Pattern Recognition 110 (2021) 107573

Contents lists available at ScienceDirect

Pattern Recognition

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

Unsupervised text-to-image synthesis

Yanlong Dong[a,1], Ying Zhang[b,1], Lin Ma[c,*], Zhi Wang[d,e], Jiebo Luo[f]

[a] Tsinghua University, China
[b] Tencent AI Lab, China
[c] Meituan-Dianping Group, China
[d] Tsinghua Shenzhen International Graduate School, China
[e] Peng Cheng Laboratory, China
[f] University of Rochester, USA

ARTICLE INFO

Article history:
Received 16 February 2020
Revised 20 June 2020
Accepted 4 August 2020
Available online 20 August 2020

Keywords:
Text-to-image synthesis
Generative adversarial network (GAN)
Unsupervised training

ABSTRACT

Recently, text-to-image synthesis has achieved great progresses with the advancement of the Generative Adversarial Network (GAN). However, training the GAN models requires a large amount of pairwise image-text data, which is extremely labor-intensive to collect. In this paper, we make the first attempt to train a text-to-image synthesis model in an unsupervised manner, which does not require any human labeled image-text pair data. Specifically, we first rely on the visual concepts to bridge two independent image and sentence sets and thereby yield the pseudo image-text pair data, based on which one GAN model can thereby be initialized. One novel visual concept discrimination loss is proposed to train both generator and discriminator, which not only encourages the image expressing the true local visual concepts but also ensures the noisy visual concepts contained in the pseudo sentence being suppressed. Afterwards, one global semantic consistency regarding to the real sentence is used to adapt the pretrained GAN model to real sentences. Experimental results demonstrate that our proposed unsupervised training strategy is able to generate favorable images for given sentences, which even outperforms some existing models trained in the supervised manner. The code of this paper is available at https://github.com/dylls/Unsupervised_Text-to-Image_Synthesis.

1. Introduction

Recently, synthesizing images from natural language descriptions [1,2] has been attracting more and more attentions in research communities, due to its great importance in many practical applications, such as cross-modal retrieval [3–6], captioning [7–10], image manipulation [11], industrial design [12], and so on.

Most prevailing models for the text-to-image synthesis relies on recently proposed Generative Adversarial Network (GAN) [13], which is usually realized in an encoder-decoder-discriminator architecture. The input sentence is first encoded as one latent vector and injected into one decoder to produce photo-realistic image [2,14,15]. The discriminator is used to ensure that the generated image is not only visually realistic but also semantically consistent with the input sentence.

For example, Reed *et.al* [14] first utilize a GAN to generate one realistic image conditioned on the visually-discriminative vector representation of one text description. StackGAN [16] and StackGAN++ [17] improve the generated image quality by stacking multiple GANs, where low-resolution and high-resolution images are generated stage-by-stage. AttnGAN [2] attempts to pay more attentions on different sub-regions conditioned on different words at different stages. Recently, in order to eliminate the ambiguity of the textual descriptions for generating images and thereby relieve the difficulties of text-to-image synthesis task, the scene graph [18] or image layout [19] is taken as the input to generate visually pleasant images, which semantically corresponds to the input scene graph or layout. Although great progresses have been achieved, in order to train a GAN for synthesizing images conditioned on the text/layout/scene graph, a large amount of image-sentence (or image-layout or image-scene graph) pairs is inevitably needed. However, compared with image classification, collecting and annotating image-sentence pairwise data is much more complicated and labor-intensive. Therefore, how to relieve the dependency on manually annotated pairwise data for the text-to-image synthesis

* Corresponding author.
E-mail addresses: dyl17@mails.tsinghua.edu.cn (Y. Dong), yinggzhang@tencent.com (Y. Zhang), forest.linma@gmail.com (L. Ma), wangzhi@sz.tsinghua.edu.cn (Z. Wang), jluo@cs.rochester.edu (J. Luo).
[1] Equal contribution

https://doi.org/10.1016/j.patcog.2020.107573
0031-3203/

task becomes increasingly important, which deserves further investigations.

Training a text-to-image synthesis model in an unsupervised manner with no manually annotated image-text pairwise data is very difficult. Different from the supervised approaches focusing on designing more effective visual-textual interaction models to improve the visual quality of generated images, unsupervised text-to-image synthesis is still far from being fully explored, which issues new challenges. First, the most critical challenge is how to train a reliable generative model and make it work without any manually labeled image-text pair data. Second, how can we ensure that the generated images express the local visual concept information, which is contained in the input text. Third, how can we guarantee that the generated image is both visually realistic and semantically consistent with the input text.

In this paper, we make the first attempt to tackle the unsupervised text-to-image synthesis task to further address the aforementioned challenges. More specifically, we start with mining visual concepts from one external sentence corpus and thereafter train one concept-to-sentence generation model. Based on the concept-to-sentence model, a pseudo sentence for each image is produced based on the visual concepts detected from the image. Afterwards, the generated pseudo image-text pairs are used to train one text-to-image generative model in one supervised manner, where a novel visual concept discrimination loss is proposed to encourage the generated image to not only express the visual concept information but also suppress the noisy concepts contained in the generated sentence. Moreover, one global semantic consistency loss is used to readjust the generative models with real sentence corpus, which further enhances the semantic consistency between the generated image and input real sentence.

In summary, our contributions lie in three-fold.

- To the best of our knowledge, we make the first attempt to train one text-to-image synthesis model in an unsupervised manner, with no reliance on any human labeled image-text pair data.
- A novel visual concept discrimination loss is proposed to train both generator and discriminator, which not only encourages the generated image expressing the local visual concepts but also ensures the noisy visual concepts contained in the pseudo sentence being suppressed. One global semantic consistency loss is used to ensure that the generated image semantically corresponds to the input real sentence.
- Experimental results on the public MSCOCO dataset demonstrate that our proposed model can generate one favorable image for one given sentence, with no reliance on any image-text pair data, which even outperforms some text-to-image synthesis models trained in the supervised manner.

