2023 IEEE International Conference on Image Processing (ICIP) | 978-1-7281-9835-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIP49359.2023.10222631

TELL YOUR STORY: TEXT-DRIVEN FACE VIDEO SYNTHESIS WITH HIGH DIVERSITY VIA ADVERSARIAL LEARNING

Xia Hou[1]*, Meng Sun*[1]*, Wenfeng Song*[1]***

[1]Computer School, Beijing Information Science and Technology University

ABSTRACT

Face synthesis is a rapidly growing area of research in computer vision. Text-driven face synthesis is particularly flexible, but challenges still exist in fusing the semantics of text and images, as well as generating diverse faces. To address these challenges, we propose a cross-modality adversarial learning framework to generate highly diverse face videos that correspond to given text descriptions. We encode text and images into a common latent space and align text and image features to control the synthesis of face attributes. We have designed a novel auto-encoder with a face identity discriminator that enlarges the margin between different individuals, increasing the variety of created faces while maintaining the semantic coherence of text and images. Our proposed method has been successfully tested on the recently released Multimodal VoxCeleb dataset. Our code is public available at https://github.com/sunmeng7/TYS.git.

***Index Terms*—** Face synthesis, generative adversarial networks, face videos, text-to-face, diversity faces

1. INTRODUCTION

Photo-realistic image synthesis is an important field in computer vision that involves creating new images or manipulating existing ones to achieve a desired result. One important application of this technology is generating facial images based on given text inputs. While many existing methods for image synthesis, such as FA-GAN, EruditeGAN, and StyleGAN [1, 2, 3, 4, 5], rely on generative adversarial networks (GANs), they lack the ability to generate facial images from text input.

Our paper focuses on a high-quality text-driven face synthesis (T2F) [6, 7, 8, 9, 10, 11] task, a new branch of image synthesis, to generate related faces given a descriptive text [12, 13, 14, 15]. Currently, there are two main approaches to improve the performance of T2F synthesis: (1) using pretrained language models such as BERT, RoBERTa [16], etc., as the text encoder and StyleGAN2 [4] as the image generator; (2) using CLIP [17], which can process both text and images, as encoders, with StyleGAN [3] enhancing the text description effect and image generation quality. However, StyleGAN-generated images may contain artifacts. T2F is a challenging task due to the following problems: (1) the cross-modality control of text to generate facial images; (2) to increase the diversity of the generated facial images.

Fig. 1. Our method can generate diverse face videos from a face-descriptive text. The first frame of each video is used as an example to demonstrate the effect in the figure above.

To address the above problems, we propose a cross-modal adversarial learning framework (Fig.2) that generates highly correlated face videos based on a given text description. We simplify this framework as Tell Your Story (TYS) in the following content. We create an image-level discriminator to expand the divide between various persons and improve face diversity. The generated results can be seen in Fig.1. We propose the following contributions:

- We propose a cross-modality adversarial learning framework (Fig. 2 A.) for text-driven facial video synthesis and reveal the relationships between cross-modality text signals and the informative facial attributes' appearance in latent feature space.
- We design a text-conditioned discriminator (Fig. 2 B.) to supervise the diversity and fidelity of facial videos under the condition of detailed descriptive attributes. Our model can produce a variety of faces with high-

This paper is supported by National Natural Science Foundation of China (62102036), Beijing Natural Science Foundation (4222024), RD Program of Beijing Municipal Education Commission (KM202211232003), Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB2022A02).

Corresponding author, Wenfeng Song, songwenfenga@163.com.

