Multi-caption Text-to-Face Synthesis: Dataset and Algorithm

Jianxin Sun[1,2], Qi Li[1,2], Weining Wang[1], Jian Zhao[3], Zhenan Sun[1,2*]
[1] Center for Research on Intelligent Perception and Computing, NLPR, CASIA, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing, China
[3] Institute of North Electronic Equipment, Beijing, China
jianxin.sun@cripac.ia.ac.cn,{qli,weining.wang,znsun}@nlpr.ia.ac.cn,zhaojian90@u.nus.edu

ABSTRACT

Text-to-Face synthesis with multiple captions is still an important yet less addressed problem because of the lack of effective algorithms and large-scale datasets. We accordingly propose a Semantic Embedding and Attention (*SEA-T2F*) network that allows multiple captions as input to generate highly semantically related face images. With a novel Sentence Features Injection Module, SEA-T2F can integrate any number of captions into the network. In addition, an attention mechanism named Attention for Multiple Captions is proposed to fuse multiple word features and synthesize fine-grained details. Considering text-to-face generation is an ill-posed problem, we also introduce an attribute loss to guide the network to generate sentence-related attributes. Existing datasets for text-to-face are either too small or roughly generated according to attribute labels, which is not enough to train deep learning based methods to synthesize natural face images. Therefore, we build a large-scale dataset named CelebAText-HQ, in which each image is manually annotated with 10 captions. Extensive experiments demonstrate the effectiveness of our algorithm.

CCS CONCEPTS

• **Computing methodologies** → *Concurrent algorithms*; • **Information systems** → **Multimedia databases**.

KEYWORDS

Dataset; Text-to-face synthesis; Vision and language

ACM Reference Format:
Jianxin Sun, Qi Li, Weining Wang, Jian Zhao, Zhenan Sun. 2021. Multi-caption Text-to-Face Synthesis: Dataset and Algorithm. In *Proceedings of the 29th ACM Int'l Conference on Multimedia (MM '21), Oct. 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475391

1 INTRODUCTION

Text-to-Image (T2I) synthesis is an emerging research topic in multimedia, which requires generating images as real as possible according to natural language descriptions as well as ensuring semantic consistency. In contrast to traditional T2I tasks which generate flowers and birds, Text-to-Face (T2F) is more challenging and rarely investigated because face images have many fine-grained features [6, 20, 41, 42, 47]. It has various potential applications such as human-computer interaction, art generation, and criminal investigation based on the victim's descriptions. Compared with attribute labels, natural languages provide richer contextual information and are much easier to interact with people. Previous methods [17, 38–40, 45] usually generate images with one caption, which is not enough to describe comprehensive content, especially for face images with complex geometric structure and detailed components as shown in Figure 1. Furthermore, due to the subjectivity of language description, there may be a discrepancy between the descriptions of different people for the same face image. Multiple captions can complement each other to provide more complete information. Thus, we propose a novel Semantic Embedding and Attention framework for Text-to-Face (SEA-T2F) synthesis, which uses multiple captions to synthesize fine-grained face images. Since there is no **large-scale manually-annotated** multi-caption dataset for text-to-face synthesis, we build a CelebAText-HQ dataset, which contains 15,010 face images from CelebAMask-HQ [16] annotated with 10 text descriptions. In contrast to previous datasets [7, 22, 26, 33, 34, 37**?**] that automatically generate text descriptions by deep learning methods, we manually annotate each sentence for the face image. Examples of our dataset are shown in Figure 1, Figure 2 and Figure 6.

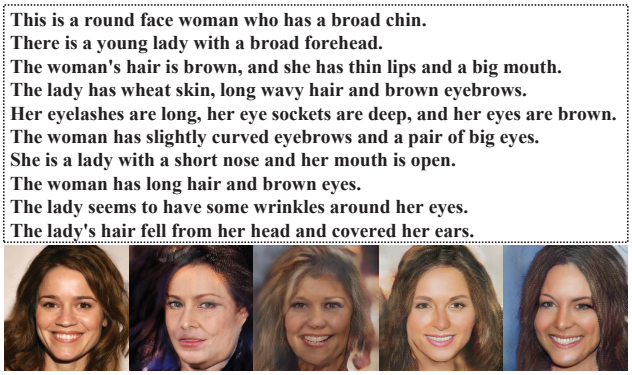


Figure 1: Captions and images. (a) is a real image; (b) is synthesized by AttnGAN [38] with the first caption; (c) is synthesized by ControlGAN [17] with the first caption; (d) is synthesized by our methods with the first caption; (e) is synthesized by our methods with all captions.

*Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China.

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475391

In current T2I methods [4, 17–19, 38–40, 44, 45], noise and sentence vectors are concatenated as the input to a multi-stage framework, which can only deal with a single caption. SEA-T2F also adopts a multi-stage framework where the first stage generates low-resolution images, and the later stages increase the resolution as well as refining the image. Inspired by recent advances in style transfer [10, 13, 14, 46], we propose a Sentence Features Injection Module (SFIM) in the first stage. SFIM takes noise as input and injects the sentence embedding from different captions into different layers of the initial network by adaptive instance normalization. Merely using sentence embedding as input will leave out fine-grained word level information [38]. In recent years, attention mechanism has been widely exploited in the intersection between vision and language modalities [2, 9, 25, 27, 29]. However, it has been rarely explored in the intersection of multiple word features and vision features. Inspired by LTMI [28], we propose a novel attention mechanism named Attention for Multiple Captions (AMC) specifically for aggregating multiple word features and combining them with vision features. Different from natural images like birds and flowers, whose descriptions are mainly related to color, human faces contain many fine-grained components such as mouth, nose and eyes, which are difficult to describe accurately in language. So, additional constraints should be employed for T2F generation. Fortunately, most face datasets contain attribute labels and motivated by recent advances in facial attribute editing [5, 8, 43], we introduce an attribute classifier in the last stage, where the attribute loss is utilized to optimize the network parameters.