2. Related work

2.1. Text-to-image synthesis

Text-to-image synthesis is newly emerged research area and has been drawing more and more attention in recent years. As a pioneer work, Reed et al. [14] introduced a GAN-based architecture to generate images conditioned on the visually-discriminative representation of an input text. Although the method has shown great potential of GAN in synthesizing photo-realistic image relevant to the text descriptions, the generated images were restricted to a low resolution of 64×64. To generate images with higher resolution, StackGAN [16] and StackGAN++ [17] stacked multiple GANs to generate images from low-resolution to high-resolution stage-by-stage. Based on the multi-stage strategy, AttnGAN [2] attempted to draw more details of sub-regions conditioned on different words at different stages. Gao et al. [20] proposed a pyramid framework which utilizes one generator to directly synthesize high-quality images, with three discriminators regularizing the generated images at different scales. MirrorGAN [21] proposed to re-describe the generated images for enhancing the semantic consistency. However, the success of some existing methods [14–16] were limited to simple scenarios, and had difficulty in presenting multiple objects and their relations when generating images from complex descriptions. To address this issue, Hong et al. [19] and Li et al. [1] construct the semantic layouts from text descriptions to guide the image generation, where a box generator first estimates the bounding box layout of an images, followed by a shape generator to refine the pixel-level synthesis.

2.2. Unsupervised generation

Unsupervised learning of generating a specific object given an input has been extensively studied in various tasks, such as machine translation [22–24], image description generation [25,26], unsupervised image-to-image translation [27,28] and person image generation [29]. Unsupervised machine translation approaches [22,23] generally build a common latent space between the two languages, and the sentences from two languages are semantically aligned to perform translation. Su et al. [24] utilized images to assist the unsupervised machine translation, based on the assumption that the description of the same visual content by different languages should be approximately similar. Feng et al. [25] attempted to conduct unsupervised image captioning with the interconnection of visual concepts, while Chen et al. [26] generates stylish image description in an unsupervised manner by learning a joint space for paired-unstylish captions and monolingual corpus of a specific style. Liu et al. [27] adopted coupled GANs for unsupervised image-to-image translation, based on the assumption that a pair of corresponding image from different domains can be mapped to a same latent representation. Stacked Cycle-Consistent Adversarial Networks (SCANs) [28] further refines the transferred images with multi-stage learning similar to [16,17]. Song et al. [29] addressed the unsupervised pose-guided person image generation via exploiting the transformation between semantic parsing maps of different person images.

While many unsupervised generation tasks have been explored to alleviate the burden of labeling data, how to generate images from natural language descriptions in an unsupervised manner is still left untouched. Compared with unsupervised image-to-image [27,29] and language-to-language generation [22–24], unsupervised text-to-image synthesis is more challenging due to the difficulty in generating visually-semantic images, as well as the significant property differences between visual and textual modalities.

3. Background

In this section, we briefly review the encoder-decoder-discriminator architecture for training the text-to-image synthesis model in a supervised manner.

Encoder. As we aim to generate one natural image for one given sentence, we need to first encoder the sentence to generate its semantic representation, where one recurrent neural network (RNN), specifically long short-term memory (LSTM) or gated recurrent unit (GRU), is naturally suitable for modeling sentence. Given one input sentence $S = \{w_1, w_2, \cdots, w_R\}$, we feed w_r at time r into the RNN unit to generate the corresponding hidden state h_r:

$$h_r = \mathrm{RNN}(w_r, h_{r-1}). \tag{1}$$

The hidden state h_R of last time step R encodes the whole information of sentence, which can be thereby regarded to express its global semantic meaning. And the hidden state sequence $\mathbf{h} =$

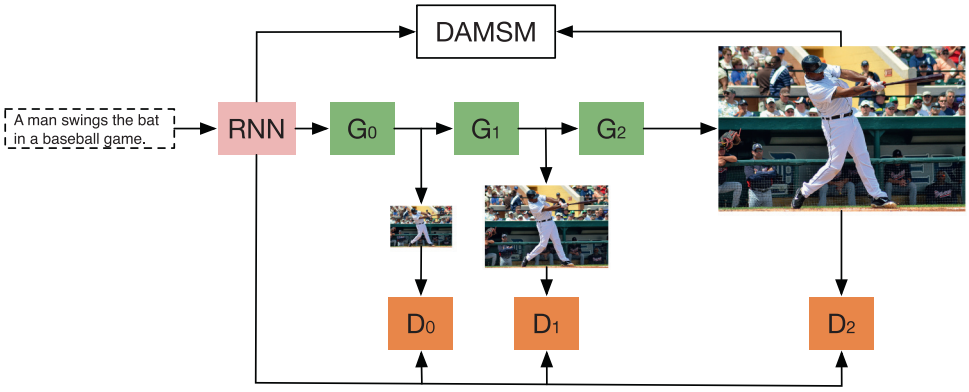

Fig. 1. The framework of AttnGAN for the text-to-image synthesis task, where RNN encodes the sentence, $\{G_0, G_1, G_2\}$ and $\{D_0, D_1, D_2\}$ are the corresponding generators and discriminators at different stages, respectively. DAMSM is used to evaluate the fine-grained image-text matching relationships.

$\{h_1, h_2, \cdots, h_R\}$ generated during the encoding process can also be regarded to contain the fine-grained word-level semantic information.

Decoder. Conditioned on the semantic representation of the given sentence, the decoder aims to generate one image, which is realized by stacking multiple generators. Taking AttnGAN [2] as one example, its framework is illustrated in Fig. 1. Specifically, three generators $\{G^0, G^1, G^2\}$ take the corresponding generated hidden states $\{x^0, x^1, x^2\}$ as input and decode images $\{\hat{I}^0, \hat{I}^1, I^2\}$ stage-by-stage from low resolution to high resolution:

$$\begin{aligned} x^0 &= F^0(z, f(h_R)), \\ x^t &= F^t(x^{t-1}, f_a^t(\mathbf{h}, x^{t-1}), \\ \hat{I}^t &= G^t(x^t), \end{aligned} \tag{2}$$

where z is one noise vector sampled from one standard normal distribution. $f(h_R)$ performs the conditioning augmentation process [16], converting the global sentence representation to one conditioning vector. f_a^t denotes the attention strategy performed at t-th stage. F^t and G^t are the neural networks to generate the image $\hat{I}^t$ at the t-th stage. Please refer to Xu et al. [2] for more detailed information about AttnGAN.

Discriminator. As the image are generated stage-by-stage, multiple discriminators, namely $\{D^0, D^1, D^2\}$ are used at different stages to discriminate the input image as real or not, as shown in Fig. 1.

The text-to-image synthesis model targets at not only synthesizing photo-realistic image but also expressing semantically consistent meaning with the input sentence. To this end, as stated in [2], each discriminator D^t is trained to classify the input image into the class of real or fake by minimizing the cross-entropy loss $\mathcal{L}^{uncond}$ [17]. For each generator, consisting of the encoder and decoder, besides the GAN loss yielded from the discriminator at t-th stage, an additional fine-grained image-text matching loss, termed as $\mathcal{L}_{\text{DAMSM}}$ is computed to measure the image-text similarity at word level for training the generator.

