IDEAL-NeRF: Cross-Identity Expression-Audio Aligned NeRF for High-Fidelity Talking Head Generation

Gan Kaiyuan, Wang Kai, Peng Xiangyu

National University of Singapore

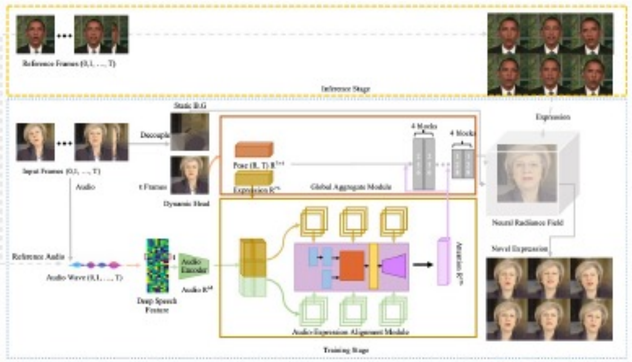
Abstract

Generating high-fidelity talking head with free pose and rich emotion has received considerable attention in recent years. However, synthesizing high consistency between audio emotion and facial expression has not been exploited yet. To tackle this issue, we propose crossIDentity Expression-Audio aLigned Neural Radiance Field (IDEAL-NeRF) to unify facial expression and audio emotion in talking head. Specifically, we design a ExpressionAudio Alignment Module (EAA-Module) based on selfattention mechanism in order to bridge domain gap between audio and visual modality. Given the expressionaudio aligned representation of EAA-Module, we build a Global Aggregation Module (GA-Module) based on neural radiance field to learn expression-audio aligned talking pattern for a specific identity. In addition, the IDEALNeRF empowers cross-identity video dubbing by providing expression-audio aligned talking head. Qualitative and quantitative results both demonstrate that our model is capable to generate high-fidelity talking head with more consistency between audio emotion and facial expression. We decrease 58% in AU existing error, 13% in pose angle error and 8% in landmark distance comparing to state-of-the-art methods. We highly recommend to view the supplementary videos for better visual understanding.

Motivation

Talking head generation [1, 7, 12, 17, 49, 59, 61] has been a long-standing important task both in the academia and the industry due to its wide applications including digital human animation [33], video dubbing of movie [13], virtual robot assistant [45] and video conference [15]. However, because of the complexity of human face and domain gap between audio and vision, this area remains challenging in lip-sync movement, head pose control and talking style restoration.

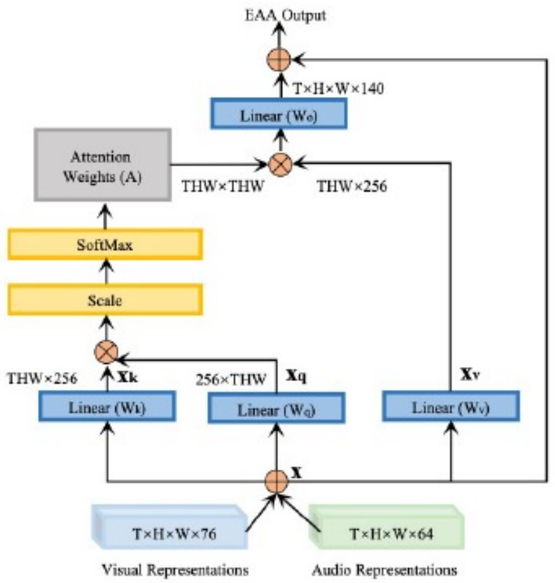
between audio emotion and facial expression.



- We propose a novel framework IDEAEL-NeRF to generate high-fidelity talking head. To our best knowledge, this is the first work to achieve high-consistency between audio emotion and facial expression in talking head generation.
- We propose the EAA-Module and the GA-Module. The EAA-Module learns the attention weights for multi-modal feature fusion and synthesizes expression-audio aligned representation. The GAModule seeks to aggregate original visual signals with the expression-audio aligned representations given by EAA-Module.
- On the strength of our proposed IDAEA-NeRF, we empower talking head generation in cross identity talking style transfer, lip movement synchronization and most importantly, audio expression alignment. Both qualitative and quantitative results demonstrate that our method can generate impressive talking head with high consistency between audio emotion and expression.

Methods

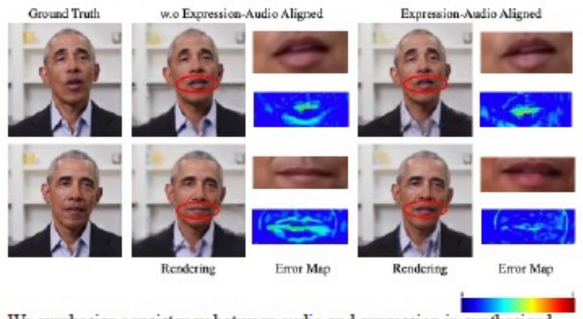
We propose IDEAL-NeRF to generate talking heads with high-consistency between audio emotion and facial expression. Our approach relies on two assumptions: (1) The background, camera pose and location must remain static inside each original videos. (2) There must be only one talking person within single video. As shown in Fig. 2, our approach consists of three modules: 1) Expression and Audio Representation Module, 2) Expression-Audio Alignment Module (EAA-Module), 3) Global Aggregate Module (GA-Module) . In the first module Sec. 3.2, we construct embedding feature spaces for audio, facial expression based on input frames and related audio. In the second module Sec. 3.3, the EAA-Module synthesizes a set of expression-audio aligned representations through audio and visual modalities. In the third module Sec. 3.4, we leverage neural radiance field to generate talking heads based on the expression-audio aligned representations. During the inference, the IDEAL-NeRF only takes audio and facial expression to synthesize high-fidelity talking heads which can perform audio-aligned facial expression of another identity.