In summary, the contributions of this paper are concluded as follows:

- We build a new dataset named CelebAText-HQ. To the best of our knowledge, it is the first large-scale manually annotated dataset for T2F task.
- We introduce a novel framework named SEA-T2F for T2F with multi-captions. The sentence features are injected to the first stage by the proposed SFIM and the word features are combined with the visual features by AMC mechanism. We also design an attribute classifier in the last stage and introduce the attribute loss to generate fine-grained face components.
- We provide a benchmark for T2F task. Experimental results show that SEA-T2F significantly outperforms previous models. Our source codes, pre-trained models and dataset will be released upon acceptance.

2 RELATED WORKS

2.1 Text-to-Image Synthesis

Text-to-image synthesis with one caption. As aforementioned, the goal of text-to-image synthesis with a single caption is to generate realistic images according to a single caption. The first GAN-based T2I work is proposed by Cheng *et al.* [4] which takes the noise and sentence embedding as the input for a one-stage GAN framework to synthesize $64 \times 64 \times 3$ images. Tao *et al.* [32] directly input the noise into a similar framework and inject the sentence features as normalization parameters to synthesize high-quality images. However, these one-stage architectures could only synthesize low-resolution images, and suffer mode collapse when generating high-resolution images. StackGAN [39] and stackGAN++ [40] further design a multi-stage framework, allowing synthesizing images with the resolution of 256×256. AttnGAN [38] introduces attention mechanism to the multi-stage framework to combine each subregion with corresponding word embedding, which generates more fine-grained details. In addition, a Deep Attentional Multimodal Similarity Model (DAMSM) is proposed to evaluate the similarity between text descriptions and images. ControlGAN [17] adds channel-wise attention to disentangle different visual attributes and a word-level discriminator to provide the generator with fine-grained training feedback. Considering the quality of final results is heavily affected by the output of the initial stage, Zhu *et al.* [45] propose a dynamic memory module to refine the fuzzy images.

Text-to-image synthesis with multi-captions. Most of T2I works focus on text-to-image with a single caption, but a single description can only describe a part of an image [23, 30, 35]. Recently, text-to-image synthesis with multi-captions has attracted increasing interest from academia. Joseph *et al.* [11] propose a deep generative model that iteratively updates its generated image features by taking into account different captions at each step. Nevertheless, it is time-consuming due to that the generator needs to be updated many times according to the number of captions. Cheng *et al.* [4] enrich the captions by retrieving compatible captions from prior knowledge automatically, and sends the enriched captions into the Bi-LSTM [24] network to obtain sentence features and word features. A Self-Attentional Embedding Mixture (SAEM) mechanism is proposed to aggregate the augmented sentence features and visual-language attention maps. Unfortunately, the information from multiple sentences is messed up after the SAEM mechanism. To address this issue, we inject the augmented sentence features by SFIM module. As for word features, we introduce an AMC module to fuse multiple word features and visual features.

2.2 Text-to-Face Synthesis

T2F work is rarely studied compared with T2I. There are two possible reasons. 1) Different from bird and flower images whose descriptions are mostly related to color, human faces have more complicated attributes, identity-related features and fine-grained textures. Thus, it is more difficult to bridge the gap between natural language descriptions and human faces. 2) There is no face dataset with accurate text descriptions so far. Most T2F works generate captions by simply converting a list of attributes to captions, thus the descriptions are imprecise and unnatural. As far as we know, the first work on T2F is a project named *T2F* on Github [31], which adopts ProGAN [12] as the generator. The experiments are conducted on Face2text dataset [7], which contains only 400 images. Nasir *et al.* [22] generate face captions based on labels in CelebA [21] and propose a one-stage conditional-GAN called Text2FaceGAN to synthesize $64 \times 64 \times 3$ face images. Both T2F and Text2FaceGAN can only generate low-resolution images that are inconsistent with the description. Cehn *et al.* [3] build the SCU-Text2face dataset and utilize a text encoder, an image decoder and three discriminators to synthesize $256 \times 256 \times 3$ face images, but the generated face images are in poor quality. Recently, StyleGAN [13, 14] has achieved amazing performance in face generation, some methods [26, 37] adopt StyleGAN as backbone to synthesize realistic face images. One

Table 1: The division of 10 captions.

Captions	Face Parts
1, 2	Gender, age and face shape.
3, 4	Facial expressions, mouth, nose and ears.
5, 6	Skin, hair style and color, beard,etc.
7, 8	Features around the eyes.
9, 10	Decorations such as eyeglasses, hats and earrings.

drawback of these methods is that StyleGAN is difficult to control precisely. The reason is that the noise usually dominates the results of StyleGAN, which can not be matched with text descriptions.

Even though T2F tasks have been studied for these years, there are still some unsolved problems. First, all of previous T2F works synthesize the face image based on a single caption which is difficult to describe the complex information. Besides, most works exploit different methods based on a small dataset, which lacks sufficient credibility.

3 DATASET

Our dataset is a subset of the CelebAMask-HQ dataset [16] with 30,000 high-fidelity images. In order to avoid the influence of imbalanced data distribution (e.g., the number of females is significantly more than males), occlusion and pose, we construct CelebAText-HQ by selecting 15,010 frontal images from the CelebAMask-HQ dataset, including 8,959 females and 6,051 males. Each image in CelebAText-HQ is manually annotated with 10 text descriptions, and each description consists of 6~44 words.

3.1 Annotation

Different from the descriptions of birds [35] and flowers [23] which are mainly related to color, human faces are more fine-grained and contain ambiguous geometric characteristics which are hard to describe in natural language. Each image is described by 10 people independently in our dataset. In order to describe a human face as comprehensive as possible and prevent information redundancy, the descriptions are divided into 5 groups according to different parts of the face. For example, the first group captures the face outline such as gender, age and face shape (see Table 1). Each person is assigned to a certain group, and the description should cover the characteristics of the group, as well as the most distinguishing features of the image. Considering that some face images may not contain a certain group of features (e.g., lack of decorations), the annotators of that group are allowed to describe other salient features. In order to ensure the accuracy of the annotations, each text has been checked at least 3 times by different people.

3.2 Comparison with Existing Datasets

Almost all T2F studies have proposed their own datasets, which can be divided into two categories, manually annotated and neural network generated. As shown in Table 2, the former has insufficient data, and the latter is based on attribute labels, which has a gap with human natural language.