4. Unsupervised text-to-image synthesis

In this paper, we focus on training one text-to-image synthesis model in an unsupervised manner. As aforementioned, the most critical challenge is how to train one generative model and make it work without any manually labeled image-text pair data. In order to tackle such a challenge, we propose to mine the visual concepts contained in the sentence corpus. Afterwards, the visual concepts act as one semantic bridge between images and one sentence to construct one pseudo image-text pair. With such pseudo pairwise data, we can thereby initialize one text-to-image synthesis model in a supervised training manner. Afterwards, the visual conception distillation and the global semantic consistency measurement are proposed to further tuning the text-to-image synthesis model.

4.1. Pseudo image-text pair generation

In order to train one text-to-image synthesis model, we resort to the visual concepts, which perform as one bridge to semantically align the unlabeled images and sentences and thereby produce pseudo image-text pair data.

Formally, under the unsupervised setting, we have an image set $\mathcal{I} = \{I_0, I_1, \ldots, I_{N-1}\}$ and one sentence corpus $\mathcal{S} = \{s_0, s_1, \ldots, s_{M-1}\}$, as well as one image object detector. Please note that the images and sentences are not semantically aligned. First, we start mining the visual concepts contained in each sentence and the class labels of the object detector, and thereby construct one visual concept dictionary $\mathcal{C}$. In order to bridge the sentence and image, the concept dictionary $\mathcal{C}$ is constructed by K visual concepts $\{c_0, c_1, \cdots, c_{K-1}\}$. Meanwhile, we also have the concept-sentence pair data, which can be used to train a sequence-to-sequence (seq2seq) [30] model, as shown in Fig. 2 (a). Please note that the ordering of the visual concepts is same as their ordering in the sentence. With such a seq2seq model, a set of individual visual concepts is transformed into a humanlike language description.

Afterwards, we rely on the existing object detector [31] to obtain the visual concepts contained in one given image. Please note the detected visual concepts are ordered by their detection confidence scores, which will be fed into the learnt seq2seq model and thereby generate one natural sentence, as shown in Fig. 2 (b). The generated sentence $\hat{s}_n$ and the given image I_n thereby form one pseudo image-text pair data.

4.2. Visual concept discrimination

With the obtained pseudo image-text pair data, we can train generative models following the traditional supervised text-to-image synthesis approaches, and readily adopt the network architectures and loss functions introduced in Section 3 to synthesize visually realistic images. While unfortunately, the generated sentence accompanied with the image may contain noisy visual concepts or miss the truly existing visual concepts, compared with the groundtruth sentence, which could easily lead to misalignment between image contents and textual information. As shown in Fig. 3, the generated sentence contains the noisy visual concept "`television`" which does not appear in the real image, while failed to express the truly existed visual concept "`boy`".

To address this issue, we propose one novel visual concept discrimination loss:

$$\mathcal{L}^{dis} = \mathcal{L}(\hat{I}_n^t, \hat{s}_n, Y_n) + \mathcal{L}(I_n, \hat{s}_n, Y_n) + \mathcal{L}(I_j, \hat{s}_n, Y_j). \tag{3}$$

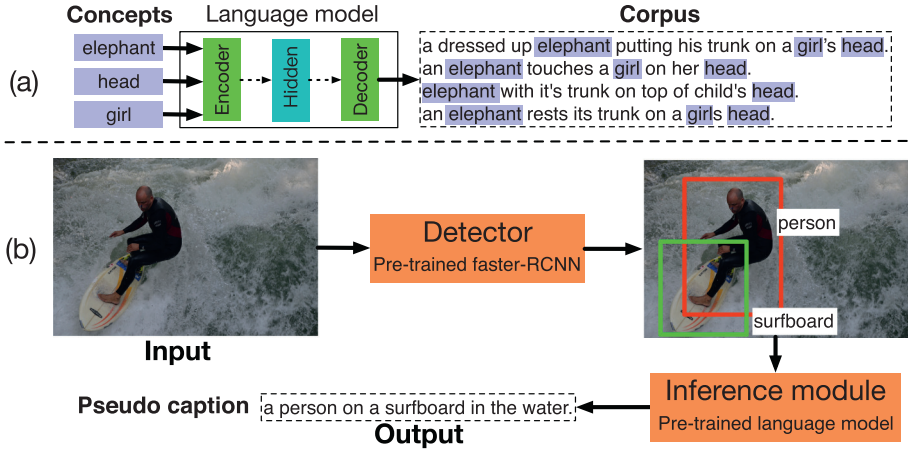

Fig. 2. The pseudo image-text pair generation pipeline. (a) We mine the visual concepts from one sentence corpus, based on which the concept-sentence pairs are collected to train a seq2seq model. (b) The learnt seq2seq model takes the concepts detected from one given image as the input and generates one sentence. The generated sentence and the given image are thereby form one pseudo image-text pair.

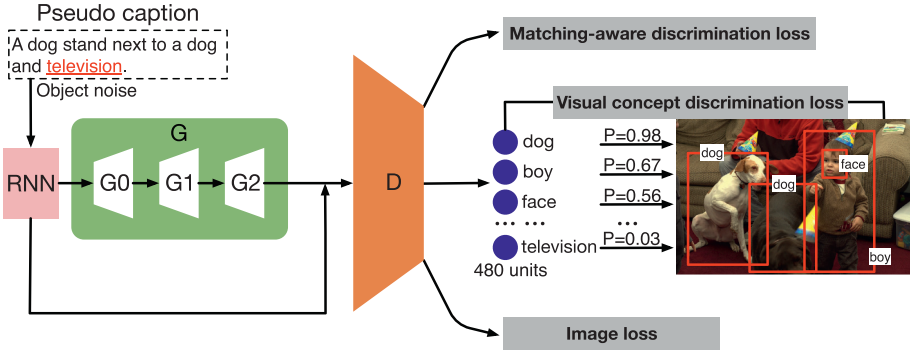

Fig. 3. The proposed visual object discrimination loss, which encourages the generative model to produce images expressing the truly existed visual concepts, such as "boy", and also suppressing the noisy visual concepts, such as "`television`", existing the pseudo generated sentence.

