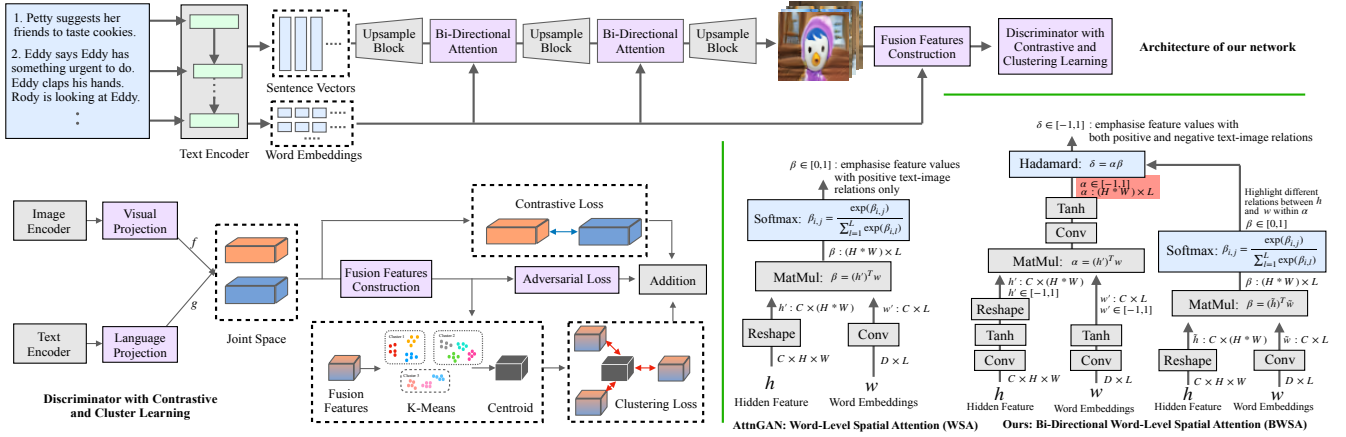


Figure 2: Illustration of our proposed framework.

training. The works [5, 49] kept the clustering step independent of the representation learning phrase, while the works [2, 3, 47] jointly learned visual representations and cluster assignments. As for implementing clustering in multi-modal tasks, XDC [1], SeLaVi [2], and MCN [7] focused on pretrained feature extractors for audio and video, and aimed to produce effective multi-modal representations used for different downstream tasks. Unlike prior works, we extend contrastive learning with clustering algorithms, and use both in the generation pipeline of a conditional GAN, aiming to capture semantic similarity across multi-modal features.

3 CONTRASTIVE AND CLUSTERING STORY GENERATIVE ADVERSARIAL NETWORKS

In this paper, we aim to build a single-GAN-based architecture for story visualization, which has a smaller number of parameters, compared to current methods, but can still ensure image quality and story consistency at the same time, instead of treating them separately. To achieve this, we propose a new contrastive-learning-and-clustering-learning-based network, where contrastive learning maximizes the mutual information between language and visual information to ensure a well-aligned text-image correspondence, and clustering learning captures semantic similarity across instances.

3.1 Model Overview

Differently from current methods [29, 31, 32, 42], which have two generators and two discriminators, with the help of segmentation masks [42] or auxiliary captioning networks [31, 32], the network of our approach is made up with a generator and a discriminator. The generator takes text embeddings encoded by a text encoder from a given multi-sentence story, and a noise vector sampled from a Gaussian distribution as inputs. First, the noise vector and text embeddings are fed into a fully connected network with a reshape operation to produce initial story image features at 4×4 . Then, we apply a series of upsampling blocks to upsample image features into the desired resolution, generating all story images at the same time. The discriminator consists of a series of downsampling blocks, and evaluates the quality of given fusion features, providing training feedback to the generator. Meanwhile, our discriminator also implements contrastive learning to maximize mutual information

between language and visual information, and adopts clustering learning to capture semantic similarity between instances.

3.2 Contrastive Losses for Story Visualization

To ensure a good image quality and story consistency at the same time, we implement contrastive learning to maximize the mutual information between corresponding pairs, including (1) story sentence and story image, (2) single sentence and single image, (3) words and image regions, and (4) words and story image regions. Here, sentence and image are global representations of words and image regions, respectively, and story sentence and story image are global representations of sentences and images from the same story, respectively. Differently from XMC-GAN, it adopts the normalized temperature-scaled cross entropy loss (NT-Xent) [8]:

$$\mathcal{L}_{\text{cont}}(v_i, s_i) = -\log \frac{\exp(\cos(v_i, s_i)/\tau)}{\sum_{j=1}^M \exp(\cos(v_i, s_j)/\tau)}, \quad (1)$$

where v and s are two representations from a sample, M is the total number of samples with $M - 1$ negative samples, $\cos(\cdot)$ denotes cosine similarity, and τ denotes a temperature hyperparameter.