Face2text. Face2Text [7] selects 400 images from the Faces in The Wild (FTW) dataset [1], and designs a website to allow participants to describe any number of pictures without offering any financial or other incentives. Finally, 1,400 descriptions are collected from 185 participants. Unfortunately, the dataset is too small for complex models, and each image contains a different number of descriptions, which limits its application.

Table 2: Datasets for text-to-face synthesis.

Datasets	Images	Captions	Manually Annotated
Face2text	400	1,400	True
SCU-Text2face	1,000	5,000	True
Text2FaceGAN	10,000	60,000	False
CelebA-HQ-TD	30,000	300,000	False
Multi-Modal CelebA-HQ	30,000	300,000	False
CelebAText-HQ	**15,010**	**150,100**	**True**

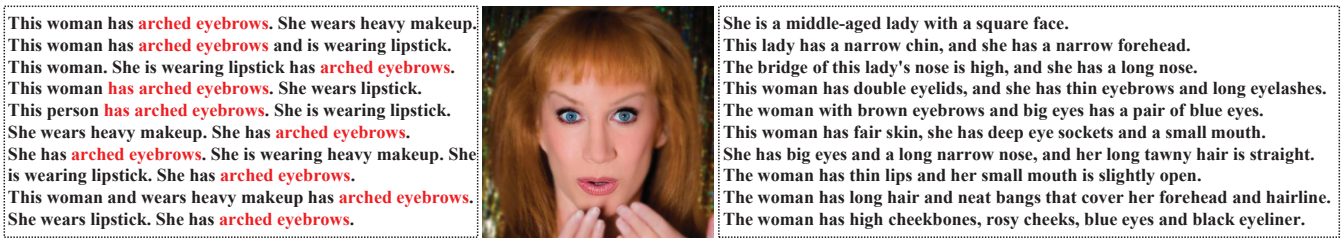
SCU-Text2face. Based on the public face dataset CelebA [21], Chen *et al.* [3] build a dataset named SCU-Text2face, which contains 1,000 images and each image is manually annotated with 5 descriptions. Even though the data size is larger than Face2Text [7], it still can not satisfy the requirements of current deep learning-based models. Furthermore, the 5 descriptions are not constrained, leading to a lot of repeated information.

Text2FaceGAN. Nasir *et al.* [22] create descriptions with the help of the attribute labels in CelebA [21] dataset. The labels are divided into 6 groups, and each group generates a description to avoid redundancy information. However, the number of labels is limited in most images, where some groups only contain one or several labels. Therefore, some descriptions only provide marginal information.

CelebA-HQ-TD and Multi-Modal CelebA-HQ. CelebA-HQ-TD [26] and Multi-Modal CelebA-HQ [37] annotate images in the same way, and both use the attribute labels in CelebAMask-HQ [16] dataset to generate descriptions. The CelebA-HQ-TD [26] is not open-source for confidentiality reasons. To illustrate that our dataset is more suitable for T2F task, we compare the proposed CelebAText-HQ dataset with Multi-Modal CelebA-HQ [37] as shown in Figure 2. We can see that the sentences in Figure 2 (a) describe limited content (e.g., only 4 attributes are described) and duplicate attributes (e.g., "arched eyebrows" appears in each sentence), resulting in information redundancy. Furthermore, salient features such as "blue eyes" and "narrow chin" are not mentioned. In contrast, the manually annotated descriptions in CelebAText-HQ complement each other, with more details and less information redundancy, as well as emphasizing salient features. More importantly, the descriptions in CelebAText-HQ are more natural.

4 ALGORITHM

Given an image x, its corresponding captions can be denoted as $S_1, S_2, ..., S_N$, where N is the number of captions. The goal of our work is to synthesize an image x' by multiple captions. A text description S_i $(i = 1, ..., N)$ is firstly extracted as a sentence feature $s_i \in \mathbb{R}^D$ and a word feature $w_i \in \mathbb{R}^{D \times L_i}$ by a Bi-LSTM [24], where D is the dimension of the sentence feature and word feature, L_i



(a) Multi-Modal CelebA-HQ **(b) CelebAText-HQ**

Figure 2: Comparison of Multi-Modal CelebA-HQ and CelebAText-HQ. (a) The text descriptions of image in Multi-Modal CelebA-HQ. (b) The text descriptions of image in CelebAText-HQ.

is number of words in the i^{th} sentence. The sentence feature s_i is further processed as the augmented sentence feature $s'_i \in \mathbb{R}^D$ by Conditioning Augmentation (CA) [39].

The architecture of SEA-T2F is shown in Figure 3 (a). We adopt a multi-stage generative network as our backbone. Different from previous works [4, 17, 38–40, 45] which concatenate the augmented sentence features with noise as the input, we directly input the noise to synthesize the initial visual feature $h_1 \in \mathbb{R}^{C \times l_1}$, where C denotes the channel of the visual feature, and l_1 denotes the product of the width and height of the visual feature. The sentences from different captions are integrated into different layers of the network through SFIM. The visual feature h_1 and word features $[w_1, w_2, \ldots, w_N]$ are integrated into an attention map $Attn_1$ via the proposed AMC mechanism. Then h_1 and $Attn_1$ are concatenated together as the input of the upsampling network F_2 to obtain the visual feature $h_2 \in \mathbb{R}^{C \times l_2}$, which can be formulated as:

$$h_2 = F_2(h_1, Attn_1) \tag{1}$$

We can also calculate $h_3 \in \mathbb{R}^{C \times l_3}$ in the same way.

4.1 Sentence Features Injection Module

The final synthesized images are influenced heavily by the first stage. To improve the quality of the images in the first stage as well as utilizing multiple captions, we take the noise as input of the first stage directly and leverage SFIM to integrate multiple sentence features to different layers. We design a six layer deep network for the first stage, where the first layer is a fully connected layer and the last five layers are composed of five upsampling and convolutional layers. The network structure of SFIM is shown in Figure 3 (b), where z is the output of the previous layer, z' is the output of SFIM, s'_i and s'_j are different sentence features corresponding to the same image. In order to reduce the complexity and adapt to the number of captions, each SFIM uses two captions as input. In the following, we will describe SFIM in detail.