Here the loss $\mathcal{L}(\hat{I}_n^t, \hat{s}_n, Y_n)$ is defined as:

$$\mathcal{L}(\hat{I}_n^t, \hat{s}_n, Y_n) = -\frac{1}{K}\sum_{k=1}^{K}\Big(Y_n^k \log(D_c(\hat{I}_n^T, \hat{s}_n)) + (1 - Y_n^k)\log(1 - D_c(\hat{I}_n^T, \hat{s}_n))\Big), \tag{4}$$

where Y_n and Y_j denotes the concept labels of image I_n and I_j, respectively, with each element $Y_n^k \in \{0, 1\}$ indicating the existence of concept c_k in the real image I_n. $D_c(\hat{I}_n^t, \hat{s}_n)$ is realized by K binary classifiers to yield the corresponding visual concepts expressed by the image-text pair $(\hat{I}_n^t, \hat{s}_n)$. As such, for the generator, $\mathcal{L}(\hat{I}_n^t, \hat{s}_n, Y_n)$ encourages the generated image $\hat{I}_n^t$ at t-th stage together with the obtained pseudo sentence contains the same visual concepts as the real image I_n. Meanwhile, the noisy visual concepts which do not appear in the real image I_n are simultaneously suppressed.

For training the discriminator, besides the generated image I_n^t, we also consider the real image I_n accompanied with the generated sentence $\hat{s}_n$ and randomly sample another real image I_j with no semantic relationships with $\hat{s}_n$. As such, the visual concept discrimination loss defined in Eq. (3) not only encourages the semantically aligned image-text pairs, specifically, $(\hat{I}_n^t, \hat{s}_n)$ and $(I_n, \hat{s}_n)$, to produce the correct concept labels, but also ensures the unrelated pair $(I_j, \hat{s}_n)$ to yield unreliable concept labels. In this paper, as the visual concept label for each image is extremely sparse, therefore, the two labels Y_n and Y_j seem not be able to overlap with each other. As such, Y_j is used and regarded as the unreliable concept label for the pair data $(I_j, \hat{s}_n)$, instead of manually creating the unreliable concept label.

Discussion. One similar visual concept distillation loss is proposed in [25] to encourage the generated caption containing the visual concepts detected from the input image. The most difference lies in that the visual concept distillation loss in [25] is only used to training the generator, while our proposed visual concept discrimination loss not only train the generator but also tune the discriminator by sampling one semantically unrelated image-text pair $(I_j, \hat{s}_n)$. Second, for the text-to-image synthesis task, we mainly focuses on the presence of the visual object while neglect the corresponding probability the object detected from the image. As such, we resort to the K binary classifiers to determine whether the image-text pair expresses the concept labels without considering the confidence score used in [25].

4.3. Global semantic consistency with respect to real sentences

Based on the constructed pseudo image-text pair and the proposed visual concept discrimination loss, we have made it possible to train a generative model for text-to-image synthesis without any labelled image-sentence pairs and improved the ability of generator to express local visual concept information. However, we notice that the generative model is trained with the generated pseudo image-text pair data, which inevitably has distribution deviation with the real sentences. As such, the global semantic consistency between generated images and real human language sentence cannot be guaranteed. In this section, we incorporate the real

sentences to readjust the generative models to make it more suitable for the text-to-image synthesis task. Specifically, we use the matching-aware discrimination loss as [14] and formulate the objective function as:

$$\begin{aligned}\mathcal{L}^{pair} = & -\mathbb{E}_{\hat{I}_m^t \sim p_{G^t}, s_m \sim p_{\mathcal{S}}}[\log(D^t(\hat{I}_m^t, s_m))] \\ & -\mathbb{E}_{I_n \sim p_{\mathcal{I}}, \hat{s}_n \sim p_{\mathcal{S}}}[\log(D^t(I_n, \hat{s}_n))] \\ & -\mathbb{E}_{\hat{I}_m^t \sim p_{G^t}, s_m \sim p_{\mathcal{S}}}[\log(1 - D^t(\hat{I}_m^t, s_m))] \\ & -\mathbb{E}_{I_n \sim p_{\mathcal{I}}, s_m \sim p_{\mathcal{S}}}[\log(1 - D^t(I_n, s_m))] \end{aligned} \tag{5}$$

where the generator G^t is encouraged to generate fake image $\hat{I}_m^t$ to match the real sentence s_m, and the discriminator D^t is learnt to judge the fake pair $(\hat{I}_m^t, s_m)$, the matched pair $(\hat{I}_n, \hat{s}_n)$, and the mismatched pair $(\hat{I}_n, s_m)$ as fake, real, and fake, respectively.

By fine-tuning the generative model with the participation of real sentences in the matching-aware loss, the global consistency between the synthesized images and real sentences can be further enhanced. More theoretically, the generator attempts to transform an sentence sample from one distribution into an image following the target distribution. The generated pseudo sentences cannot accurately approximate the distribution of real sentences and may lead to deviation when we attempt to generate images from real sentences. By incorporating $\mathcal{L}^{pair}$ for fine-tuning the generative model enables the generator to learn more accurate information from real language descriptions, which can further improve the generation results.

4.4. Training

In this paper, we target at training one text-to-image synthesis model without any manually labeled data. In previous sections, we introduce how to construct the pseudo image-text pair to train one generative model. In order to ensure the image express the local visual concept information, one visual concept discrimination loss is introduced. For adapting to real sentence, one global semantic consistency loss is also proposed. In this section, we will provide one brief description of the training pipeline, with detailed information of the initialization and training of each component.

As illustrated in Algorithm 1, training one text-to-image synthesis model is performed in five steps. First, we need to construct one visual concept dictionary $\mathcal{C}$ by jointly considering the class labels of the object detectors and the words containing in the sentences. Such a constructed visual concept dictionary is also useful for the following visual concept discrimination loss. Second, we rely on the visual concepts as one bridge to generate one pseudo sentence for each image, thereby constructing pseudo image-text pairs. Detailed information can be referred to Section 4.1. In order to further training one generative model for text-to-image synthesis, we follow the same procedure in [2] to train one deep attentional multimodal similarity model (DAMSM) based on the constructed pseudo image-text pair. The DAMSM module stays fixed during the training of the text-to-image synthesize model. Afterwards, we train one generative model, specifically AttnGAN, based on the pseudo image-text pair. Please note that the visual concept discrimination loss as introduced in Section 4.2 is also used for both training the generator and discriminator. Finally, the real sentence is taken as the input to finetune the pretrained AttnGAN model, which further considers the global semantic consistency loss introduced in Section 4.3. Specifically, the training losses for the generator and discriminator are formulated as:

$$\begin{aligned}\mathcal{L}_G &= \mathcal{L}^{uncond} + \mathcal{L}^{pair} + \lambda_1 \mathcal{L}^{dis} + \lambda_2 \mathcal{L}^{\text{DAMSM}}, \\ \mathcal{L}_D &= \mathcal{L}^{uncond} + \mathcal{L}^{pair} + \lambda_1 \mathcal{L}^{dis}, \end{aligned} \tag{6}$$

where the unconditional adversarial loss $\mathcal{L}^{uncond}$ and $\mathcal{L}^{\text{DAMSM}}$ are the same as Xu et al. [2]. Please note that $\mathcal{L}^{\text{DAMSM}}$ is evaluated by the DAMSM model trained on the pseudo image-text pair data. The matching-aware discrimination loss $\mathcal{L}^{pair}$ is defined in Eq. (3), and the visual concept discrimination loss $\mathcal{L}_{dis}$ is defined in Eq. (5).

