idea INTERNATIONAL DIGITAL ECONOMY ACADEMY
粤港澳大湾区数字经济研究院

Audio-driven Talking Face Generation: A Brief Survey

Ling-Hao CHEN
chenlinghao@idea.edu.cn

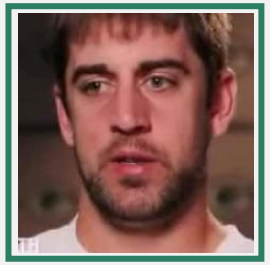
Outline

- Problem Definition
- Demo
- Road Map
  - End-to-end, GAN-based
  - 3D model based
  - (2D/3D) landmark based
- Summary
- Challenges

Problem Definition

- Driven source: audio

Input:



Identity Reference

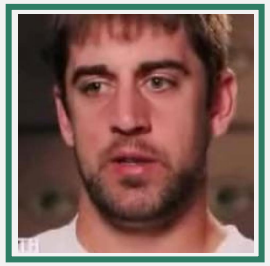
Audio Source

Output:

Problem Definition

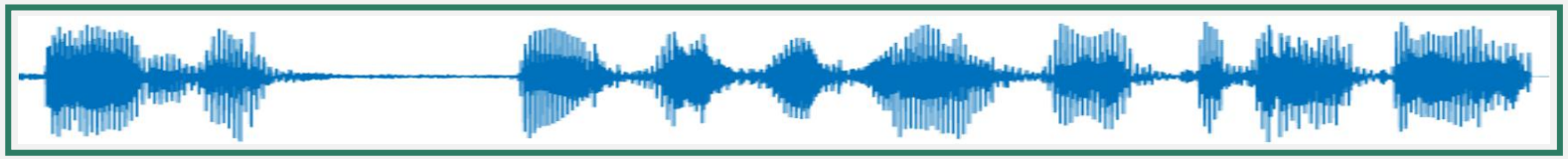
- Driven source: audio & video

Input:



Identity Reference

Video Source

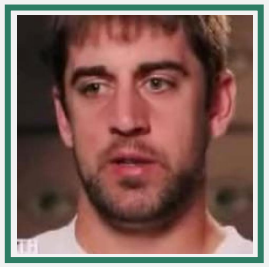


Audio Source

Output:

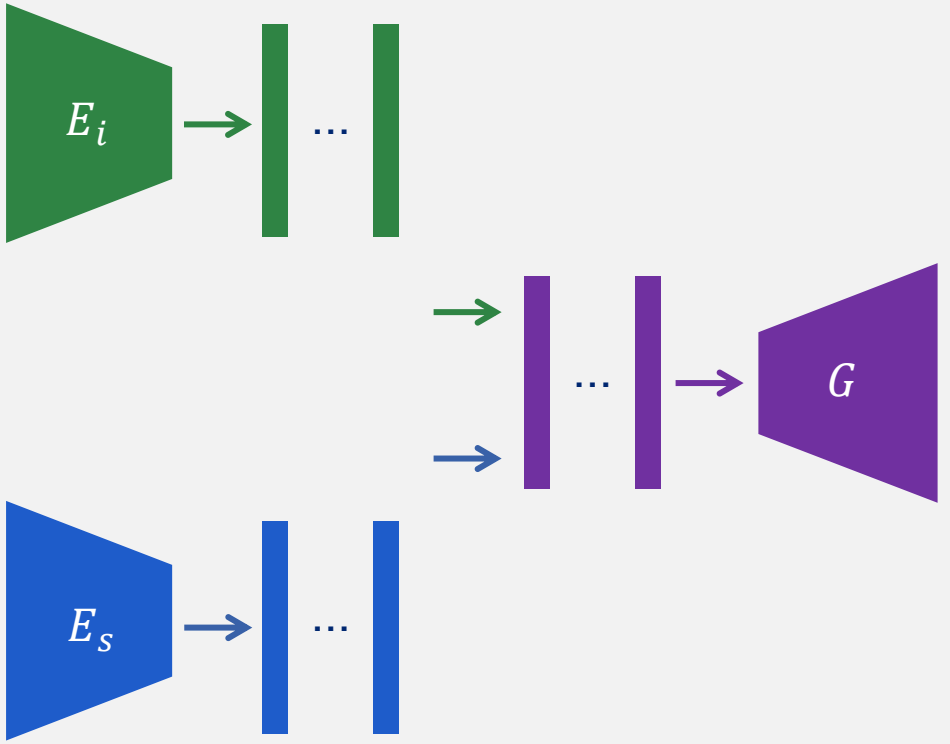
Problem Definition

Preliminaries



Identity Reference I

Driven Source S_k



Result R

$$R = G(E_i(I), E_s(\{S_k\}))$$

Demo

ATVGNet

[1]: Chen, Lele, et al. "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

Demo

FACIAL

[1]: Zhang, Chenxu, et al. "Facial: Synthesizing dynamic talking face with implicit attribute learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

Demo

PCAVS

Identity Reference | Output | Identity Reference | Audio Source

[1]: Hang Zhou et al. “Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation.” Computer Vision and Pattern Recognition (2021).