First, the input z is normalized to zero mean and unit deviation by instance normalization:

$$\bar{z} = \frac{z - \mu(z)}{\sigma(z)} \tag{2}$$

where $\mu(z)$ and $\sigma(z)$ are channel-wise mean and variance of z. In order to integrate the sentence feature, we compute the output activation by denormalizing $\bar{z}$ as follows:

$$y = \alpha_i(s'_i) \odot \bar{z} + \beta_i(s'_i) \tag{3}$$

where $\odot$ represents channel-wise multiplication, $\alpha_i(s'_i)$ and $\beta_i(s'_i)$ are two learned modulation parameters of SFIM convolved from s'_i, which can be implemented using a simple fully connected network, y is the activation after the denormalization process. Since we have two sentences for each SFIM, another sentence feature s'_j is injected into SFIM in a similar way to further improve the efficiency of the injection module. This process can be represented as:

$$\bar{y} = \frac{\mathrm{ReLU}(y) - \mu(\mathrm{ReLU}(y))}{\sigma(\mathrm{ReLU}(y))} \tag{4}$$

$$z' = \alpha_j(s'_j) \odot \bar{y} + \beta_j(s'_j) \tag{5}$$

where $\alpha_j(s'_j)$ and $\beta_j(s'_j)$ are two learned modulation parameters of SFIM convolved from s'_j, z' is the activation after the denormalization process.

4.2 Attention for Multiple Captions

Attention mechanism is widely used to integrate word features with intermediate visual features to synthesize fine-grained details. However, most of these methods fail to handle multiple captions since an attention mechanism can only deal with interactions between two modalities. Concatenating multiple captions will lead to huge consumption of computing resources and have difficulty in catching long-range dependence. To address this problem, we propose an AMC module, as shown in Figure 4, which takes the visual feature from the previous stage and word features from multiple sentences as input. In the k^{th} $(k = 2, 3)$ stage, a word feature is mapped into the same dimension with the visual feature, *i.e.*, $w'_i = Uw_i$ $(i = 1, ..., N)$, where $U \in \mathbb{R}^{C \times D}$ denotes the mapping matrix. In the following, we will refer to the visual features h_1, h_2 as h for brevity, and the product of the width and height of the visual feature h is represented as l. The transposed visual feature is multiplied by the word feature w'_i and then normalized by softmax to get the attention weight:

$$\gamma_i = softmax(h^T w'_i) \tag{6}$$

where $\gamma_i \in \mathbb{R}^{l \times L_i}$, the element of γ_i indicates the relationship between the sub-region of the visual feature and the corresponding word in the i^{th} caption. The sub-attention map FM_i can be obtained by multiplying the transformed word feature and the transpose of the attention weight:

$$FM_i = w'_i(r_i)^T \tag{7}$$

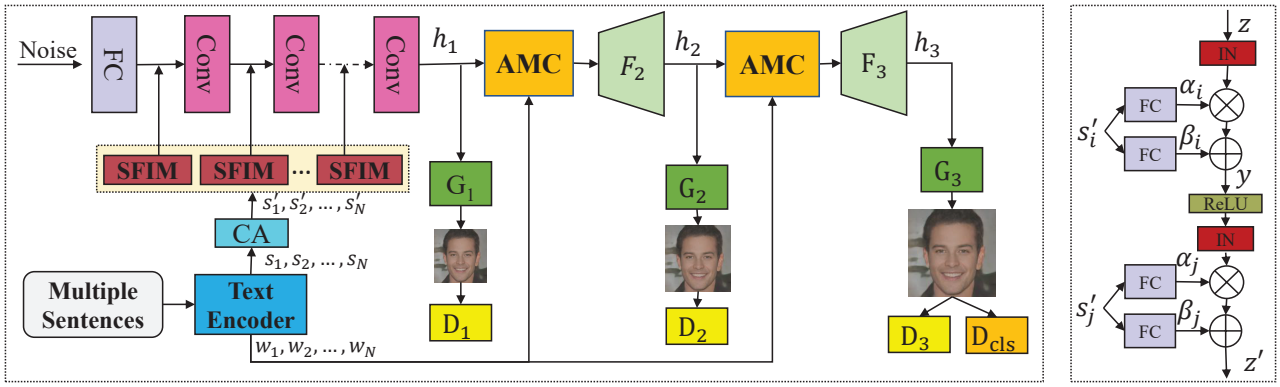


Figure 3: The framework of our method. (a) The overall architecture of SEA-T2F. (b) SFIM module, where IN denotes instance normalization.

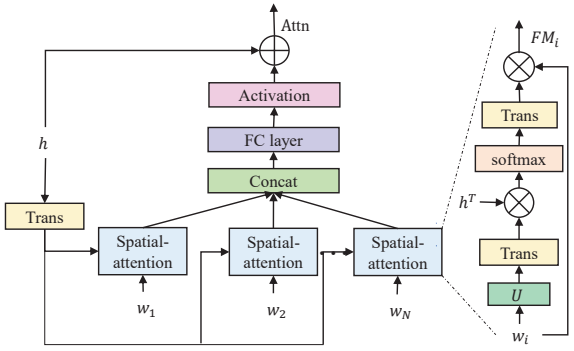


Figure 4: Attention for Multiple Captions (AMC).

To integrate all of the sub-attention together, we concatenate them to get FM_{concat}:

$$FM_{concat} = [FM_1, FM_2, ..., FM_N] \tag{8}$$

where $FM_{concat} \in \mathbb{R}^{C\times(N*l)}$. We apply a linear transformation $W \in \mathbb{R}^{(N*l)\times l}$ and a ReLU activation layer, followed by adding the visual feature h and a LayerNorm activation to get the final Attention map $Attn$:

$$Attn = LayerNorm(ReLU(FM_{concat}W) + h) \tag{9}$$

4.3 Objective Functions

Attribute Loss. Text-to-face synthesis is actually an ill-posed problem, which means that there may be many face images corresponding to the same text description. Hence, we need additional constraints to restrain the synthesize process. ControlGAN [17] introduces perceptual loss to mitigate the uncertainty of the generated images. However, the constraint is not suitable for face images since it will have a negative effect on the diversity of the generated face images. To address this problem, we make use of the available attribute labels and exploit an attribute classifier D_{cls} on top of the discriminator to ensure the attribute consistency of the generated images.