Algorithm 1 The whole pipline for training one text-to-image synthesis model in an unsupervised manner.

Input: one image set $\mathcal{I}$, one sentence set $\mathcal{S}$, and one existing visual concept detector.

Output: one learnt text-to-image synthesis model.

1: Construct the visual concept dictionary $\mathcal{C}$ by jointly considering the class labels of image detectors and the words existing in the sentences;
2: Construct the pseudo image-text pair as illustrated in Sec. 4.1.
3: Relying on $\mathcal{C}$ to contruct the concept-sentence pair data.
4: Training one seq2seq model based on the concept-sentence pair data.
5: Use the learnt seq2seq model to generate one pseudo sentence for each image based on its detected visual concepts.
6: Train one deep attentional multimodal similarity model (DAMSM)~[2] based on the pseudo image-text pair.
7: Train one generative model, specifically AttnGAN~[2], based on the pseudo image-text pair data and incorporating the proposed visual object discrimination loss in Sec. 4.2.
8: Train the generative model, specifically AttnGAN, by taking the real sentence as input and further incorporating the global semantic consistency loss introduced Sec. 4.3.

5. Experiments

5.1. Datasets and settings

Datasets. We conduct experiments on the MSCOCO [32] dataset, which is a large-scale dataset widely used for object detection, image captioning, and text-to-image synthesis. The MSCOCO dataset contains a training set of 82,783 images and a test set of 40,504 images, with each image annotated with 5 sentences. For the unsupervised setting, we split the original training set into 50,000 and 32,783 images, and take the 50,000 images (denoted as $\mathcal{I}$) to train our model, and utilize the sentences from another 32,783 images as external text corpus (denoted as $\mathcal{S}$). The 40,504 images from the original test set is used to evaluate the proposed approach.

Implementation Details. Our model is implemented in PyTorch with a NVIDIA Tesla V100 GPU. For visual concepts mining, we utilize the 480 object categories of the OpenImage dataset [33] to build the visual concept dictionary. For the concept-to-sentence model, we employ a seq2seq [30] model, where the concept encoder and sentence decoder are all realized with one single-layer LSTM with the input size and hidden size setting as 512. For visual concept detection, we adopt Faster R-CNN [31] and select top 5 detected concepts by the detection confidence score. For text-to-image generation, we rely on the architecture of AttnGAN [2] to generate images with resolution as 256×256. We set the hyperparameter as $\lambda_1 = 0.5$, $\lambda_2 = 50$, $\lambda_3 = 1$. The Adam [34] optimizer is employed for optimization with batch size of 32. The model is initialized by train 50 epochs with $lr = 0.0002$. Afterwards, by further incorporating the global semantic consistency loss with respect to the real sentences, the model further fine-tuned with the same learning rate for another 25 epochs.

Evaluation Metrics. Two quantitative metrics are employed to evaluate the proposed method.

- **Inception Score (IS)** [35] is a measurement of objectiveness and diversity of the generated image, which is defined by the