Demo

PCAVS-Eye

Identity Reference | Output | Identity Reference | Audio Source

[1]: Hang Zhou et al. “Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation.” Computer Vision and Pattern Recognition (2021).

Road Map

- End-to-end, GAN-based
- 3D model based
- (2D/3D) landmark based

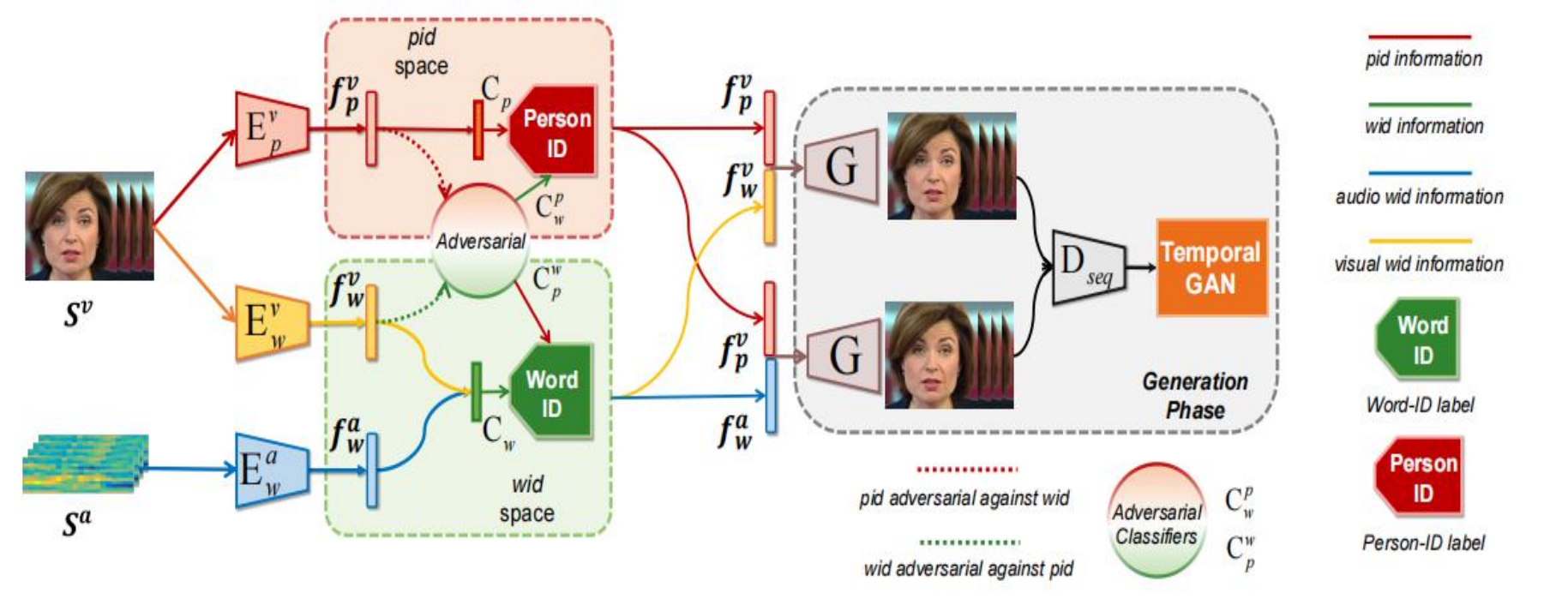
Road Map

- End-to-end, GAN-based
- 3D model based
- (2D/3D) landmark based

Road Map

➢ Adversarial Talking Face Generation

PROBLEM: How to decouple the voice-related actions in a video from the identity information of the face?