Given a real face image x and its corresponding attribute labels c, the attribute loss is imposed when optimizing both the generator and the discriminator. The attribute loss is decomposed into two terms: an attribute classification loss of real images used to optimize D, and an attribute preservation loss of fake images employed to optimize G. The former can be represented as:

$$\mathcal{L}^D_{cls} = E_{x,c}[-logD_{cls}(c|x)] \tag{10}$$

By minimizing $\mathcal{L}^D_{cls}$, the attribute classifier learns to classify the real image and its corresponding attribute labels. The latter can be formulated as:

$$\mathcal{L}^G_{cls} = E_{x',c}[-logD_{cls}(c|x')] \tag{11}$$

We fixed the classifier $\mathcal{L}^D_{cls}$ to optimize G, which tries to minimize this objective function by generating images that can be classified as real images with corresponding attributes.

Adversarial Loss. The generator loss is defined as follows:

$$\mathcal{L}_G = \frac{\lambda_1}{N}\sum_{k=1}^{3}\sum_{i=1}^{N}\mathcal{L}_{G_{k,i}} + \lambda_2\mathcal{L}^G_{\text{cls}} + \lambda_3\mathcal{L}_{DAMSM} \tag{12}$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters, $\mathcal{L}_{DAMSM}$ is $DAMSM$ loss [38], x'_k is the generated image in the k_{th} stage, and x_k is the real image with the same size with x'_k. $\mathcal{L}_{G_{k,i}}$ is the generator loss in the k_{th} stage with the i^{th} caption:

$$\mathcal{L}_{G_{k,i}} = -\mathbb{E}_{x'_k\sim p_{G_k}}\left[\log\left(D_k\left(x'_k\right)\right)\right] - \mathbb{E}_{x'_k\sim p_{G_k}}\left[\log\left(D_k\left(x'_k, s_i\right)\right)\right] \tag{13}$$

The discriminator loss is computed as:

$$\mathcal{L}_D = -\frac{1}{2N}\sum_{k=1}^{K}\sum_{i=1}^{N}\mathcal{L}_{D_{k,i}} + \mathcal{L}^D_{cls} \tag{14}$$

where $\mathcal{L}_{D_{k,i}}$ is the discriminator loss in the k_{th} stage with the i^{th} caption:

$$\begin{aligned}\mathcal{L}_{D_{k,i}} = & \mathbb{E}_{x_k \sim p_{\text{data}_k}} \left[\log D_k\left(x_k\right)\right] + \mathbb{E}_{x'_k \sim p_{G_k}} \left[\log\left(1 - D_k\left(x'_k\right)\right)\right] \\ & + \mathbb{E}_{x_k \sim p_{\text{data}_k}} \left[\log D_k\left(x_k, s_i\right)\right] \\ & + \mathbb{E}_{x'_k \sim p_{C_k}} \left[\log\left(1 - D_k\left(x'_k, s_i\right)\right)\right]\end{aligned} \tag{15}$$

5 EXPERIMENTS

5.1 Implementation Details

We adopt a 3 stage framework which consists of a fully connected layer, 5 upsampling layers, 5 convolutional layers and 2 upsampling modules. The generator consists of a 3×3 convolutional layer, a batch normalization layer and a gated linear layer. The attribute classifier consists of 6 residual blocks, which is similar to StarGAN [5]. The sentence features and word features are extracted by Bi-LSTM [24], and the image features are extracted by a *CNN* similar to AttnGAN [38]. The hyperparameters λ_1, λ_2, λ_3, are set to 0.5, 0.02, 10. We choose Adam [15] to be the optimizer with learning rate and batch-size setting to 0.0002 and 12 respectively. Since imbalance exists between the discriminator and the generator, the discriminator is updated every 5 iterations.

For the T2F task, we test our method on the proposed CelebAText-HQ and Multi-Modal CelebA-HQ [37] datasets. Table 3 lists the splits of different datasets. We also evaluate our model on the CUB [35] dataset to prove the generality of our model.

Table 3: Splits of datasets.

Dataset	Train	Test
CelebAText-HQ	13,710	1,300
Multi-Modal CelebA-HQ	28,000	2,000
CUB	8,855	2,933

5.2 Comparison with Other Methods

Evaluation Metrics. Existing T2I methods apply Inception Score (IS) and R_precision as quantitative evaluation metrics to measure the reality of generated images and the relevance between generated images and the corresponding captions. However, we find R_precision is not suitable for T2F task, since R_precision is even less than 10% for original images. Instead, we introduce another evaluation metric, top-5 image retrieval accuracy (Top-5 Acc), to evaluate the identity similarity between the synthesized image and the original image. To be specific, we use *LightCNN* [36] to extract the features of both original and generated face images, and find the top five original images with the highest cosine similarity for each generated image. The reason we choose Top-5 instead of Top-1 ACC is that it is hard for the current algorithm to restore the identity information of original image merely based on text, and can only synthesize face images as similar as possible.

5.2.1 Quantitative Evaluation. We compare our method with AttnGAN [38] and ControlGAN [17] which both adopt a multi-stage framework and attention mechanism. We do not compare with previous multi-caption models [4, 11] for T2I tasks because they did not have open source code and training details, which makes it difficult to make a fair comparison. We evaluate them on Multi-Modal CelebA-HQ [37] and the proposed CelebAText-HQ datasets, all results have a resolution of $256 \times 256 \times 3$. Since almost all the existing T2I and T2F methods are based on a single caption, we also synthesize and evaluate the images based on one caption for fair comparison, where each caption is copied 10 times to cope with the input dimension of our model. The quantitative results are shown in Table 4. We can find that, even though our model is not designed for single caption, our method overwhelms the previous methods in both IS and Top-5 Acc in Multi-modal CelebA-HQ [37] dataset. It is observed that the Top-5 Acc of our method on CelebAText-HQ dataset is also better than AttnGAN except the inception score. We argue that CelebAText-HQ dataset is more challenging, and we need to strike a balance between image diversity and identity preservation. We believe that the latter is more important for the T2F task.