In our work, we propose to use a masked margin softmax function (MMS) [7, 17] to define the similarity between representations from visual and language information, with respect to their learned embeddings' dot product within a batch N . This is because NT-Xent may miss opportunities to learn against a wider set of negative examples, namely, all those in the batch that are not known to be positively associated, while MMS can exploit these additional negatives. For simplicity, in the following, V represents the visual representation (i.e., story image, image, or image regions), v represents corresponding visual features, extracted by an image encoder, S represents the language representation (i.e., story sentence, sentence, or words), and s represents corresponding text features, encoded by a text encoder. In this task, the image encoder takes an image as input to extract image regional features and a global image feature, and then we average a series of global image features within the same story as a story image feature. The text encoder takes multiple sentences from a story as input to encode a global story sentence vector, sentence vectors, and word embeddings.

Given a set of pairs of associated story image and story sentence, we first construct parametrized mappings that derive embedding

representations from extracted visual and language features. We use the transform $f : v \rightarrow \mathbb{R}^D$ to convert given visual features into a joint space \mathbb{R}^D , and the transform $g : s \rightarrow \mathbb{R}^D$ to map given language features into the same space. Thus, for a pair of information (e.g., single sentence and single image), the loss $\mathcal{L}_{\text{cont}}$ is:

$$-\frac{1}{N} \sum_{i=1}^N \left[\log \frac{\exp(f(v_i) \cdot g(s_i) - \delta)}{\exp(f(v_i) \cdot g(s_i) - \delta) + \sum_{j=1, j \neq i}^N \exp(f(v_j) \cdot g(s_i))} \right] + \left[\log \frac{\exp(f(v_i) \cdot g(s_i) - \delta)}{\exp(f(v_i) \cdot g(s_i) - \delta) + \sum_{k=1, k \neq i}^N \exp(f(v_i) \cdot g(s_k))} \right], \quad (2)$$

where N is the batch size, δ is a hyperparameter. By projecting all features to the same space and ensuring that their similarities are maximized pairwise, we ensure that the mutual information is maximized between corresponding pairs, and thus output results can be both realistic and consistent.

3.3 Clustering Losses for Story Visualization

As contrastive learning treats two samples as a negative pair as long as they occur at different batches, regardless of their semantic similarity, it can (1) push instances from the same story far away from each other, and then may fail to keep the consistency between these instances, and (2) also push instances from different stories away without considering their semantic similarity. By doing this, the learned representation can be considerably affected. So, to ensure instances with similar semantic meaning being close in the learned joint multimodal space, we incorporate clustering learning in the training process. In the following, we will explore both online and offline clustering learning in story visualization.

3.3.1 Fusion Features. First, we define the inputs for the clustering algorithm. If we simply cluster multi-modalities features into different clusters, these features may be only grouped into their own modalities separately. To address this, we use fusion features $X \in \mathbb{R}^{(H*W) \times L}$ as the input for clustering learning:

$$X = f(v_{\text{region}}) \cdot g(s_{\text{word}}^T), \quad (3)$$

where T denotes the transpose, $f(v)_{\text{region}}$ represents image regional features, $g(s)_{\text{word}}$ represents word embeddings, $H * W$ denotes the total number of spatial locations in image regional features, and L denotes the total number of words in a sentence. Here, we take the matrix multiplication over visual and language fine-grained features to represent a multimodal instance, which preserves visual and language information to ensure a more precise clustering.

3.3.2 Online K-Means Clustering. We adopt the standard clustering algorithm K-means in the training process. To implement it, we cluster fusion features into k distinct groups. So, the objectives of the clustering algorithm are to partition T fusion features X_1, X_2, \dots, X_T into k ($< T$) sets $S = \{S_1, S_2, \dots, S_k\}$, and to minimize the within-cluster distances:

$$\arg \min_S \sum_{i=1}^k \sum_{X \in S_i} \|X - C_i\|^2, \quad (4)$$

where C_i is the mean of fusion feature points in S_i . Here, to have enough instances and to cover variant semantic information for

clustering, following [7], we use fusion features from the previous batches to gather sufficient instances for online clustering learning.

So, we get k centroids C_1, C_2, \dots, C_k , and then we use these centroids as contrastive loss reference targets. This target pulls the fusion features belonging to the same cluster closer to the centroid, and pushes the features far away from the other centroids:

$$\mathcal{L}_{\text{cluster}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\cos(X_i, C_i)/\tau)}{\sum_{j=1}^k \exp(\cos(X_i, C_k)/\tau)}. \quad (5)$$

Finally, the network learns to project both visual and language features to be closer to their centroids, and also learns to capture semantic similarity across modalities, which, together with contrastive learning, comprehensively maximizes the mutual information between instances.