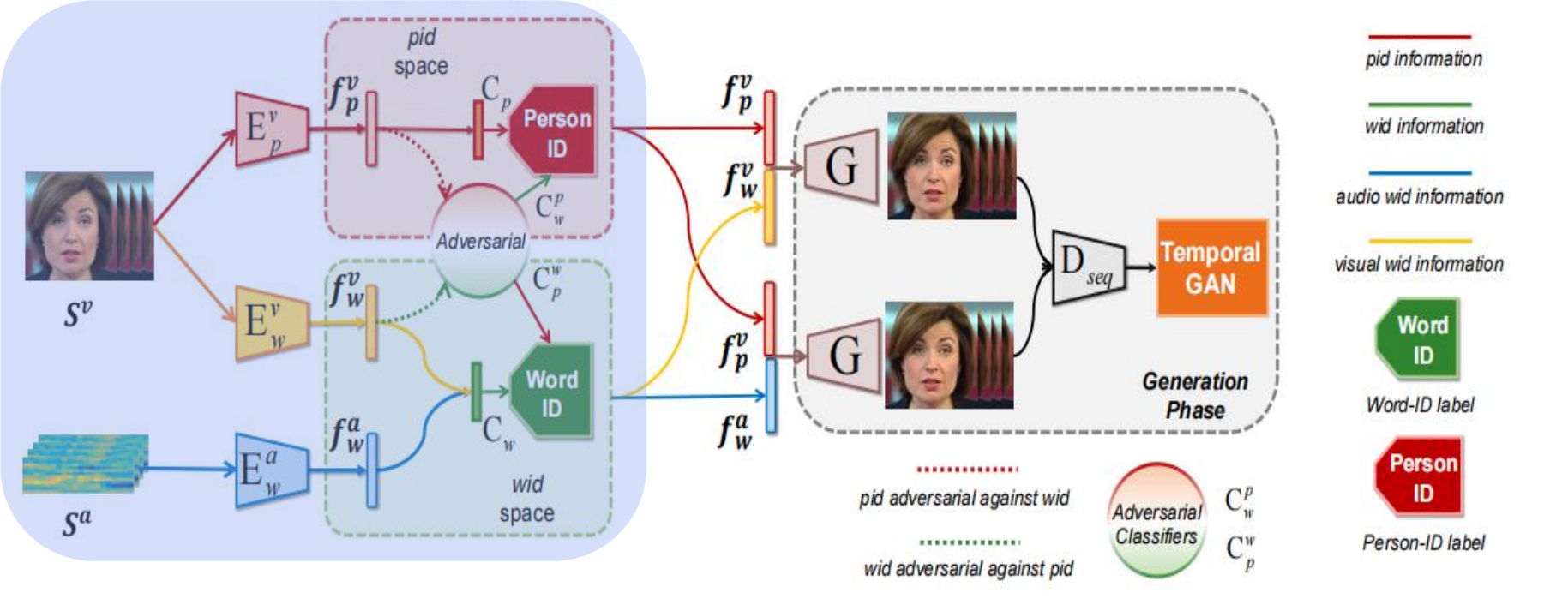


[1]: Zhou, Hang, et al. "Talking face generation by adversarially disentangled audio-visual representation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.

Road Map

➢ Adversarial Talking Face Generation

PROBLEM: How to decouple the voice-related actions in a video from the identity information of the face?

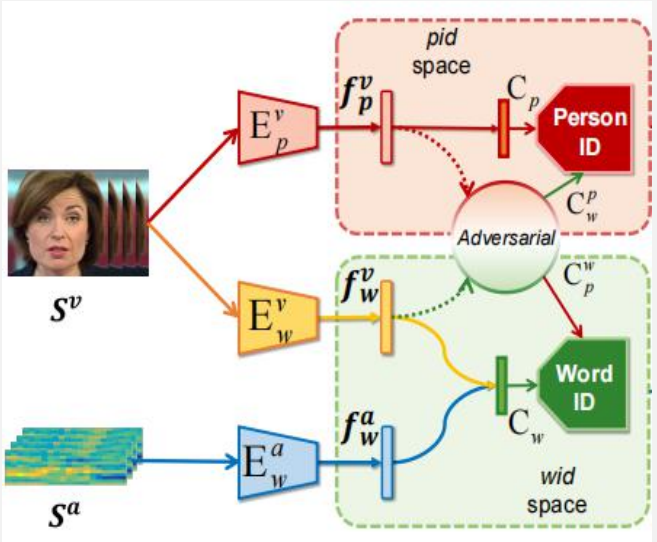


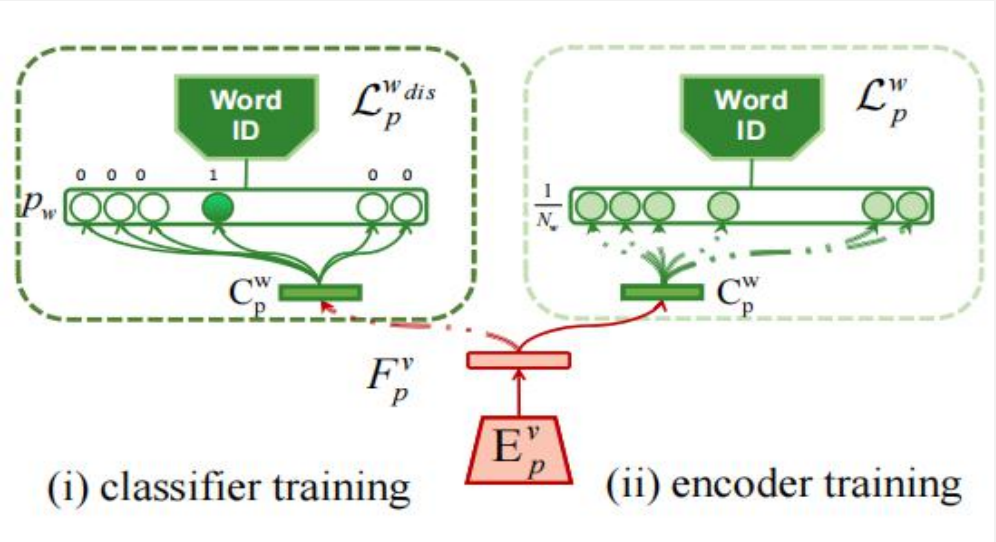
[1]: Zhou, Hang, et al. "Talking face generation by adversarially disentangled audio-visual representation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.