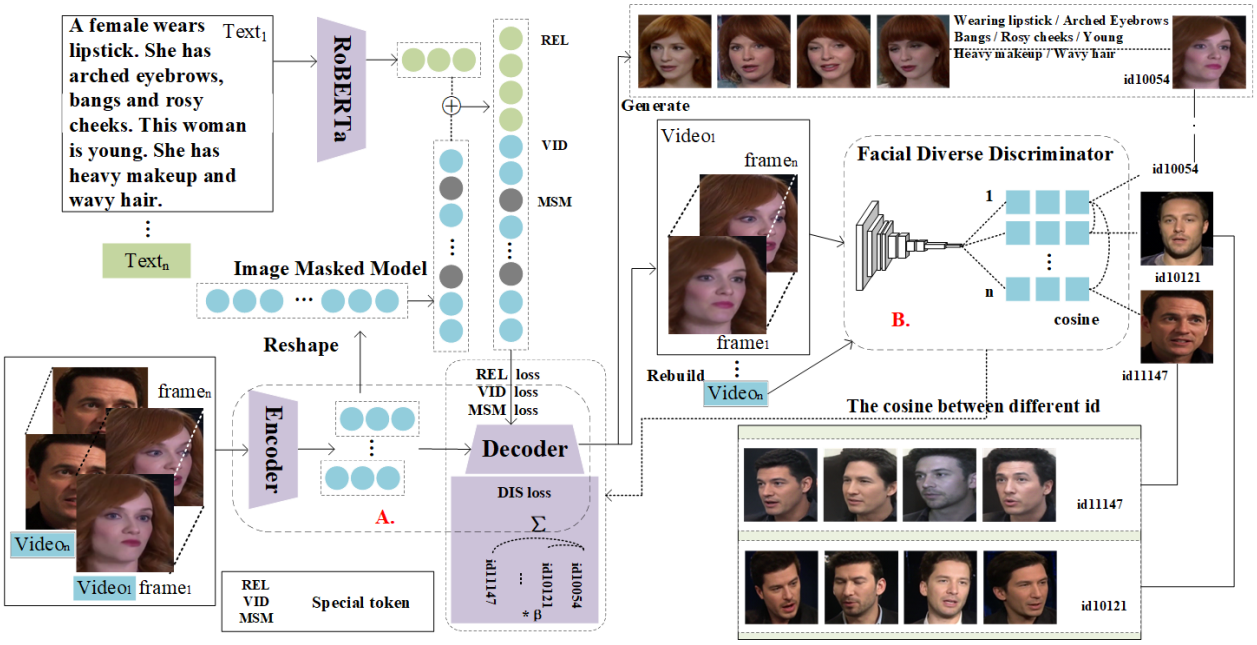

Fig. 2. The overview of our method. The text embedding is obtained by RoBERTa and the image embedding is obtained by auto-encoder. Our discriminator (B.) consists of a convolutional neural network trained using reconstructed face images. Loss generated by discriminator discriminating images in collaboration with special REL, VID, MSM loss optimization generator, constituting adversarial learning (A.), decode text-controlled sequence features and masked image sequence features to generate diverse faces.

quality, simultaneously diverse, and match the semantics of text descriptions.

- We have conducted extensive experiments to verify the quality and diversity of our model synthetic face videos, and our code is pubic available.

2. METHOD

2.1. Overview

In order to generate T2F diversity videos, we first analyze the text using the pre-training model RoBERTa to obtain text word embedding features for text control. Before the text embedding, the [REL] token is introduced as a flag for multi-modal control. We use an encoder-decoder network with the same network structure as VQGAN [18] to obtain a quantized representation of a video and reconstruct it, introduce [VID] token before the image embedding, and reshape the representation of multiple single-frame images, and random masking of representation of multi-frame videos by the masked sequence modeling (MSM) task. We introduce an image-level discriminator made up of VggFace [19] to broaden the diversity of the generated faces. The reconstructed single-frame face image R_1 is input to the discriminator to recognize the face features, and a new representation r_1 of the face embedding is obtained. We aim to increase the gap between similar features faces. Therefore we use cosine similarity to calculate the similarity between the same set of face feature vectors and optimize the encoder-decoder architecture by a new loss function.

2.2. Cross-modality Adversarial Learning

In this section, we revisit the Generative adversarial network (GAN). The GAN framework trains two models simultaneously: the discriminative model D and the generative model G. The discriminative model estimates the likelihood that the samples come from the training data rather than G.

$$\min_G \max_D V(D,G) = E_{i\sim p_{\text{data}}(i)}[\log D(i)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))] \quad (1)$$

Where G(z; g) represents the mapping of the data space, and G is a differential function. D(i) represents the probability that i came from the data rather than p_g.

The inputs are the original data i and a random noise signal z. The output is a probability value or a scalar value. $V(D,G)$ is the value function of the discriminant network D and the generating network G, min represents the minimum value of the generating function G, max represents the maximum value of the discriminant function $D(i)$, E represents the expectation value about the specified distribution in the subscript, i is the real data, which conforms to the $p_{data}(i)$ distribution, z is the noisy data, which conforms to the p_z(z) distribution, log$D(i)$ denotes the logarithm of $D(i)$, $\log(1-D(G(z)))$ denotes the logarithm of $1-D(G(z))$.