3.3.3 Offline K-Means Clustering. In online K-means clustering, centroids are not fixed and are adjusted along with the training of the GAN, which may cost time to find centroids. Instead, we can precompute the centroids and then fix these centroids during training. To achieve this, we first pretrain an image encoder (e.g., VGG-16 [41]) and a text encoder (e.g., LSTM) to directly map both visual and language inputs into a joint semantic space \mathbb{R}^D . The objectives to train both image and text encoders are defined below:

$$\begin{aligned} \mathcal{L}_{\text{pretrain}} = & \mathcal{L}_{\text{cont}}(v_{\text{s_image}}^r, s_{\text{s_sent}}) + \mathcal{L}_{\text{cont}}(v_{\text{image}}^r, s_{\text{sent}}) \\ & + \mathcal{L}_{\text{cont}}(v_{\text{regions}}^r, s_{\text{words}}) + \mathcal{L}_{\text{cluster}}(X^r) \\ & + \mathcal{L}_{\text{cont}}(v_{\text{s_regions}}^r, s_{\text{words}}), \end{aligned} \quad (6)$$

where r represents features extracted from a real sample, $\text{cont}(v, s)$ denotes applying contrastive losses between visual representation v and language representation s , $\mathcal{L}_{\text{cluster}}(X)$ represents applying clustering learning on fusion features X , which are created by paired image and sentence features, $v_{\text{s_image}}$ denotes visual features from a story image, $s_{\text{s_sent}}$ denotes text features from a story sentence, and $v_{\text{s_regions}}$ denotes regional features for a story image. To pretrain both text and image encoders, we use paired story images and sentences from the given dataset. Therefore, after finishing the pretraining of both text and image encoders, we can find the centroids and then fix them during the training of the generative model to speed up the training process.

3.4 Discriminator

As mentioned in Section 3.3.1, we use fusion features to combine features from different modalities, and then use them in clustering algorithms to pull different-modal features with similar semantic meaning together into the same cluster. To encourage our network to produce better fusion features, which in turn lead to better visual and language feature representations, we suggest to feed fusion features into the discriminator, where the discriminator compares the quality of fusion features created from synthetic images and given sentences to fusion features created from ground-truth images and the sentences. Feeding fusion features into the discriminator can help regularize training and stabilize clustering training. Also, as our discriminator only controls the quality of fusion features, we do not need to separate the discriminator network into two pathways to evaluate image quality and text-image alignment separately, which satisfies our goals to build a concise network.

The discriminator consists of a series of downsampling blocks, and inputs for our discriminator are the fusion features (see Eq. 3). Fusion features are created by fine-grained image regional features and word embeddings. So, the objectives to train the discriminator \mathcal{L}_D are defined as follows:

$$\begin{aligned}\mathcal{L}_D = & \mathcal{L}_{\text{cont}}(v_{\text{s_image}}^r, s_{\text{s_sent}}) + \mathcal{L}_{\text{cont}}(v_{\text{image}}^r, s_{\text{sent}}) \\ & + \mathcal{L}_{\text{cont}}(v_{\text{regions}}^r, s_{\text{words}}) + \mathcal{L}_{\text{cont}}(v_{\text{s_regions}}^r, s_{\text{words}}) \quad (7) \\ & + \mathcal{L}_{\text{cluster}}(X^r) + \mathcal{L}_{\text{GAN}}^D,\end{aligned}$$

$$\mathcal{L}_{\text{GAN}}^D = -\frac{1}{L} \sum_{i=1}^L \log \left(\mathbb{E} \left[\log(D(X_i^r)) \right] + \mathbb{E} \left[\log(1 - D(X_i^f)) \right] \right), \quad (8)$$

where L is the length of a story (i.e., the number of text-image pairs in a story), X^r are real fusion features created by sentences and real image features that are sampled from the real image distribution, X^f are synthetic fusion features created by sentences and fake image features that are sampled from the model distribution. Here, following [50], we use the real stories and their descriptions to train discriminator projection heads. This is because the generated images may have a poor quality, especially at the start of training, and using such low-quality synthetic image and sentence pairs can hurt the training of the projection heads f and g . So, the contrastive losses from fake images are only applied to the generator.

3.5 Generator

As the discriminator adopts word-level information to provide fine-grained training feedback, we propose to utilize it in the generator to improve its capabilities of capturing relations between texts and images. To achieve this, we adopt word-level spatial attention (WSA) [48]. Then, we further extend WSA and propose a bi-directional word-level spatial attention (BWSA), shown in Fig. 2. WSA is based on the implementation of the softmax function to produce a weight matrix, and thus almost all values in WSA are greater than 0. This means that the attention utilizes the scales of positive values to highlight or ignore different image regions, i.e., giving high (or low) positive weights to word-matched (or word-mismatched) image regions. However, not all words have a positive impact on image regions, even the impact is small. So, to keep both negative and positive effects, BWSA only highlights or ignores the relations between words and image regions, rather than the actual values of hidden features. Main functions of BWSA are shown in Fig. 2, and more details are discussed in the supplementary material.