Table 4: Inception Scores and Top-5 Acc of different methods evaluate on CelebAText-HQ and Multi-modal CelebA-HQ test set.

Methods	CelebAText-HQ		Multi-modal CelebA-HQ	
	IS	Top-5 Acc	IS	Top-5 Acc
AttnGAN	**2.68±0.12**	7.69%	2.69±0.07	9.85%
ControlGAN	1.91±0.04	7.77%	2.30±0.06	8.15%
SEA-T2F	1.94 ±0.04	**8.54%**	**2.91±0.12**	**10.2%**

5.2.2 Qualitative Results. The visual comparisons for CelebAText-HQ and Multi-modal CelebA-HQ dataset are shown in Figure 5. We find that our method produces results with much better visual quality and fewer artifacts, while the results of AttnGAN [38] and ControlGAN [17] are accompanied by blurry textures and color distortions. Furthermore, our method can not only generate rough information such as gender, skin color, hair style and color, etc. in the sentences, but also generate fine-grained components like eyelids, lip size, mouth open and closing, etc.

We can also see that compared to Multi-modal CelebA-HQ [37], the quality of the face image generated in our CelebAText-HQ is slightly lower. This is because the sentence in the former dataset is generated by attribute labels, which are simpler and easier to learn by the network. In other words, our dataset is more challenging.

5.3 Multiple Captions

SEA-T2F is designed for synthesizing images with arbitrary captions. In order to verify our algorithm, we take the first, the first two, the first five and all ten captions as inputs to synthesize images respectively. Figure 6 shows the visual results generated by SEA-T2F with different captions in CelebAText-HQ dataset. Different from existing T2I datasets [23, 30, 35], ten captions in our dataset focus on different aspects of face images. Therefore, the results generated by fewer captions have a certain degree of randomness and differ from original images. Despite the result generated by ten captions having different identity with the original face image, the attributes and facial components are almost the same as the

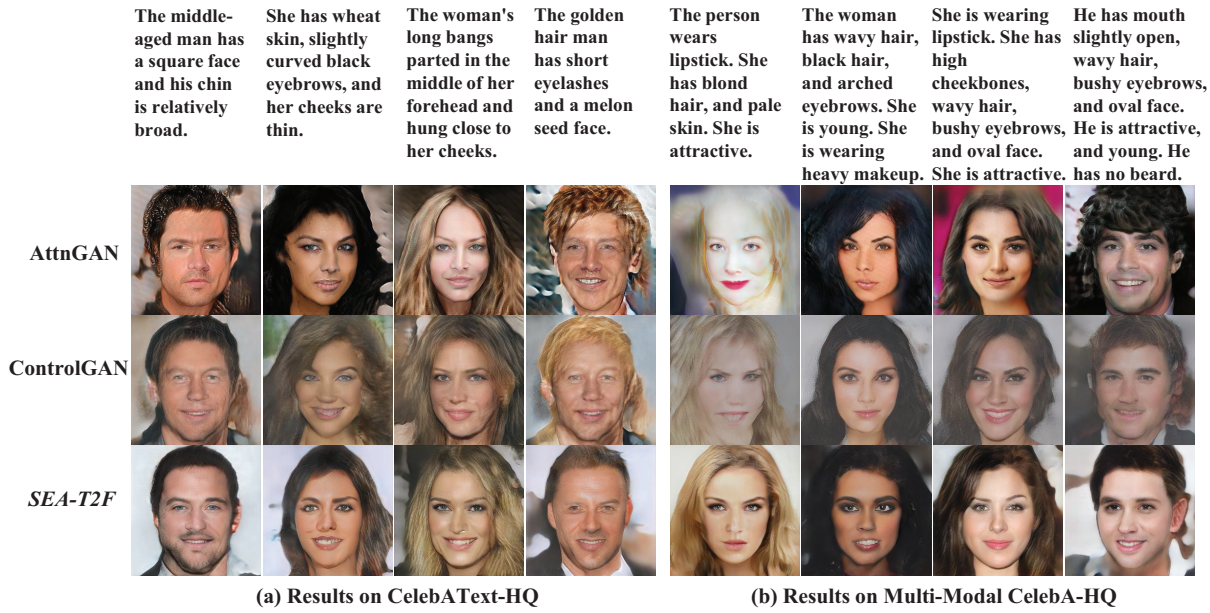


Figure 5: Qualitative results with one caption, (a) denotes the results generated on CelebAText-HQ dataset and (b) denotes the results generated on Multi-Modal CelebA-HQ dataset.

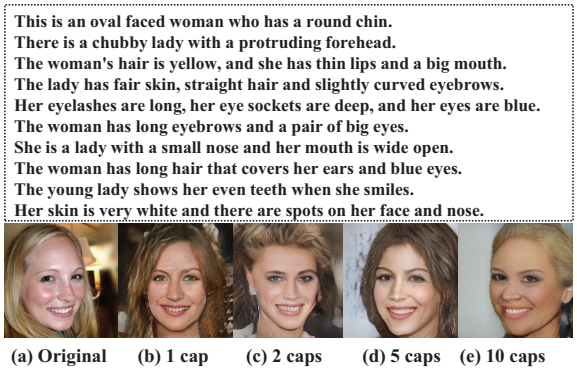


Figure 6: Qualitative results on the CelebAText-HQ dataset with different captions. (a) represents the original image, (b), (c), (d) and (e) denote the synthesized images with the first caption, the first two, the first five and all ten captions respectively.

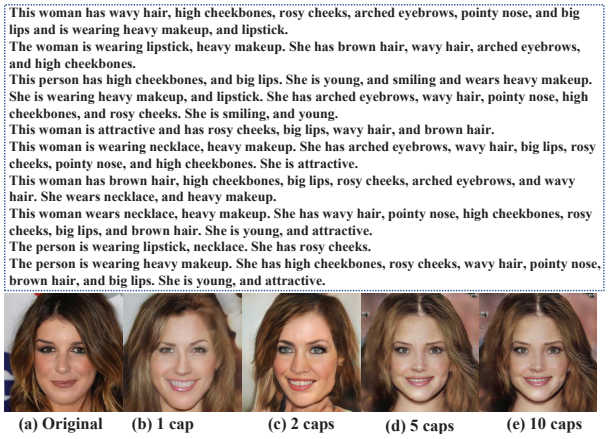


Figure 7: Qualitative results on the Multi-Modal CelebA-HQ dataset with different captions. (a) represents the original image, (b), (c), (d) and (e) denote the synthesized images with the first caption, the first two, the first five and all ten captions respectively.

original image and faithfully preserve the semantic information of the captions.