We design our TYS based on the GAN pipeline, the generated images are denoted as $g(I)$. To enhance the recon-

struction effect of images, we employ the MSM masks image tokens instead of masking text tokens. The L_{msm} about masked sequence modeling task can be obtained by the following equation,

$$L_{msm} = -\frac{1}{|M|}\sum_{i\in M}\log P\left(z^{(i)} \mid \mathbf{z}^m(g(I)), \mathbf{c}\right) \quad (2)$$

where M is the masking indices, z^m is the masks sequence, and $\mathbf{c}$ is the text control sequence.

2.3. Text-Conditioned Facial Diverse Discriminator

In order to rebuild and regenerate the original images, we put the input images into an encoder-decoder network with the same network topology as VQGAN. The discriminator receives the first frame of the reconstructed set of images and uses it to determine the embedding representation of the face image, mostly using VggFace for face recognition. The separation between these embeddings is determined by the cosine similarity.

In our experiments, we discover that when the cosine value is higher than 0.88, the face images are too similar to each other so that the faces are more homogeneous at this time. Therefore, we assign a weight value of 4 to a cosine value greater than 0.88. The similarity between facial images is too small when the cosine value is less than 0.54. A weight value of 3 is given for a cosine value smaller than 0.54 in order to prevent the semantic coherence of image separation from the text description. All the other cases are assigned a weight value of 0.5. The L_{dis} about discriminator can be obtained by the following equation,

$$L_{dis} = \sum F(g(I(i,j))) * \beta \quad (3)$$

where the cosine values are stored in a similarity matrix represented by F, i and j is the matrix row and column index, β is the weight, and the candidate values are 4, 3, 0.5.

2.4. Text-Image Relevance Estimation

In order to learn the semantic consistency between the input text description and the target video sequence, a special token REL [20] is added before the text embedding. The L_{rel} about relevance estimation task can be obtained by the following equation,

$$L_{rel} = -\log P\left(1 \mid \mathbf{z}^m(g(I)), \mathbf{c}\right) - \log P\left(0 \mid \mathbf{z}^m(g(I)), \overline{\mathbf{c}}\right) \quad (4)$$

where $\overline{\mathbf{c}}$ is the negative text control sequence.

2.5. Adversarial Training Losses

The model picks out positive and negative sequences and learns to produce time-consistent videos by adding a specific token VID before the target image embedding. The L_{vid} about video consistency estimation task can be obtained by the following equation.

$$L_{vid} = -\log P\left(1 \mid \mathbf{z}^m(g(I)), \mathbf{c}\right) - \log P\left(0 \mid \overline{\mathbf{z}}^m, \mathbf{c}\right) \quad (5)$$

To integrate all of the former losses, we formulate the final loss L as $L = L_{dis} + \beta_{msm} * L_{msm} + \beta_{rel} * L_{rel} + \beta_{vid} + L_{vid}$, the β_* is expressed as the weights of different losses.

Fig. 3. Each set of data from left to right is the description of the attributes, the TYS result, and the MMVID result. The text description add(red font) or replace(blue font) one or two facial characteristics.

3. EXPERIMENTS

3.1. Settings

Dataset. We use the dataset of Multimodal VoxCeleb, which is taken from VoxCeleb. Multimodal VoxCeleb contains a total of 19,522 videos and 3,437 interview scenarios with 453 persons.

Evaluation Metrics. We use FVD [21] and PRD [22] as evaluation metrics. To evaluate the fidelity of our generated images, we follow the work in Section 2.3 and use cosine similarity. To compare with the T2F image comparison method StyleT2I [10], we use FID [23] to measure the quality of the first frame of our generated videos.

3.2. Comparison with State-of-the-art Methods

3.2.1. Qualitative Research

Attributes Quality. We start with the initial face description text, "The male has black and wavy hair," and gradually add or replace one or two face description attributes with the statement, in Fig.3. It is clear that face images generated by our method (left) continue changing as more facial attributes are added or modified, while the face images created by using MMVID (right) do not significantly change. In the figure, properties added are shown in red, and properties modified are shown in blue. In Fig.4, StyleT2I has errors in generating attributes such as wearing a hat and hair color. In addition, the StyleT2I faces have unnatural colors. The faces generated by TYS are more accurate and realistic than StyleT2I.