Our generator only consists of a series of upsampling blocks to generate story images at the desired resolution from given descriptions. Meanwhile, it uses BWSA to utilize word-level information to improve its capabilities of capturing different relations. The objectives to train the generator \mathcal{L}_G are defined as follows:

$$\begin{aligned}\mathcal{L}_G = & \mathcal{L}_{\text{cont}}(v_{\text{s_image}}^f, s_{\text{s_sent}}) + \mathcal{L}_{\text{cont}}(v_{\text{image}}^f, s_{\text{sent}}) \quad (9) \\ & + \mathcal{L}_{\text{cont}}(v_{\text{regions}}^f, s_{\text{words}}) + \mathcal{L}_{\text{cluster}}(X^f) + \mathcal{L}_{\text{GAN}}^G,\end{aligned}$$

$$\mathcal{L}_{\text{GAN}}^G = -\frac{1}{L} \sum_{i=1}^L \log \left(\mathbb{E} \left[\log(D(X_i^f)) \right] \right), \quad (10)$$

where f represents features extracted from a synthetic sample.

4 EXPERIMENTS

We evaluate our approach by comparing it with other methods. We adopt StoryGAN [29], CP-CSV [42], DUCO [32], and VLC [31] as baselines. StoryGAN is a sequential conditional generative adversarial network, CP-CSV uses StoryGAN as backbone and adds the other segmentation mask generation pipeline to improve character consistency, and DUCO and VLC are built on top of StoryGAN as well, which rely on auxiliary video captioning networks to keep consistency. StoryGAN, CP-CSV, DUCO, and VLC all have two sets of generative adversarial networks, one for image generation, and one for story generation.

4.1 Datasets

We evaluate our approach on the Pororo-SV dataset. Pororo-SV is a modified version of PororoQA, a dataset for video question answering [19]. In Pororo-SV, each story has five consecutive images, along with their text descriptions. There are 13,000 story samples in the training set, and 2,336 story samples in the testing set.

Here, we do not evaluate our approach on CLEVR-SV [29], as there are only 15 different words in the entire CLEVR-SV dataset, which might fail to fully explore the multi-modal story visualization task. So, we adopt Abstract Scenes [54, 55] to further evaluate our approach. Abstract Scenes was proposed for studying semantic information, which contains over 1,000 sets of 10 semantically similar scenes of children playing outside. The scenes are composed of 58 clip-art objects, and there are six sentences describing different aspects of a scene. In this dataset, we treat scenes from the same set as a story, as they are all created from the same seed scene, sharing similar semantic information. Following [55], we reserve 1000 samples as the testing set and 497 samples for validation.

4.2 Implementation

We evaluate our approach at the resolution 64×64 on Pororo-SV and 256×256 on Abstract Scenes, as Abstract Scenes provides larger-scale ground-truth images. To work on images at the resolution 256×256 , we repeat the same upsampling blocks in the generator and downsampling blocks in the discriminator for baseline methods. For implementing online clustering learning, the text encoder is a pretrained bi-directional LSTM [48], and the image encoder is a VGG-16 [41], pretrained on ImageNet. For offline clustering, we adopt the same text and image encoders, but train both using Eq. 6 from scratch, and freeze both during the training of the generative model. The image batch size is $N = 40$. To do clustering, we combine 10 batches of instances together, and the number of clusters k in K-means is set to 8. The hyperparameter is $\delta = 0.001$. The network is trained for 240 epochs on Pororo-SV and Abstract Scenes. The Adam optimizer [20] is adopted with learning rate 0.0002. We evaluate our approach on a single Quadro RTX 6000 GPU.

4.3 Evaluation Metrics

The Fréchet inception distance (FID) [15] and the Fréchet story distance (FSD) [42] are adopted. FID computes the Fréchet distance between the distribution of real images and the distribution of fake images. As FID focuses on single image, FSD is proposed for the story visualization task, which takes the sequence of images into account. FSD is built on the principle of FID by using $R(2+1)$ [44]

Table 1: Quantitative evaluation between different methods on Pororo-SV and Abstract Scenes. For FID, FSD, and the number of parameters on Pororo-SV, lower is better; for text-image cosine similarity (Cosine), higher is better.

Method	Pororo-SV dataset			Abstract dataset			Number of Trainable Parameters	
	FID↓	FSD↓	Cosine↑	FID↓	FSD↓	Cosine↑	Generator	Discriminator
StoryGAN [29]	78.64	94.53	0.22	135.16	55.80	3.59	47.0M	47.2M
CP-CSV [42]	67.76	71.51	0.32	-	-	-	86.9M	70.9M
DUCO [32]	95.17	171.70	0.08	142.34	49.16	3.95	53.2M	47.2M
VLC [31]	94.30	122.07	0.21	-	-	-	54.5M	47.2M
Ours	39.17	41.18	11.93	76.51	19.74	9.28	72.4M	10.4M

as backbone model, where $R(2+1)$ has a flexible sequence length and the strong ability to capture temporal consistency.

However, as both FID and FSD cannot reflect the semantic alignment between sentences and story images, we compute the average cosine similarity (Cosine) between pairs of sentence and synthetic image over the testing set, and further scale the value by 100.

Besides, we show the number of parameters in the generator and the discriminator for different methods on Pororo-SV to compare the size of different networks.