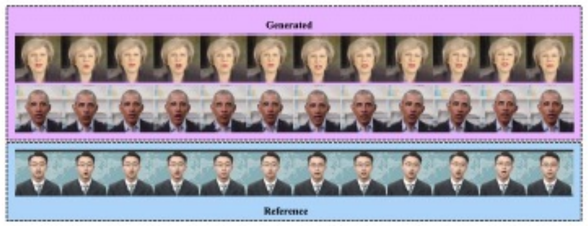
In Style-Avatar [53], the approach takes an expression $\delta \in R\ 76$ along with the transformation matrix $P \in R\ 3\times4$ as style-code which represents talking style for monocular targets in single video. Then it utilizes ResNet-based [21] Module to aggregate audio and the style-code which only process local neighbourhood audio and style-code in feature space. Considering audio and expression show high randomness when talking, this method is constrained to capture the connection between audio and expression at local position thus resulting in misalignment between two modalities. To address this problem, we propose EAA-Module to capture holistic dependencies by computing affinity of two modalities. With such dependencies, the two modalities comprehend each other in a global manner where each position in the feature space knows the importance of all the positions in the other feature space. Therefore, the response of EAA-Module can be incorporated into generation process as an expression-audio aligned condition.

Experiments

Dataset. The public datasets e.g. TCD-TIMIT [20], CREMA-D [5] and LRW [11] contain video clips of hundreds of different targets each of which is too short for our framework to learn a specific talking pattern. Moreover, the camera pose and background do not remain static in these dataset which require previous methods [8, 9, 49, 60, 61] to crop and align target faces before the training process. To address the problems, we collect our training video with static background and camera pose from YouTube in public domain. Each video is cut into 3-5 minutes including more than 5000 frames. Specifically, in our experiment, we split each video frames into 90% for training and 10% for validation, while during inference stage, we use videos of different target to demonstrate our ability of cross-identity talking head generation.



We emphasize consistency between audio and expression in synthesized talking head by incorporating EAAModule and we conduct an ablasive experiment by naive concatenation of audio and expression. As aforementioned in Sec. 3.3, to achieves expression-audio consistency, EAAModule adaptively fuses audio and expression by the affinity of two modalities. We compare the rendering head based on EAA-Module and naive concatenation of audio and expression, where we simply stack the audio and expression to NeRF. In Fig. 5. We visualize the photo-metric error map of rendering head and ground truth which demonstrates that the EAA-Module plays a vital role in generating expression-audio consistent talking head.



We compare our approach with two main-stream talking head generation methods: image-based methods and model based methods. We propose both qualitative and quantitative experiments to evaluate the results of these approaches. For quantitative evaluation, we employ photo-metric error (L1 distance), Peak Signal-to-Noise Ratio (PSNR), Structure Similarity Index Measure(SSIM), Learned Perceptual Image Patch Similarity (LPIPS) [58] with AlexNet [28] to evaluate generated image quality. To measure facial expression consistency with audio, we adopt action unit (AU) error, landmarks distance (LMD) and pose angle error (Pose-D) computed by OpenFace [2] to evaluate the distance between generated and driven talking head. Based on landmarks, we compute landmark velocity error (LMDV) which presents the landmark movement in consecutive frame. We report quantitative comparison with competitive methods in Tab. 1 and we highly recommend to view the supplementary videos for better comparison.

Conclusions

In this paper, we propose the IDEAL-NeRF to address inconsistency between audio emotion and facial expression in talking head generation. By implement self-attention based EAA-Module, we obtain expression-audio aligned representations which are further used to be a condition for talking head generation. This not only enables IDEALNeRF to generate expression-audio aligned talking head, but also makes it possible to synthesize novel talking style of arbitrary targets.

Limitation. Though our approach can generate highfidelity talking head with consistency between audio and expression, it stills remains some unsolved problems. The IDEAL-NeRF, based on time-consuming volume rendering, can not generate real-time talking head videos. At the same time the eye gaze in our results remains static which reveals some unnatural details. In the follow-up work, we plan to compensate for aforementioned limitations and generate more vivid results in an efficient way.

Broader Impact. The purpose of talking head generation is to facilitate virtual reality applications, e.g. film dubbing and video conference, but this technique generate inexistent speech content of people which can be misused. The generated talking head of some celebrities may give rise to some social problem e.g. misleading public opinion and obtaining illegal profit. To counteract these potential disadvantages, we can make use of methods that focus on distinguishing authenticity of images [41]. The social issue of talking head merits further research and consideration and we hope the public be aware of the potential risk of talking head and avoid misuse of this novel technique.

Reference

[1] Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F. Cohen. Bringing portraits to life. ACM Trans. Graph., 36(6), Nov. 2017. 1

[2] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and LouisPhilippe Morency. Openface 2.0: Facial behavior analysis toolkit. In 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). 7

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In ICCV, 2017. 4

[4] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In CVPR, 2020. 3

[5] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing, 5(4):377–390, 2014. 7

[6] Yao-Jen Chang and Tony Ezzat. Transferable videorealistic speech animation. In Proceedings of the 2005 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA '05, page 143–151, New York, NY, USA, 2005. Association for Computing Machinery. 3

[7] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In ECCV, pages 35–51, 2020. 1, 3

Contact

kaiyuan1706@gmail.com

kai.wang960112@gmail.com

Xiangyu.peng@comp.nus.edu.sg