➢ Adversarial Talking Face Generation

PROBLEM: How to decouple the voice-related actions in a video from the identity information of the face?

Training adversarially between P-id Encoder and W-id classifier.



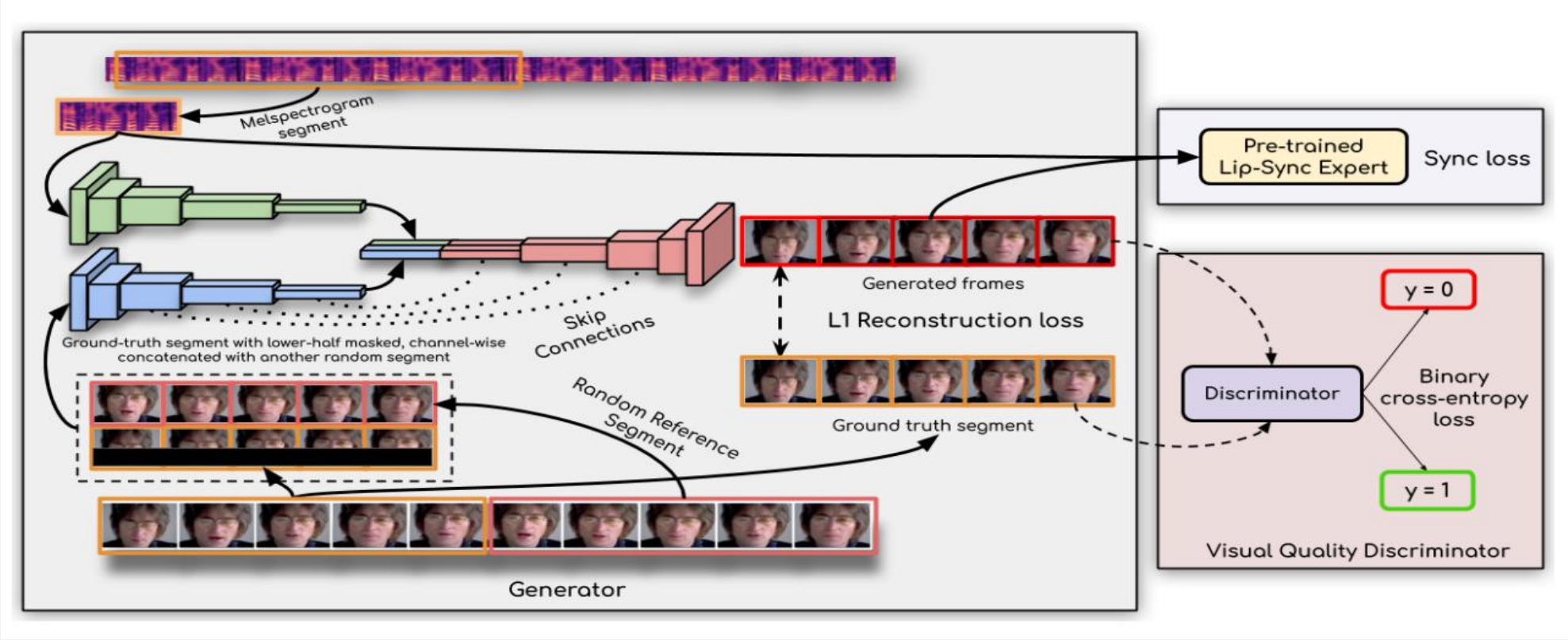


[1]: Zhou, Hang, et al. "Talking face generation by adversarially disentangled audio-visual representation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.

Road Map

➢ Wav2Lip

PROBLEM: How to do accurate lip alignment?