4.4 Quantitative Evaluation

We conduct a quantitative comparison between different methods on both Pororo-SV and Abstract, shown in Table 1. We find that our approach achieves the best results in all evaluation metrics on both datasets. Our approach improves the FID score from 78.64 to 39.17, and the FSD score from 94.53 to 41.18 on Pororo-SV, and creates a strong benchmark FID score of 76.51 and FSD score of 19.74 on Abstract, neither relying on segmentation masks nor complex video captioning auxiliary networks. Note that we do not evaluate CP-CSV and VLC on Abstract Scenes, as CP-CSV requires segmentation masks and VLC has not released the complete code.

Moreover, in Table 1, we show the number of parameters in the generator and the discriminator for different methods. Our approach has a smaller number of parameters. Compared to StoryGAN, the total number of parameters in the network reduces about 12.1%, especially reducing about 77.9% in the discriminator.

4.5 Qualitative Evaluation

In Fig. 3, we show the qualitative comparison between different methods on both Pororo-SV and Abstract. Although our approach has a single GAN with a fewer number of parameters, the synthetic story images have better image details: characters produced by our approach can be easily identified (e.g., penguin Pororo and pink bear Loopy) with clear outline and finer appearance (e.g., glasses and a hat on Pororo). As for Abstract, objects produced by our approach have a fine-grained shape (e.g., realistic characters boy Mike and girl Jenny), and ensure a good consistency between images within a story (e.g., the word sun only appears in the fifth sentence, but all images generated by our approach contain a sun in the sky).

4.6 Human Evaluation

A human evaluation study on Pororo-SV is conducted to further evaluate the effectiveness of our approach. Following StoryGAN [29], three evaluation criteria are chosen: (1) visual quality, (2) text-image semantic alignment, and (3) consistency across story images. In

Table 2: Human evaluation on Pororo-SV between VLC, DUCO, and Ours based on three criteria.

Choice (%)	Ours	VLC	DUCO
Visual Quality	78.67	11.33	10.00
Alignment	73.67	16.00	10.33
Consistency	82.00	9.67	8.33

this human evaluation, we ask workers to decide which sample is the best, where each sample contains story images and corresponding sentences. 100 randomly selected samples are assigned to three workers to reduce the human variance. As shown in Table 2, workers prefer the synthetic story images that are generated by our approach on these three evaluation criteria.

4.7 Ablation Study

In Table 3, we thoroughly evaluate different losses and analyze their impact on the Pororo-SV testing set, including: (1) story image-story sentence (SI-SS), (2) image-sentence (I-S), (3) region-word (R-W), (4) story region-word (SR-W), and (5) clustering (cluster).

4.7.1 Individual contrastive and clustering losses. Table 3 shows that using any of the contrastive and clustering losses improves all metrics compared to the baseline. The largest FID and FSD improvements come from the region-word contrastive loss, which improves FID from 96.54 to 50.09 and FSD from 103.72 to 67.17, respectively. However, using clustering loss alone would not significantly improve the performance. The ablations highlight the effectiveness of contrastive and clustering losses, where different contrastive losses maximize the mutual information at different representation levels, and the clustering loss helps capture semantic similarity and prevent contrastive learning pushing instances with similar semantic meaning far away from each other.

4.7.2 Combined contrastive and clustering losses. Combining contrastive and clustering losses further improves all metrics. Using all contrastive losses achieves a better performance (FID 49.22 and FSD 54.47) than each individual. This demonstrates that different representation level contrastive losses are complementary, and using all of them can comprehensively explore mutual information between pairs to achieve the best improvement. Furthermore, contrastive and clustering losses also complement each other: the best FID and FSD scores come from combining all contrastive and clustering losses together, to obtain our final results.

4.7.3 Different choices of contrastive methods. As discussed in Section 3.2, we suggest to adopt the masked margin softmax function