The visual results with different captions synthesized on Multi-Modal CelebA-HQ [37] are shown in Figure 7. Although as captions increase, the generated images are closer to the original image and the details are more abundant, the improvement is limited. In particular, the images generated with five captions and ten captions are almost the same. This is because all of the captions are synthesized by attribute labels, and the information between captions is highly redundant.

5.4 Ablation Study

As shown in table 5, we conduct an ablation study to show the effectiveness of the proposed modules. Three variants of our method are designed. SEA-T2F$_A$ means to replace the SFIM module of SEA-T2F with an upsampling network similar to the first stage in AttnGAN [38]. SEA-T2F$_B$ means to replace the AMC module with sparse attention and SEA-T2F$_C$ represents SEA-T2F without attribute loss. SEA-T2F$_A$ has the lowest inception score, which means the SFIM module has a great impact on image quality. In contrast,

the inception score is much higher in SEA-T2F$_C$ but it gets the lowest Top-5 Acc value, which proves the effectiveness of attribute loss in acquiring identity information. Relatively, SEA-T2F$_B$ has the smallest performance degradation, but it is still inferior to SEA-T2F. These experiments further verify the effectiveness of the proposed components.

Table 5: Quantitative results of different variants of our method on CelebAText-HQ and Multi-modal CelebA-HQ dataset.

Methods	CelebAText-HQ		Multi-modal CelebA-HQ	
	IS	Top-5 Acc	IS	Top-5 Acc
SEA-T2F$_A$	1.63±0.04	7.77%	2.20±0.08	9.05%
SEA-T2F$_B$	1.70±0.05	7.92%	2.42±0.07	9.45%
SEA-T2F$_C$	1.77±0.05	7.38%	2.50±0.12	8.2%
SEA-T2F	**1.94±0.04**	**8.54%**	**2.91±0.12**	**10.2%**

5.5 Results on CUB Dataset

5.5.1 Quantitative Evaluation. In order to prove that our method can also be applied to T2I tasks, we evaluate it on the CUB dataset [35]. Considering that the bird images have no attribute labels, we remove the attribute classifier and attribute loss in experiments. As shown in Table 6, our method improves the IS from 4.58 to 4.62 and the R_precision from 69.33 to 71.79 on the basis of AttnGAN [38] and ControlGAN [17].

Table 6: Inception Scores and R_precision evaluated on CUB test set.

Methods	IS	R_precision
AttnGAN	4.36 ± .03	67.82
ControlGAN	4.58 ± .09	69.33
SEA-T2F	**4.62±0.19**	**71.79**

5.5.2 Qualitative Results. The qualitative results of CUB dataset with one caption are shown in Figure 8. It is clear that our method achieves the best visual results in terms of both image quality and semantic consistency. To be specific, our method can not only synthesize salient features (e.g., colors and wings) but also generate more clear background and more fine-grained details (e.g., strip and specks).

5.5.3 Multiple Captions. We also synthesize images with different numbers of captions as shown in Figure 9. As the number of captions increases, the generated birds have more prominent textures and contain richer details, which means SEA-T2F learns more information. Further, the image synthesized by 10 captions has the most similar features to the original image.

6 CONCLUSIONS

In this paper, we propose a Semantic Embedding and Attention method for Text-to-Face synthesis with multiple captions named SEA-T2F. Firstly, we build a novel multi-stage framework, which consists of a Sentence Features Injection Module to inject multiple sentence features into the network, and an Attention for Multiple Captions to integrate multiple word features and synthesize fine-grained details. Secondly, we introduce an attribute classifier in the discriminator and propose an attribute loss to ensure attribute consistency. Lastly, we build a new dataset named CelebAText-HQ which each image is annotated with 10 captions. Experimental results have shown the effectiveness of our method.

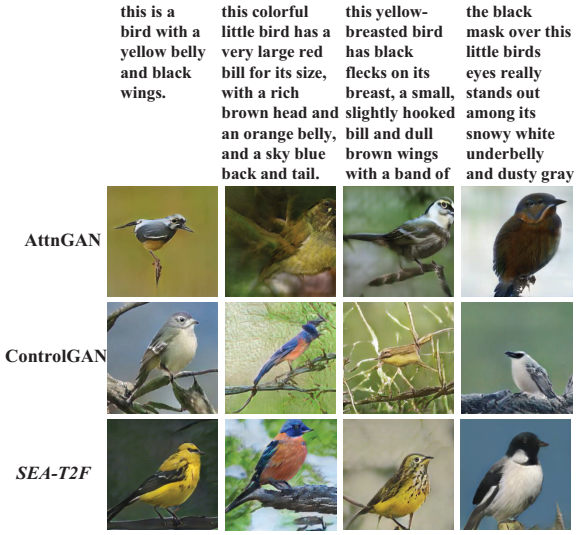


Figure 8: Qualitative results on the CUB dataset.

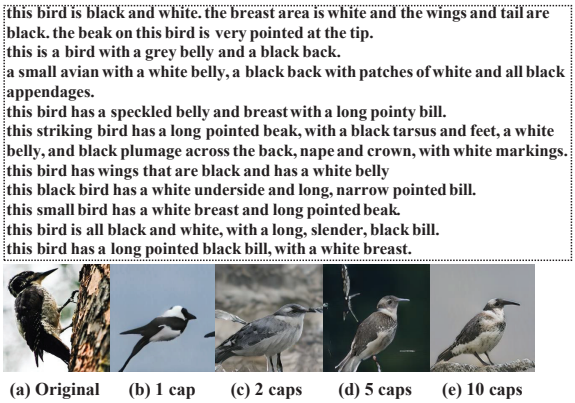


Figure 9: Qualitative results on the CUB dataset with different captions. (a) represents the original image, (b), (c), (d) and (e) denote the synthesized images with the first caption, the first two, the first five and all ten captions respectively.