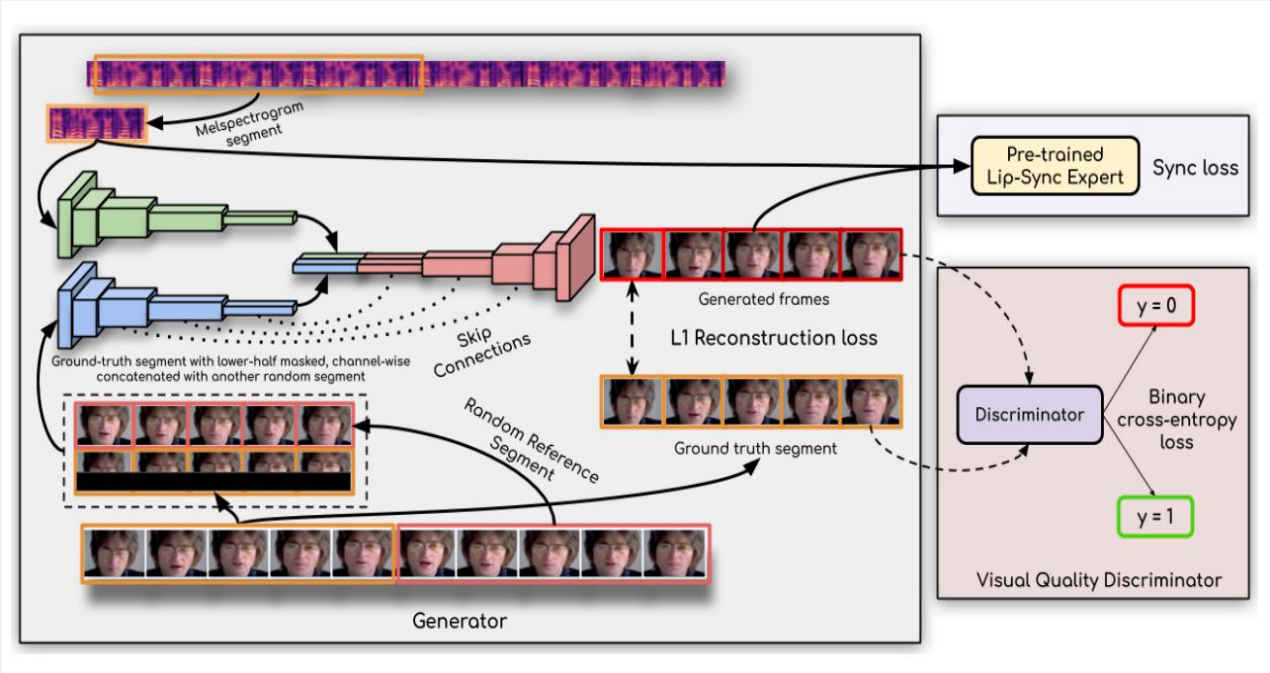
[1]: Prajwal, K. R., et al. "A lip sync expert is all you need for speech to lip generation in the wild." Proceedings of the 28th ACM International Conference on Multimedia. 2020.

Road Map

➢ Wav2Lip

PROBLEM: How to do accurate lip alignment?

- ✓ Introduce the SyncNet as Lip-Sync Expert.
- ✓ Training with audio and video jointly.
- ✓ Data Augmentation.
  - Image occlusion.
  - Random Sampling.



[1]: Prajwal, K. R., et al. "A lip sync expert is all you need for speech to lip generation in the wild." Proceedings of the 28th ACM International Conference on Multimedia. 2020.

Road Map

➢ PCAVS

PROBLEM: How to model the pose of the head?

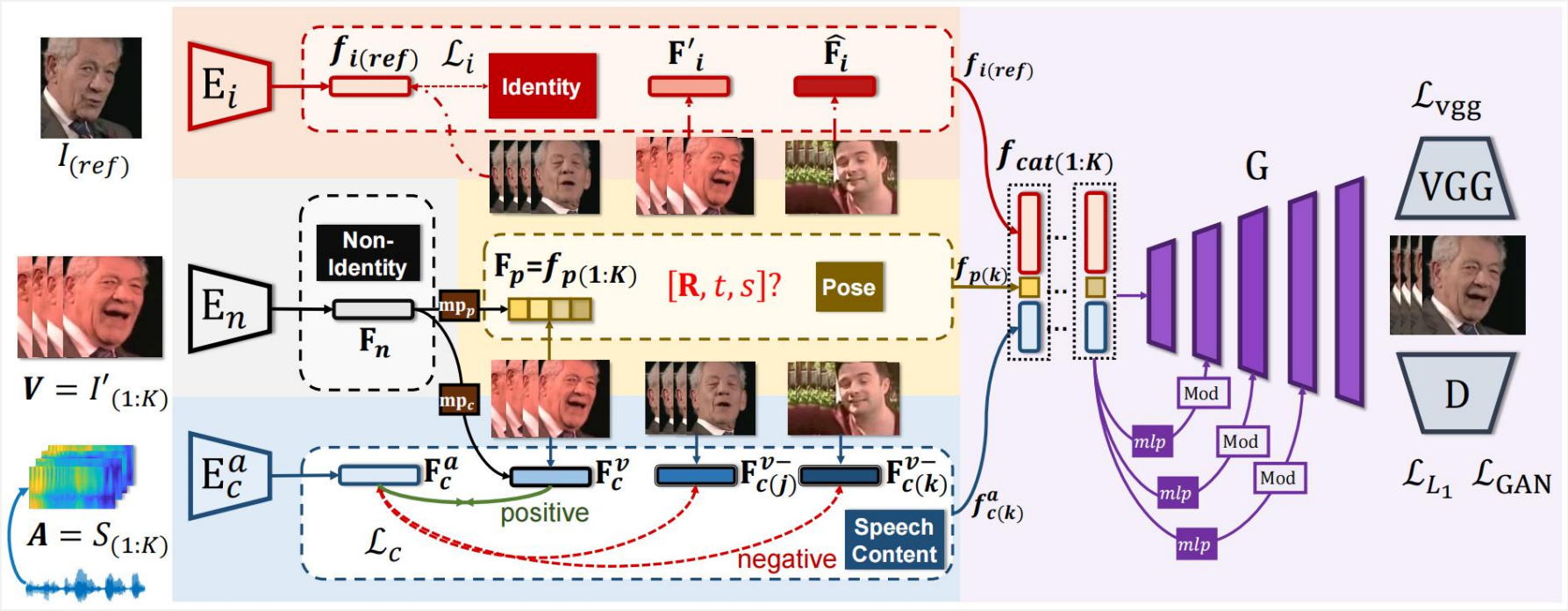
Pose Controllable:

[1]: Hang Zhou et al. “Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation..” Computer Vision and Pattern Recognition (2021).

Road Map

➢ PCAVS

Freamwork Overview:



[1]: Hang Zhou et al. “Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation..” Computer Vision and Pattern Recognition (2021).

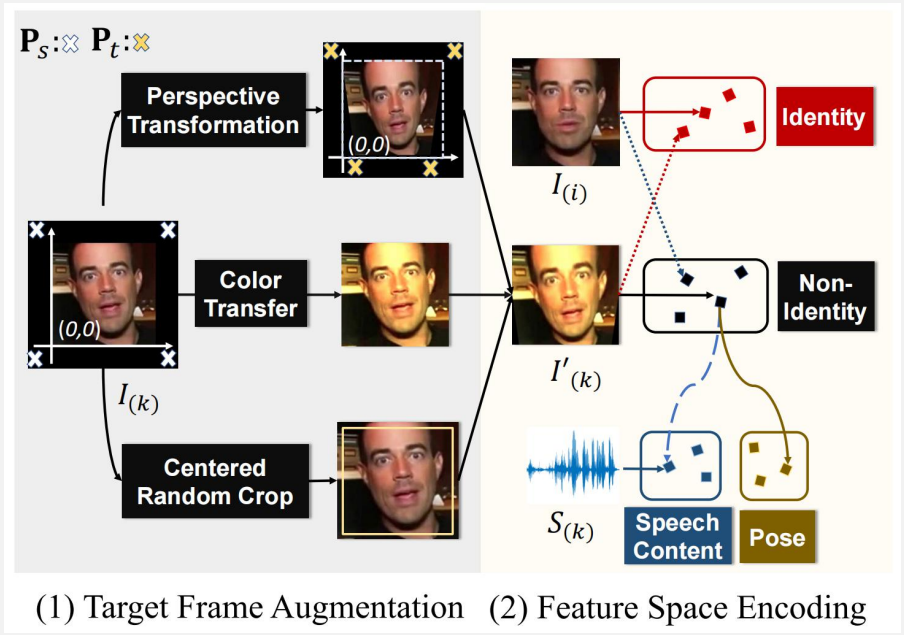
Road Map

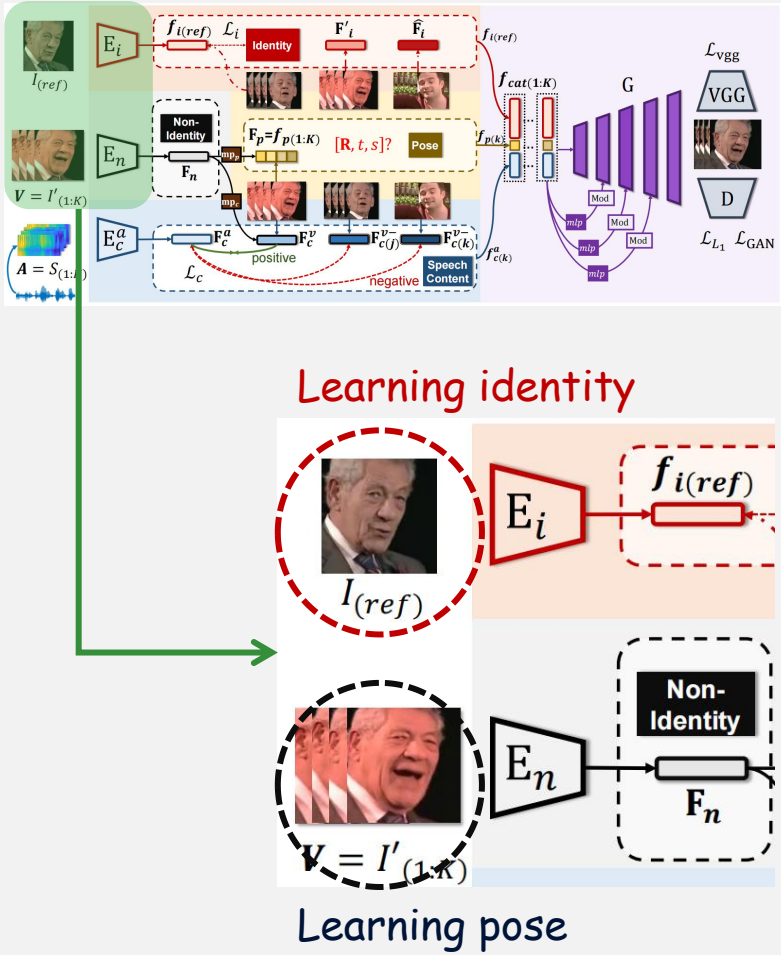
➢ PCAVS

Identity and Non-Identity Encoder (pre-train jointly):