Diversity Quality. The diversity of generated faces can be increased using TYS in addition to making it more sensitive

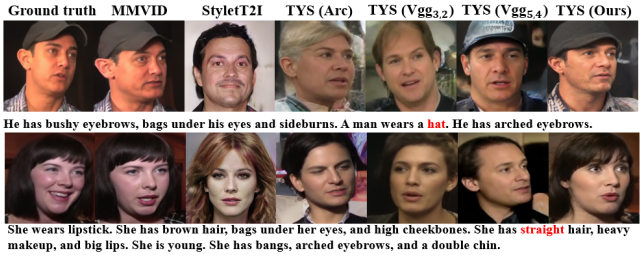

Fig. 4. Compared with TYS, MMVID lacks diversity and is consistent with ground truth; StyleT2I and others are less sensitive in some attributes. TYS is consistent with the text and different from the ground truth.

Fig. 5. In contrast to MMVID, TYS can generate a diversity of faces that differ from the ground truth.

to specific facial characteristics. TYS can generate highly-diverse face images while keeping the semantic properties of the original images, in contrast to MMVID, which real face images that overlap with the genuine images in the dataset, as shown by the face generation effect (Fig.5).

MMVID can also generate faces that do not exist in the dataset, but the face images they generate are too homogeneous and lack diversity, while TYS can generate at least four different sets of faces while ensuring semantic consistency with the text description (Fig. 6).

3.2.2. Quantitative Research

We train our model on NVIDIA GeForce RTX 3090 with batch size set to 6 and iterations set to 20, 0000, and the results of the evaluation metrics are shown in the Table 1. The FVD value indicates the overall quality of the generated videos; the $F_{\frac{1}{8}}$ value indicates the images quality, they have a great improvement, and the F_8 value indicates the diversity. In Fig.7, graphs of the same color and shape represent features of the same description text generated images. We want the same

Fig. 6. TYS is more diverse than MMVID.

Table 1. Quantitative comparison of different methods on Multimodal VoxCeleb dataset. MMVID equals TYS w/o Discriminator.

Method	FVD↓	F_8 ↑	$F_{\frac{1}{8}}$ ↑	CS↑	FID↓
MMVID(CVPR2022)	168.793	0.894	0.672	0.5547	43.108
StyleT2I(CVPR2022)	-	-	-	-	574.023
TYS (ArcFace)	149.198	0.904	0.749	0.6189	47.021
TYS ($VggFace_{3,2}$)	149.900	0.934	0.769	0.6087	38.527
TYS ($VggFace_{5,4}$)	145.965	0.919	0.761	0.6159	43.886
TYS (Ours)	**62.682**	**0.958**	**0.952**	**0.7739**	**24.038**

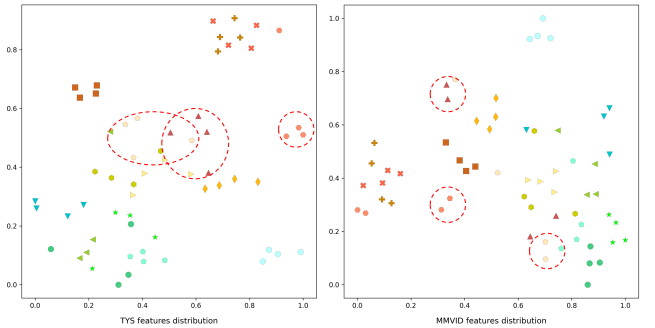

Fig. 7. TYS features distribution graph (Left) and MMVID features distribution graph (Right). In the red circle, TYS can generate three to four semantically similar images for the same text description, but MMVID can only generate at least two, compared to TYS, which has better diversity and fidelity. The feature distribution of TYS is better than that of MMVID.

feature to be aggregated to ensure semantic consistency while being able to potentially disperse to increase diversity. From the figure, TYS is better at maintaining semantic features and generating more images.