ACKNOWLEDGMENTS

This work was partially support by the National Key R&D Program of China under Grant 2020AAA0140002, the Natural Science Foundation of China under Grant No. U1836217, Grant No. 62076240, Grant No.62006244, and Grant No. 61721004.

REFERENCES

[1] Tamara L Berg, Alexander C Berg, Jaety Edwards, and David A Forsyth. 2005. Who's in the picture. *Advances in Neural Information Processing Systems (NerIPS)* 17 (2005), 137–144.

[2] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5659–5667.

[3] Xiang Chen, Lingbo Qing, Xiaohai He, Xiaodong Luo, and Yining Xu. 2019. FTGAN: A fully-trained generative adversarial networks for text to face generation. *arXiv preprint arXiv:1904.05729* (2019).

[4] Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, and Dapeng Tao. 2020. RiFeGAN: Rich feature generation for text-to-image synthesis from prior knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 10911–10920.

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8789–8797.

[6] Qiyao Deng, Jie Cao, Yunfan Liu, Zhenhua Chai, Qi Li, and Zhenan Sun. 2020. Reference-guided face component editing. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 502–508.

[7] Albert Gatt, Marc Tanti, Adrian Muscat, Patrizia Paggio, Reuben A. Farrugia, Claudia Borg, Kenneth P. Camilleri, Mike Rosner, and Lonneke van der Plas. 2021. Face2Text: Collecting an Annotated Image Description Corpus for the Generation of Rich Face Descriptions. *arXiv preprint arxiv:1803.03827* (2021).

[8] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. 2019. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing (TIP)* 28, 11 (2019), 5464–5478.

[9] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 4634–4643.

[10] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 1501–1510.

[11] KJ Joseph, Arghya Pal, Sailaja Rajanala, and Vineeth N Balasubramanian. 2019. C4synth: Cross-caption cycle-consistent text-to-image synthesis. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 358–366.

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*.

[13] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4401–4410.

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8110–8119.

[15] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

[16] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. MaskGAN: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[17] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. 2019. Controllable text-to-image generation. In *Advances in Neural Information Processing Systems (NeuIPS)*. 2065–2075.

[18] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. 2020. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7880–7889.

[19] Jiadong Liang, Wenjie Pei, and Feng Lu. 2019. CPGAN: Full-spectrum content-parsing generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:1912.08562* (2019).

[20] Yunfan Liu, Qi Li, and Zhenan Sun. 2019. Attribute-aware face aging with wavelet-based generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 11877–11886.

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 3730–3738.

[22] Osaid Rehman Nasir, Shailesh Kumar Jha, Manraj Singh Grover, Yi Yu, Ajit Kumar, and Rajiv Ratn Shah. 2019. Text2FaceGAN: face generation from fine grained textual descriptions. In *IEEE International Conference on Multimedia Big Data (BigMM)*. 58–67.

[23] Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics & Image Processing*. 722–729.

[24] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing (TSP)* 45, 11 (1997), 2673–2681.

[25] Yale Song and Mohammad Soleymani. 2019. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1979–1988.

[26] David Stap, Maurits Bleeker, Sarah Ibrahimi, and Maartje ter Hoeve. 2020. Conditional image generation and manipulation for user-specified content. *arXiv preprint arXiv:2005.04909* (2020).

[27] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. 2020. Attention is all You need in speech separation. *arXiv preprint arXiv:2010.13154* (2020).

[28] Masanori Suganuma, Takayuki Okatani, et al. 2020. Efficient attention mechanism for visual dialog that can handle all the interactions between multiple inputs. In *European Conference on Computer Vision (ECCV)*. 223–240.

[29] Xiaoshuai Sun, Xuying Zhang, Liujuan Cao, Yongjian Wu, Feiyue Huang, and Rongrong Ji. 2020. Exploring language prior for mode-sensitive visual attention modeling. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. 4199–4207.

[30] S. Belongie J. Hays P. Perona D. Ramanan P. Dollr T.-Y. Lin, M. Maire and C. L. Zitnick. 2014. Microsoft coco: Common objects in context. (2014), 740–755.

[31] T2F. [n.d.]. https://github.com/akanimax/T2F. Accessed: 2021-07-22.

[32] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Fei Wu, and Xiao-Yuan Jing. 2020. DF-GAN: Deep fusion generative adversarial networks for Text-to-Image synthesis. *arXiv preprint arXiv:2008.05865* (2020).

[33] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky. 2016. Texture networks: Feed-forward synthesis of textures and stylized images. In *International Conference on Machine Learning (ICML)*, Vol. 1. 4.

[34] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* (2016).

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

[36] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. 2018. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security (TIFS)* 13, 11 (2018), 2884–2896.

[37] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. TediGAN: Text-guided diverse image generation and manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2256–2265.

[38] 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 (CVPR)*. 1316–1324.

[39] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 5907–5915.

[40] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2018. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 41, 8 (2018), 1947–1962.

[41] Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Shengmei Shen, Junliang Xing, et al. 2018. Towards pose invariant face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2207–2216.

[42] Jian Zhao, Lin Xiong, Yu Cheng, Yi Cheng, Jianshu Li, Li Zhou, Yan Xu, Jayashree Karlekar, Sugiri Pranata, Shengmei Shen, et al. 2018. 3D-Aided deep pose-invariant face recognition. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 1184–1190.

[43] Jian Zhao, Lin Xiong, Jayashree Karlekar, Jianshu Li, Fang Zhao, Zhecan Wang, Sugiri Pranata, Shengmei Shen, Shuicheng Yan, and Jiashi Feng. 2017. Dual-Agent gans for photorealistic and identity preserving profile face synthesis. In *Advances in Neural Information Processing Systems (NeuIPS)*, Vol. 30.

[44] Bin Zhu and Chong-Wah Ngo. 2020. CookGAN: Causality based text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5519–5527.

[45] 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 (CVPR)*. 5802–5810.

[46] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5104–5113.

[47] Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. 2021. One shot face swapping on megapixels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4834–4844.