Data augmentation:

- ✓ Perspective Transformation
- ✓ Color Transfer
- ✓ Centered Random Crop





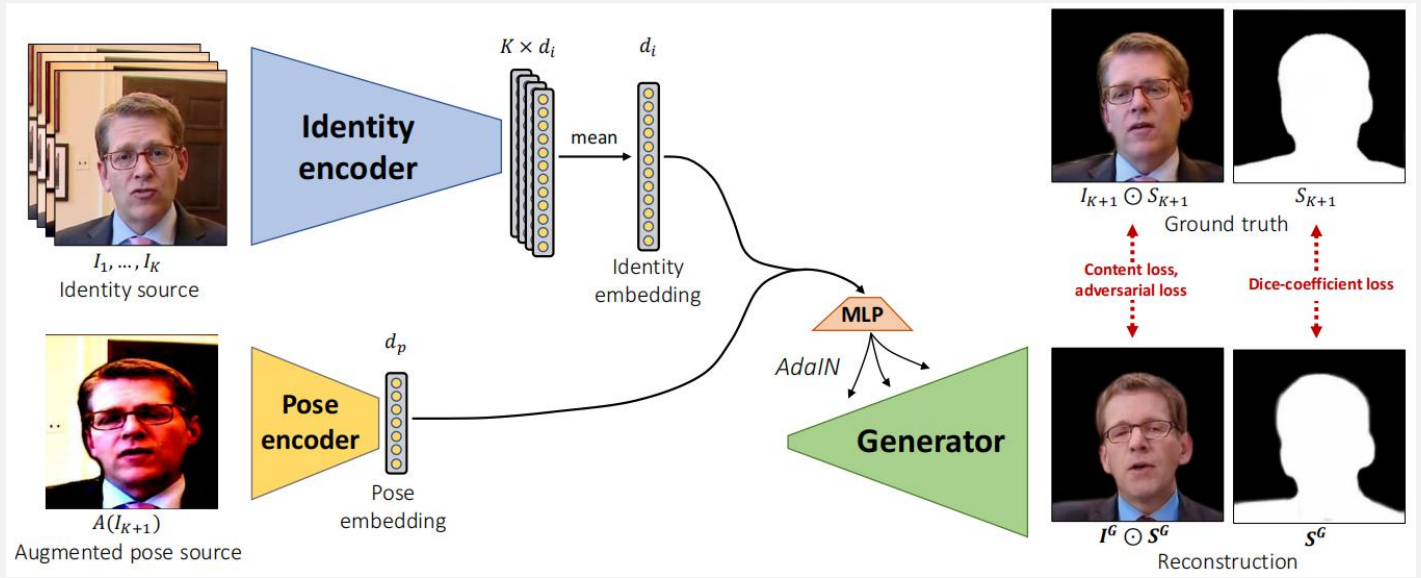
[1]: Hang Zhou et al. “Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation..” Computer Vision and Pattern Recognition (2021).

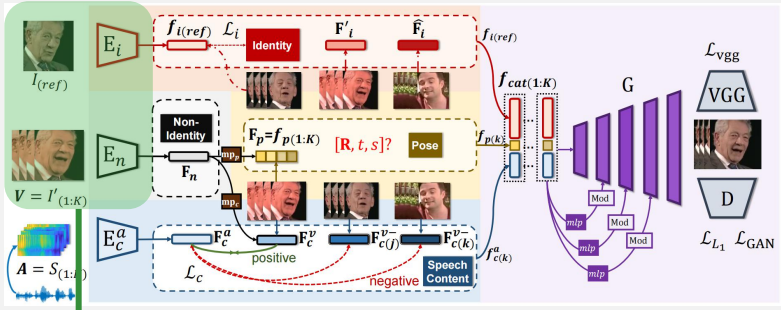
Road Map

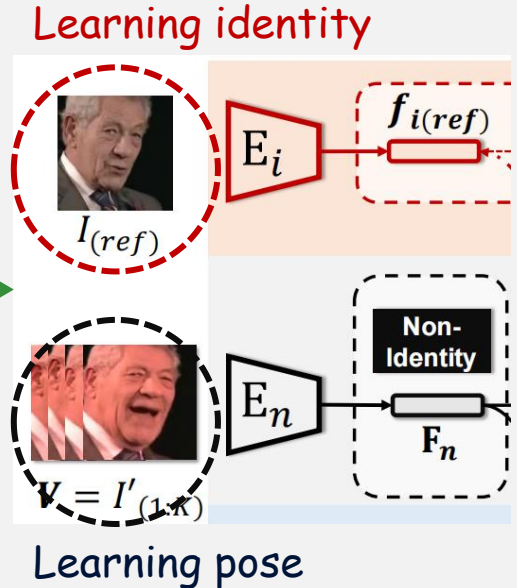
➢ PCAVS

Identity and Non-Identity Encoder (pre-train jointly):

Disentangled representation learning:







d_p: low-dimension is better

[1]: Hang Zhou et al. “Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation..” Computer Vision and Pattern Recognition (2021).