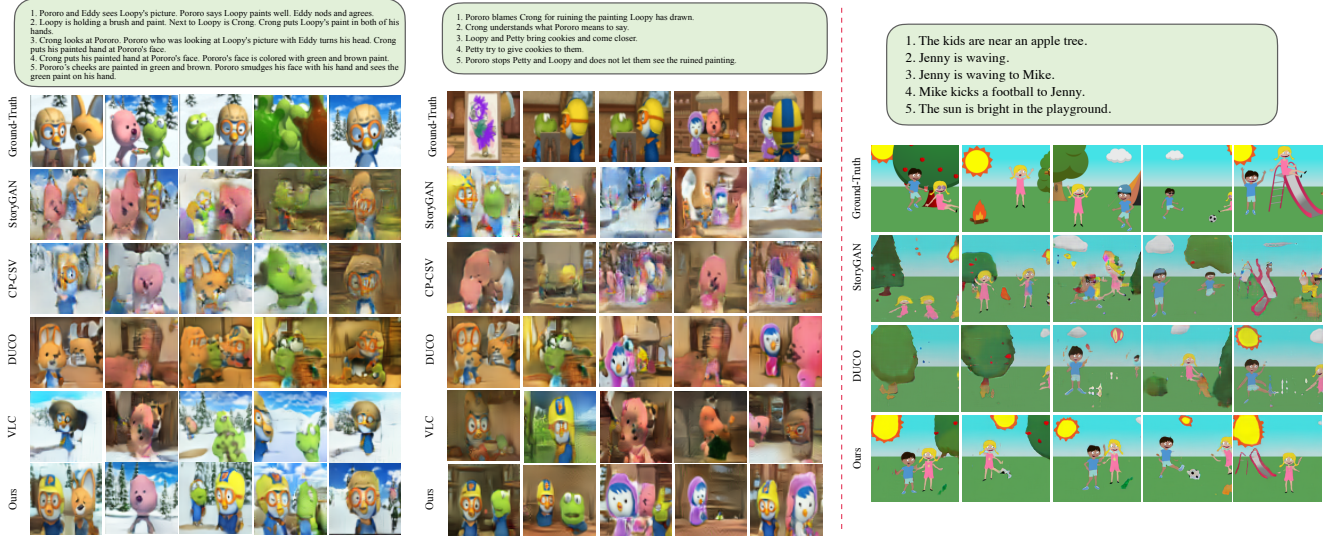


Figure 3: Qualitative comparison between different methods on Pororo-SV and Abstract.

Table 3: Ablation results with different contrastive losses and clustering losses on Pororo-SV, including: (1) story image-story sentence (SI-SS), (2) image-sentence (I-S), (3) region-word (R-W), (4) story region-word (SR-W), and (5) clustering (Cluster).

SI-SS	I-S	R-W	SR-W	Cluster	FID	FSD	Cosine
					96.54	103.72	1.53
✓					72.36	76.14	2.01
	✓				79.95	81.27	2.29
		✓			50.09	67.17	7.64
			✓		86.83	79.80	1.62
				✓	90.09	89.20	1.68
✓	✓	✓	✓		49.22	54.47	10.37
✓	✓	✓	✓	✓	39.17	41.18	11.93

(MMS) [17], instead of the normalized temperature-scaled cross-entropy loss (NT-Xent) [8], to define the similarity between representations from visual and language information and thus to exploit more additional negative. Here, we conduct a comparison study by replacing MMS with NT-Xent in all computation of contrastive losses, shown in Table 4. As we can see, using NT-Xent achieves worse scores on all metrics, degrading FID from 39.17 to 51.82, and FSD from 41.18 to 52.13. This indicates that comprehensively learning against a wider set of negative examples, namely, all those in the batch that are not known to be positively associated, can better explore mutual information across modalities, and thus improve image quality, text-image alignment, and story consistency.

4.7.4 Different choices of clustering methods. We evaluate different choices of clustering learning in our approach, including (1) selection of different clustering algorithms, such as hierarchical clustering and K-means, (2) inputs for clustering algorithms, i.e., fusion features and individual features from their own modality, and (3) online and offline clustering learning. Detailed descriptions are included in the supplement. As shown in Table 4, first, there are no significant changes on all metrics by using different clustering algorithms, so the choices of different clustering algorithms depend on the specific requirements of the tasks, and the strengths and weaknesses of each algorithms. Second, using individual features

degrades the performance of our method, reflected by worse FID and FSD. This indicates that using fusion features encourages each modality feature to move closer to the semantic centroid, by explicitly encouraging semantically close features from different domains to cluster together. Third, online clustering learning achieves a better performance compared to offline learning, reflected by the improvement on all metrics, which means that the training prediction heads along with generative models can better convert different modal features into a joint space to achieve a better translation from language inputs into visual contexts.

4.7.5 Number of Clusters. Table 4 shows the results using different number of cluster sizes for K-means. As we can see, results considerably improve when the number of clusters increases from 4 to 8, and become stable after 8 on Pororo-SV. Based on this, we set the number of clusters $k = 8$ for our clustering algorithm.

4.7.6 Visualization of effects of clustering learning. We visualize synthetic story images on Pororo-SV using t-SNE plots, shown in Fig. 4. For this plot, we randomly choose 16 story sentences and feed them into three methods separately, to generate 16 stories with 80 images for each method. Then, we use a pretrained VGG-16 to extract a global image feature from each image with dimension \mathbb{R}^{356} and visualize these features using t-SNE. Basically, images within the same story are close to each other, as they have a certain

Table 4: Comparison of our proposed approach with different choices on Pororo-SV.