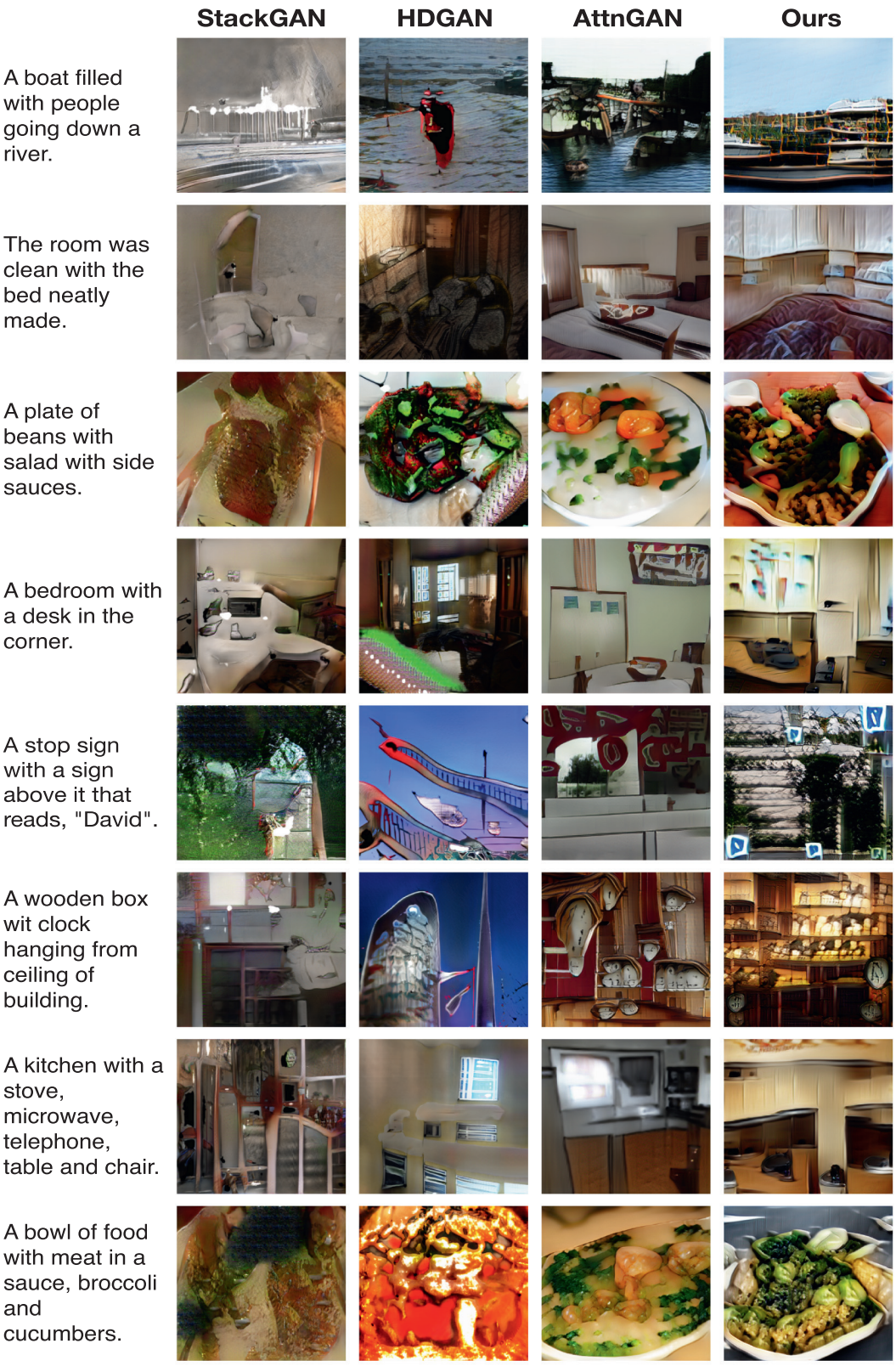

Fig. 4. Qualitative result comparisons of different text-to-image synthesis models. Please note that the StackGAN, HDGAN, and AttnGAN are all trained in one supervised manner, while ours is trained in an unsupervised manner.

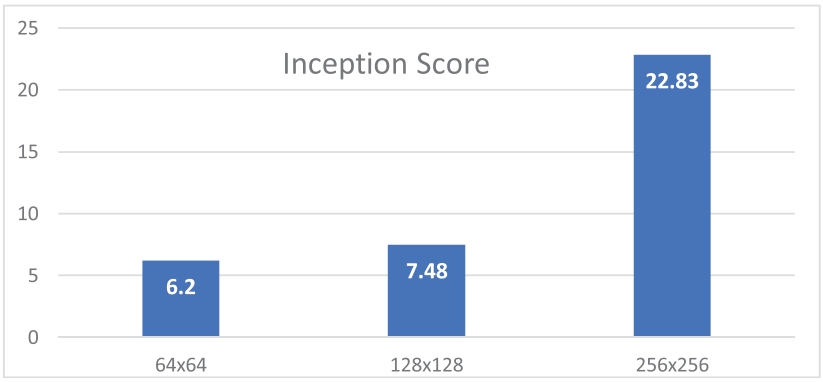

Fig. 5. The quantitative results of our approach with respect to different resolutions.

Table 1
Inception score of different methods on the COCO dataset. Note that our model is trained without any paired data, while others are trained with 40k annotated image-sentence pairs.

Methods	Resolution	Inception Score
GAN-INT-CLS [14]	64 × 64	7.88 ± 0.07
SceneGraph [18]	64 × 64	6.70 ± 0.01
StackGAN [16]	256 × 256	8.45 ± 0.03
PPGN [37]	256 × 256	9.58 ± 0.21
LayoutSynthesis [19]	128 × 128	11.46 ± 0.09
HDGAN [38]	256 × 256	11.86 ± 0.18
AttnGAN [2]	256 × 256	25.89 ± 0.47
Ours(AttnGAN)	256 × 256	22.83 ± .44

KullbackLeibler (KL) divergence between the conditional distribution $p(y|x)$ and marginal distribution $p(y)$:

$$IS(G) = exp\Big(\mathbb{E}_{x\sim p_G} D_{KL}(p(y|x)||p(y))\Big). \quad (7)$$

Same as the existing text-to-image synthesis modes, we also adopt inception-v3 [36] pre-trained on ImageNet to compute IS.

- **R-precision** [2] measures the semantic consistency between generated images and input descriptions. R-precision indicates the probability of an image to correctly retrieve the relevant text from 100 candidates composed by 1 ground truth and 99 mismatched descriptions. Here we adopt the similarity model provided in [2] to perform image-to-text retrieval.

5.2. Experimental results on MSCOCO

Comparison with existing text-to-image synthesis models. Table 1 compares the proposed unsupervised approach against 7 existing supervised method on the MSCOCO dataset. It can be observed that our proposed model, yielding an IS of 22.83, achieves significantly better results than previous supervised methods of [16,19,38]. One reason for the superior performance can be attributed that the network architecture of AttnGAN model. As illustrated in Section 5.3, our proposed visual concept discrimination loss and global semantic consistency loss also play important roles for training the text-to-image synthesis model. Compared with AttnGAN, with the same network architecture, our approach achieves slightly inferior results, (22.83 vs. 25.89). Please note that our approach is trained without any paired data, while the other competitor models are trained with over 400,000 annotated image-sentence pairs. Such significant achievements demonstrate the effectiveness of our proposed strategy for unsupervised text-to-image synthesis.

Moreover, Fig. 4 illustrates some qualitative results, which are generated by different text-to-image models. It can be observed that our approach can generate meaningful images even trained in an unsupervised manner. For some cases, our approach can generate more visually pleasant images than StackGAN and HDGAN, such as the images shown the last two columns.

Quantitative results on multiple resolutions. We can also observe that approaches that can generate higher-resolution images usually perform better than algorithms limited to low-resolution images. Specifically, StackGAN [16], HDGAN [38], AttnGAN [2], and our proposed method produces 256 × 256 images, achieving much higher inception score than GAN-INT-CLS [14] and SceneGraph [18] which only generate 64 × 64 images. As such, we examine the ability of our approach on generating images with different resolutions. Specifically, three stages in AttnGAN trained with our proposed unsupervised strategy yields three different images with the resolutions as 64 × 64, 128 × 128, and 256 × 256. We compare the IS of the images at different resolutions, which is illustrated in Fig. 5. It can be observed that the IS of 256 × 256 significantly outperforms those of 64 × 64 and 128 × 128. The main reason is that the visual conception discrimination loss $\mathcal{L}^{dis}$ and the DAMSM loss $\mathcal{L}^{\text{DAMSM}}$ are only performed on the generated 256 × 256 image. Moreover, the mean IS values of 64 × 64 and 128 × 128 are 6.2 and 7.48, respectively, which are also competitive with the existing models, namely GAN-INT-CLS, SceneGraph, StackGAN, and PPGN.

5.3. Ablation studies

In this section, we perform a series of ablation studies to evaluate the effectiveness of each component in our proposed framework, including the network initialization with pseudo caption, visual concept discrimination, and global semantic consistency referring to real sentences.