3.3. Ablation Study

To prove the effectiveness of our facial diverse discriminator, we also conduct ablation experiments. Specifically, we compare the generation effect of replacing VggFace with ArcFace (TYS (Arc) in Fig.4), and following our work in Section 2.3, we compare the different weights in VggFace (TYS ($Vgg_{3,2}$) and TYS ($Vgg_{5,4}$) in Fig.4). The quantitative results are in Table 1. TYS has excellent results under both figure and table comparisons, and the other comparison methods have errors that cannot be ignored, such as gender and hat attributes.

4. CONCLUSION

In this paper, we propose a model called TYS that uses a face recognition discriminator to produce multiple face images with semantic coherence. According to qualitative and quantitative experimental evaluation, our generated videos have better video quality, semantic consistency between text and videos, and visual diversity. In the future, we will investigate the generation of emotive faces and the decoupling of face attributes to generate detail-rich faces.

5. REFERENCES

[1] Eunyeong Jeon, Kunhee Kim, and Daijin Kim, "Fa-gan: Feature-aware gan for text to image synthesis," in *ICIP*, 2021, pp. 2443–2447.

[2] Zhiqiang Zhang, Wenxin Yu, Ning Jiang, and Jinjia Zhou, "Text to image synthesis with erudite generative adversarial networks," in *ICIP*, 2021, pp. 2438–2442.

[3] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019, pp. 4401–4410.

[4] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of stylegan," in *CVPR*, 2020, pp. 8110–8119.

[5] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Alias-free generative adversarial networks," *NIPS*, vol. 34, pp. 852–863, 2021.

[6] Yikun Wang, Liang Chang, Yuhua Cheng, Lihua Jin, Zhengxin Cheng, Xiaoming Deng, and Fuqing Duan, "Text2sketch: Learning face sketch from facial attribute text," in *ICIP*, 2018, pp. 669–673.

[7] Danlan Huang, Xiaoming Tao, Jianhua Lu, and Minh N Do, "Geometry-aware gan for face attribute transfer," in *ICIP*, 2019, pp. 729–733.

[8] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu, "Tedigan: Text-guided diverse face image generation and manipulation," in *CVPR*, 2021, pp. 2256–2265.

[9] Jianxin Sun, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, and Zhenan Sun, "Anyface: Free-style text-to-face synthesis and manipulation," in *CVPR*, 2022, pp. 18687–18696.

[10] Zhiheng Li, Martin Renqiang Min, Kai Li, and Chenliang Xu, "Stylet2i: Toward compositional and high-fidelity text-to-image synthesis," in *CVPR*, 2022, pp. 18197–18207.

[11] Qiyu Wei, Xulei Yang, Tong Sang, Huijiao Wang, Zou Xiaofeng, Cheng Zhongyao, Zhao Ziyuan, and Zeng Zeng, "Latent vector prototypes guided conditional face synthesis," in *ICIP*, 2022, pp. 3898–3902.

[12] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao, "Semantics disentangling for text-to-image generation," in *CVPR*, 2019, pp. 2327–2336.

[13] Miriam Cha, Youngjune L Gwon, and HT Kung, "Adversarial learning of semantic relevance in text to image synthesis," in *AAAI*, 2019, pp. 3272–3279.

[14] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao, "Mirrorgan: Learning text-to-image generation by redescription," in *CVPR*, 2019, pp. 1505–1514.

[15] Jun Cheng, Fuxiang Wu, Yanling Tian, Lei Wang, and Dapeng Tao, "Rifegan: Rich feature generation for text-to-image synthesis from prior knowledge," in *CVPR*, 2020, pp. 10911–10920.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *PMLR*, 2021, pp. 8748–8763.

[18] Patrick Esser, Robin Rombach, and Bjorn Ommer, "Taming transformers for high-resolution image synthesis," in *CVPR*, 2021, pp. 12873–12883.

[19] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," 2015.

[20] Zhu Zhang, Jianxin Ma, Chang Zhou, Rui Men, Zhikang Li, Ming Ding, Jie Tang, Jingren Zhou, and Hongxia Yang, "Ufc-bert: Unifying multi-modal controls for conditional image synthesis," *NIPS*, vol. 34, pp. 27196–27208, 2021.

[21] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv preprint arXiv:1812.01717*, 2018.

[22] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly, "Assessing generative models via precision and recall," *arXiv preprint arXiv:1806.00035*, 2018.

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *NIPS*, vol. 30, 2017.