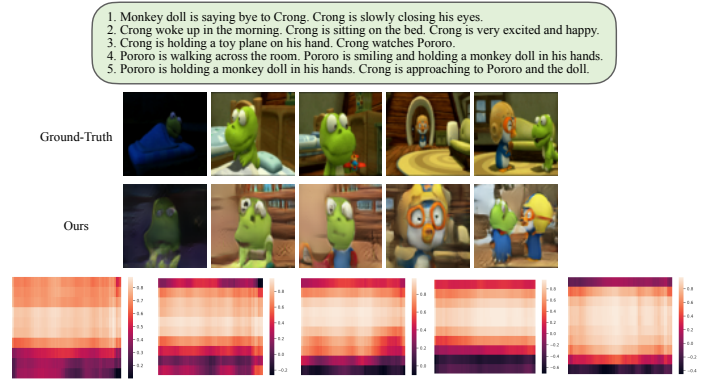
Choice	FID	FSD
w/ NT-Xent [8]	51.82	52.13
w/ Hierarchical Clustering	46.31	43.10
w/ Individual Feature	58.71	59.14
w/ Offline Learning	51.32	50.54
w/o BWSA	61.12	64.23
w/ WSA	50.13	53.25
4	186.77	282.61
8	39.17	41.18
16	43.95	42.46
32	42.87	45.29
64	45.34	45.83
Ours: MMS + K-means (8) + Fusion + Online + w/ BWSA	39.17	41.18

**Figure 4: t-SNE plots of effects of clustering learning. Images from the same story have the same color.**

semantic similarity, and images from different stories may be close if they have similar semantic meaning. We observe that with the help of contrastive learning only (middle), semantically related instances (e.g., instances from the same story) tend to be more tightly related than without using both learning methods (left), and using clustering learning (right) can further pull similar instances together than using contrastive learning only. Also, image features from different stories are clearly more separable using clustering learning, if the stories have a low semantic similarity.

4.7.7 Bi-directional attention. By Table 4, without BWSA, the evaluation results on metrics degrade. This demonstrates that BWSA is complementary to the discriminator, where the discriminator uses information from story-level to word-level to provide comprehensive training feedback to the generator, and under the guidance of such fine-grained feedback, the generator uses BWSA to capture relations between language and visual information.

Furthermore, we replace our BWSA with word-level spatial attention [48] (WSA) to verify its effect. Compared to w/o BWSA, w/ WSA improves all scores, but the scores are still worse than the full model, which indicates that the negative relation in our proposed BWSA can further improve the output results. We attribute this improvement to the comprehensive capture of different relations between words and image regions, instead of the positive effects that are only contained in WSA. We think that our proposed BWSA is a complete version of word-level spatial attention that captures both positive and negative relations between image regions and different words, where a positive effect of BWSA works similarly to WSA to highlight image regions corresponding to semantic words, and a negative effect of BWSA works as complementary component to capture relations that cannot be observed by the positive part.

**Figure 5: Heat map of bi-directional attention.**

4.8 Visualization of Bi-Directional Attention

In Fig. 5, we show the heat map of the relation matrix $\alpha \in \mathbb{R}^{(H*W) \times L}$ (shown in Fig. 2, red region) to visualize both positive and negative effects between image regions and words that are captured by our proposed BWSA. The bottom is the heat map of relations between image regions and words, corresponding to each synthetic story image. We extract the relation matrix α at the resolution 64×64 . However, there are a total of 4096 spatial locations in the hidden features, which is not good for visualization. So, we use 64 image regions to represent the whole spatial locations by computing the average in each region. Also, the number of words varies, and thus we choose the first 10 words for visualization. Thus, the y and the x axis of each heat map represent 10 different words and 64 different image regions, respectively. As we can see, the values in the heat map range from -1 to 1 , instead of ≥ 0 , and dark colors (< 0) appear in multiple locations, which indicates that there exist both positive and negative relations between words and image regions, and thus to comprehensively capture all relations can further improve the generation of story images.

5 CONCLUSION

We studied the task of story visualization, and proposed a concise single-GAN-based network, based on contrastive and clustering learning. Contrastive learning maximizes mutual information between language and visual information, and clustering learning captures semantic similarity across modalities. Besides, we further extended word-level spatial attention to comprehensively capture both positive and negative relations between words and image regions. Experimental results demonstrate the superior performance of our proposed approach, even if it has a small number of parameters compared to the baselines.

ACKNOWLEDGMENTS

This work was supported by the UKRI Turing AI Fellowship EP/W002981/1 and the EPSRC/MURI grant EP/N019474/1. We also thank the Royal Academy of Engineering and FiveAI. This work was also supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1, by the AXA Research Fund, and by the EPSRC grant EP/R013667/1. We also acknowledge the use of the EPSRC-funded Tier 2 facility JADE (EP/P020275/1) and GPU computing support by Scan Computers International Ltd.