Quantitative results. Table 2 reports the inception score and R-precision of four training strategies: (i) "Baseline" denotes training an AttnGAN with pseudo image-caption pairs; (ii) "Baseline + VCD" denotes adding the visual concept discrimination loss on "Base-

Table 2
Impact of different components in the proposed approach with AttGAN as the backbone network.

Methods	Inception Score	R-precision
Baseline	14.87 ± 0.32	27.00 ± 0.60
Baseline+VCD	18.21 ± 0.19	28.96 ± 0.72
Baseline+VCD+GSC	22.83 ± 0.44	32.87 ± 0.58
Ours w/o init	18.19 ± 0.27	32.81 ± 0.89

Fig. 6. Comparison of qualitative results impacted by different components.

line", (iii) "Baseline + VCD + GSC" denotes further fine-tuning the model with real sentences based on "Baseline + VCD". (iv) "Ours w/o init" denotes directly training AttnGAN with real sentences as described in Section 4.3, without pre-training on pseudo image-caption pairs. From the table we can see that training a AttnGAN with only pseudo image-caption pairs has outperformed many existing method by achieving an inception score of 14.87, which may due to two reasons: AttnGAN provides a strong baseline to generate high quality images, and the concept-to-sentence model is able to generate pseudo captions semantically aligned with real

Table 3
Impact of different components in the proposed approach with HDGAN as the backbone network.

Methods	Inception Score
Baseline	10.16 ± 0.12
Baseline+VCD	11.23 ± 0.28
Baseline+VCD+GSC	9.00 ± 0.21
Ours w/o init	8.08 ± 0.15

images. When we add the proposed visual concept discrimination loss, the inception score can be improved to 18.21, and R-precision is improved to 32.87, demonstrating that explicitly encouraging the generated images to express visual concepts could facilitates the image generation. It also shows that fine-tuning the network with real sentences can greatly enhance the global semantic consistency, with an improvement of 4.62 on inception score and 3.91 on R-precision. Here we also report the results of training the model from scratch with real sentences, which only produces an inception score of 18.19. This underlines the importance of initializing model by training with pseudo-caption pairs.

Moreover, in order to further demonstrate the effectiveness of our proposed components for unsupervised text-to-image synthesis, we retrain the unsupervised text-to-image generation with HDGAN as the backbone network, with the results listed in Table 3. It can be observed that, the Inception Score achieves 10.16 by training on the pseudo image-caption pairs. With further incorporating the visual concept discrimination (VCD) loss the Inception Score can be improved to 11.23, which shows that VCD can guide GAN to generate images with more accurate visual concepts. However, by further incorporating global semantic consistency (GSC), the Inception Score decreases by 2.23. Such a result is different from that of AttnGAN as the backbone network. The reason is mainly due to that GSC relies on DAMSM, which has not been considered in HDGAN. However, training HDGAN with no pseudo image-caption pairs for initialization, the Inception Score only achieves 8.08, which validates the effectiveness of the proposed component for constructing the pseudo image-caption pairs.

Qualitative results. Fig. 6 illustrates qualitative comparison of different components. We can see that images generated by "Baseline" only contains rough shape and color but lacks the objects described in input sentences. Visual concept discrimination loss enables the images to express more visual concepts such as "cloths", "trays" and "boat". Fine-tuning with real sentences produce images that are more visually-realistic and show higher semantic relevance with language descriptions.

6. Conclusion

In this paper, we proposed to train one text-to-image synthesis model in an unsupervised manner, with no reliance on any pairwise image-text data. To the best of our knowledge, this is the first attempt to tackle such an unsupervised text-to-image synthesis task. We rely on the visual concepts to bridge two sets of images and sentences, and thereby yield pseudo image-text pairs. Afterwards, one generative model is initialized on the constructed pseudo pair data by incorporating our proposed visual concept discrimination loss. Finally, the global semantic consistency loss is further used to refine the pretrained generative model to adapt to the real sentence. Experimental results demonstrate that our proposed unsupervised training method can yield promising results, which even outperforms some text-to-image models trained in the supervised manner.

Declaration of Competing Interest

None.

Acknowlgedgments

This work is supported in part by NSFC under Grant No. 61872215, SZSTI under Grant No. JCYJ20180306174057899, and Shenzhen Nanshan District Ling-Hang Team Project under Grant No. LHTD20170005.

References

[1] W. Li, P. Zhang, L. Zhang, Q. Huang, X. He, S. Lyu, J. Gao, Object-driven text–to-image synthesis via adversarial training, in: CVPR, 2019, pp. 12174–12182.
[2] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, X. He, Attngan: Fine–grained text to image generation with attentional generative adversarial networks, in: CVPR, 2018, pp. 1316–1324.
[3] N. Zhou, J. Fan, Automatic image–text alignment for large-scale web image indexing and retrieval, Pattern Recognit. 48 (1) (2015) 205–219.
[4] Y. Liu, Y. Guo, L. Liu, E.M. Bakker, M.S. Lew, Cyclematch: a cycle-consistent embedding network for image-text matching, Pattern Recognit. 93 (2019) 365–379.
[5] L. Ma, Z. Lu, L. Shang, H. Li, Multimodal convolutional neural networks for matching image and sentence, ICCV, 2015.
[6] H. Wang, Z. Ji, Z. Lin, Y. Pang, X. Li, Stacked squeeze-and-excitation recurrent residual network for visual-semantic matching, Pattern Recognit. (2020) 107359.
[7] J. Wang, W. Jiang, L. Ma, W. Liu, Y. Xu, Bidirectional attentive fusion with context gating for dense video captioning, CVPR, 2018.
[8] B. Wang, L. Ma, W. Zhang, W. Liu, Reconstruction network for video captioning, CVPR, 2018.
[9] X. Xiao, L. Wang, K. Ding, S. Xiang, C. Pan, Dense semantic embedding network for image captioning, Pattern Recognit. 90 (2019) 285–296.
[10] W. Zhang, B. Wang, L. Ma, W. Liu, Reconstruct and represent video contents for captioning via reinforcement learning, in: TPAMI, 10.1109/TPAMI.2019.2920899
[11] A. Dash, J.C.B. Gamboa, S. Ahmed, M. Liwicki, M.Z. Afzal, TAC-GAN - Text conditioned auxiliary classifier generative adversarial network, CoRR abs/1703.06412 (2017).
[12] S. Zhu, S. Fidler, R. Urtasun, D. Lin, C.C. Loy, Be your own prada: Fashion synthesis with structural coherence, in: ICCV, 2017, pp. 1689–1697.
[13] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A.C. Courville, Y. Bengio, Generative adversarial nets, in: NeurIPs, 2014, pp. 2672–2680.
[14] S.E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, in: ICML, 2016, pp. 1060–1069.
[15] L. Gao, D. Chen, Z. Zhao, J. Shao, H.T. Shen, Lightweight dynamic conditional gan with pyramid attention for text-to-image synthesis, Pattern Recognit. (2020) 107384.
[16] H. Zhang, T. Xu, H. Li, Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, in: ICCV, 2017, pp. 5908–5916.
[17] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D.N. Metaxas, Stackgan++: realistic image synthesis with stacked generative adversarial networks, IEEE Trans. PAMI 41 (8) (2019) 1947–1962.
[18] J. Johnson, A. Gupta, L. Fei-Fei, Image generation from scene graphs, in: CVPR, 2018, pp. 1219–1228.
[19] S. Hong, D. Yang, J. Choi, H. Lee, Inferring semantic layout for hierarchical text-to-image synthesis, in: CVPR, 2018, pp. 7986–7994.
[20] L. Gao, D. Chen, J. Song, X. Xu, D. Zhang, H.T. Shen, Perceptual pyramid adversarial networks for text-to-image synthesis, in: AAAI, 2019, pp. 8312–8319.
[21] T. Qiao, J. Zhang, D. Xu, D. Tao, Mirrorgan: Learning text-to-image generation by redescription, in: CVPR, 2019, pp. 1505–1514.
[22] M. Artetxe, G. Labaka, E. Agirre, K. Cho, Unsupervised neural machine translation, ICLR, 2018.
[23] G. Lample, A. Conneau, L. Denoyer, M. Ranzato, Unsupervised machine translation using monolingual corpora only, ICLR, 2018.
[24] Y. Su, K. Fan, N. Bach, C.J. Kuo, F. Huang, Unsupervised multi-modal neural machine translation, in: CVPR, 2019, pp. 10482–10491.
[25] Y. Feng, L. Ma, W. Liu, J. Luo, Unsupervised image captioning, in: CVPR, 2019, pp. 4125–4134.
[26] C. Chen, Z.F. Pan, M. Liu, M. Sun, Unsupervised stylish image description generation via domain layer norm, in: AAAI, 2019, pp. 8151–8158.
[27] M. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, in: NeurIPS, 2017, pp. 700–708.
[28] M. Li, H. Huang, L. Ma, W. Liu, T. Zhang, Y. Jiang, Unsupervised image-to-image translation with stacked cycle-consistent adversarial networks, in: ECCV, 2018, pp. 186–201.
[29] S. Song, W. Zhang, J. Liu, T. Mei, Unsupervised person image generation with semantic parsing transformation, in: CVPR, 2019, pp. 2357–2366.
[30] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, NeurIPS, 2014.
[31] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: NeurIPS, 2015, pp. 91–99.

[32] T. Lin, M. Maire, S.J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: ECCV, 2014, pp. 740–755.
[33] I. Krasin, T. Duerig, N. Alldrin, A. Veit, S. Abu-El-Haija, S. Belongie, D. Cai, Z. Feng, V. Ferrari, V. Gomes, Openimages: A public dataset for large-scale multi-label and multi-class image classification, 2016.
[34] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, ICLR, 2015.
[35] T. Salimans, I.J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, in: NeurIPS, 2016, pp. 2226–2234.
[36] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: CVPR, 2016, pp. 2818–2826.
[37] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, J. Yosinski, Plug & play generative networks: Conditional iterative generation of images in latent space, in: CVPR, 2017, pp. 3510–3520.
[38] Z. Zhang, Y. Xie, L. Yang, Photographic text-to-image synthesis with a hierarchically-nested adversarial network, in: CVPR, 2018, pp. 6199–6208.

Yanlong Dong is a third-year master student with Department of Computer Science in Tsinghua University, Beijing, China. Before that, he received the B.E degrees in Software Engineering from the Northwestern Polytechnical University, Xian, China, in 2016. His research areas involve generative adversarial networks and vision+language, especially, text-to-image synthesis, human image synthesis and image/video captioning.

Ying Zhang received her Ph.D. degree in Signal and Information Processing from Dalian University of Technology (DUT) in 2019, and she obtained her B.E. and M.E. degrees in Electronics and Information Engineering from DUT in 2013 and 2015 respectively. She is currently a senior researcher at Tencent AI Lab, Shenzhen, China. Her research interests include person re-identification, image-text matching and text-to-video retrieval.

Lin Ma received the B.E. and M.E. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2006 and 2008, respectively, and the Ph.D. degree from the Department of Electronic Engineering, The Chinese University of Hong Kong, in 2013. He was a Researcher with the Huawei Noahs Ark Laboratory, Hong Kong, from 2013 to 2016. He was a Principal Researcher with the Tencent AI Laboratory, Shenzhen, China. He is currently a Principal Researcher with Meituan-Dianping Group, China. His current research interests lie in the areas of computer vision, multimodal deep learning, specifically for vision and language, image/video understanding, and quality assessment. Dr. Ma received the Best Paper Award from the Pacific-Rim Conference on Multimedia in 2008. He was a recipient of the Microsoft Research Asia Fellowship in 2011. He was a finalist in HKIS Young Scientist Award in engineering science in 2012.

Zhi Wang is currently an associate professor with Tsinghua Shenzhen International Graduate School, Shenzhen, China. His research areas include multimedia networks, mobile cloud computing, and large-scale machine learning systems. He received the Outstanding Doctoral Dissertation Award from China Computer Federation in 2014, Best Paper Award at ACM Multimedia 2012, and Best Student Paper Award at MMM 2015. He is a recipient of the Second Prize of National Natural Science Award and the First Prize of Natural Science Award of Ministry of Education in 2017. He is an Associate Editor of IEEE TMM and Guest Editor of ACM TIST and JCST.

Jiebo Luo joined the Department of Computer Science, University of Rochester, in 2011, after a prolific career of more than 15 years with Kodak Research. He has authored more than 400 technical papers and holds more than 90 U.S. patents. His research interests include computer vision, machine learning, data mining, social media, and biomedical informatics. He has served as the Program Chair for the ACM Multimedia 2010, the IEEE CVPR 2012, ACM ICMR 2016, and IEEE ICIP 2017, and on the Editorial Boards of the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Multimedia, the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Transactions on Big Data, Pattern Recognition, Machine Vision and Applications, and ACM Transactions on Intelligent Systems and Technology. He is also a fellow of ACM, IEEE, AAAI, SPIE, and IAPR.