REFERENCES

- [1] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. 2020. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems* 33 (2020), 9758–9770.
- [2] Yuki Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. 2020. Labelling unlabelled videos from scratch with multi-modal self-supervision. *Advances in Neural Information Processing Systems* 33 (2020), 4660–4671.
- [3] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. 2019. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371* (2019).
- [4] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. 2019. Conditional GAN with Discriminative Filter Generation for Text-to-Video Synthesis.. In *IJCAI*, Vol. 1. 2.
- [5] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*. 132–149.
- [6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems* 33 (2020), 9912–9924.
- [7] Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, et al. 2021. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8012–8021.
- [8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.
- [9] Yu Deng, Jialong Yang, Dong Chen, Fang Wen, and Xin Tong. 2020. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5154–5163.
- [10] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. 2017. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 5706–5714.
- [11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2672–2680.
- [12] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. 2015. Draw: A recurrent neural network for image generation. In *International Conference on Machine Learning*. PMLR, 1462–1471.
- [13] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. 2018. Imagine this! scripts to compositions to videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 598–613.
- [14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.
- [15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*. 6626–6637.
- [16] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. 2019. Semantic object accuracy for generative Text-to-Image synthesis. *arXiv preprint arXiv:1910.13321* (2019).
- [17] Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. 2019. Large-scale representation learning from visually grounded untranscribed speech. *arXiv preprint arXiv:1909.08782* (2019).
- [18] Minguk Kang and Jaesik Park. 2020. Contragan: Contrastive learning for conditional image generation. *Advances in Neural Information Processing Systems* 33 (2020), 21357–21369.
- [19] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. Deepstory: Video story qa by deep embedded memory networks. *arXiv preprint arXiv:1707.00836* (2017).
- [20] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [21] Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
- [22] Kwot Sin Lee, Ngoc-Trung Tran, and Ngai-Man Cheung. 2021. Infomax-gan: Improved adversarial image generation via information maximization and contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3942–3952.
- [23] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. 2019. Controllable text-to-image generation. In *Advances in Neural Information Processing Systems*. 2063–2073.
- [24] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. 2020. ManiGAN: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7880–7889.
- [25] Bowen Li, Xiaojuan Qi, Philip Torr, and Thomas Lukasiewicz. 2020. Image-to-Image Translation with Text Guidance. *arXiv preprint arXiv:2002.05235* (2020).
- [26] Bowen Li, Xiaojuan Qi, Philip Torr, and Thomas Lukasiewicz. 2020. Lightweight generative adversarial networks for text-guided image manipulation. *Advances in Neural Information Processing Systems* 33 (2020), 22020–22031.
- [27] Bowen Li, Philip Torr, and Thomas Lukasiewicz. 2021. Memory-Driven Text-to-Image Generation. (2021).
- [28] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. 2020. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966* (2020).
- [29] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuxin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6329–6338.
- [30] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. 2018. Video generation from text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [31] Adyasha Maharana and Mohit Bansal. 2021. Integrating Visuospatial, Linguistic and Commonsense Structure into Story Visualization. *arXiv preprint arXiv:2110.10834* (2021).
- [32] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2021. Improving Generation and Evaluation of Visual Stories via Semantic Consistency. *arXiv preprint arXiv:2105.10026* (2021).
- [33] Louis Mahon, Eleonora Giunchiglia, Bowen Li, and Thomas Lukasiewicz. 2020. Knowledge graph extraction from videos. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 25–32.
- [34] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. 2018. Text-adaptive generative adversarial networks: Manipulating images with natural language. In *Advances in Neural Information Processing Systems*. 42–51.
- [35] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).
- [36] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. 2017. To create what you tell: Generating videos from captions. In *Proceedings of the 25th ACM international conference on Multimedia*. 1789–1798.
- [37] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. 2020. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*. Springer, 319–345.
- [38] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1505–1514.
- [39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.
- [40] Scott Reed, Zeynep Akata, Xinchun Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396* (2016).
- [41] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [42] Yun-Zhu Song, Zhi Rui Tam, Hung-Jen Chen, Huiao-Han Lu, and Hong-Han Shuai. 2020. Character-Preserving Coherent Story Visualization. In *European Conference on Computer Vision*. Springer, 18–33.
- [43] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. 2020. DF-GAN: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865* (2020).
- [44] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6450–6459.
- [45] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. 2016. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*. 4790–4798.
- [46] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints* (2018), arXiv:1807.
- [47] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. 2020. Scan: Learning to classify images without labels. In *European Conference on Computer Vision*. Springer, 268–285.
- [48] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1316–1324.
- [49] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. 2020. Clusterfit: Improving generalization of visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6509–6518.
- [50] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of*

- the *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 833–842.
- [51] Han Zhang, Tao Xu, Hongsheng Li, Shaoqing Zhang, Xiaoqiang Wang, Xiaoqi Huang, and Dimitris N. Metaxas. 2017. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 5907–5915.
- [52] Han Zhang, Tao Xu, Hongsheng Li, Shaoqing Zhang, Xiaoqiang Wang, Xiaoqi Huang, and Dimitris N. Metaxas. 2018. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 8 (2018), 1947–1962.
- [53] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5802–5810.
- [54] C Lawrence Zitnick and Devi Parikh. 2013. Bringing semantics into focus using visual abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3009–3016.
- [55] C Lawrence Zitnick, Devi Parikh, and Lucy Vanderwende. 2013. Learning the visual interpretation of sentences. In *Proceedings of the IEEE International Conference on Computer Vision*. 1681–1688.