[2]: Burkov, Egor, et al. "Neural head reenactment with latent pose descriptors." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

Road Map

➢ PCAVS

Contrastive learning:

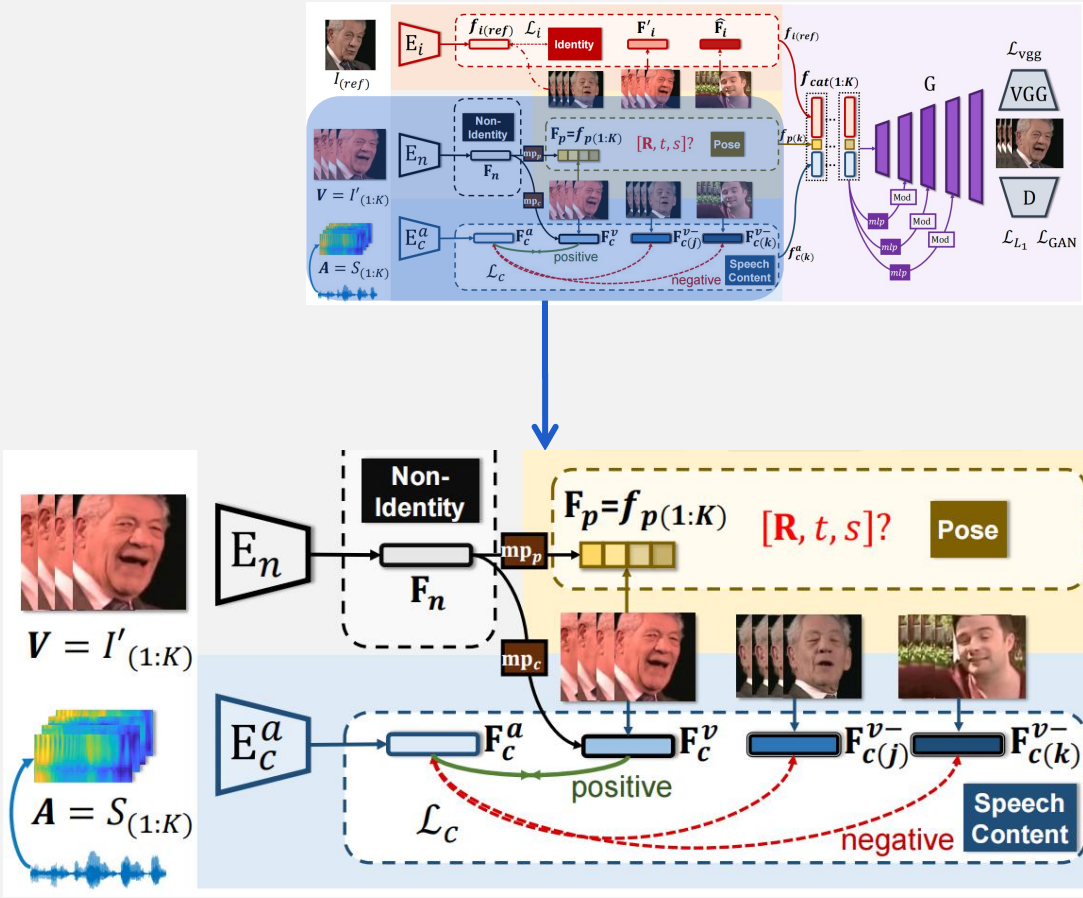
$$\mathcal{L}_c^{v2a} = -\log\left[\frac{\exp(\mathcal{D}(\mathbf{F}_c^v, \mathbf{F}_c^a))}{\exp(\mathcal{D}(\mathbf{F}_c^v, \mathbf{F}_c^a)) + \sum_{j=1}^{N^-} \exp(\mathcal{D}(\mathbf{F}_c^v, \mathbf{F}_{c(j)}^{a-}))}\right]$$

positive nagative

where $\mathcal{D}(\mathbf{F}_1, \mathbf{F}_2) = \frac{\mathbf{F}_1^{\mathbf{T}} * \mathbf{F}_2}{|\mathbf{F}_1| \cdot |\mathbf{F}_2|}$

Trainging in the speech content space:

$$\min \mathcal{L}_c^{v2a} + \mathcal{L}_c^{a2v}$$



[1]: Hang Zhou et al. “Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation..” Computer Vision and Pattern Recognition (2021).
[2]: Egor Burkov et al. “Neural Head Reenactment with Latent Pose Descriptors” Computer Vision and Pattern Recognition (2020).

Road Map

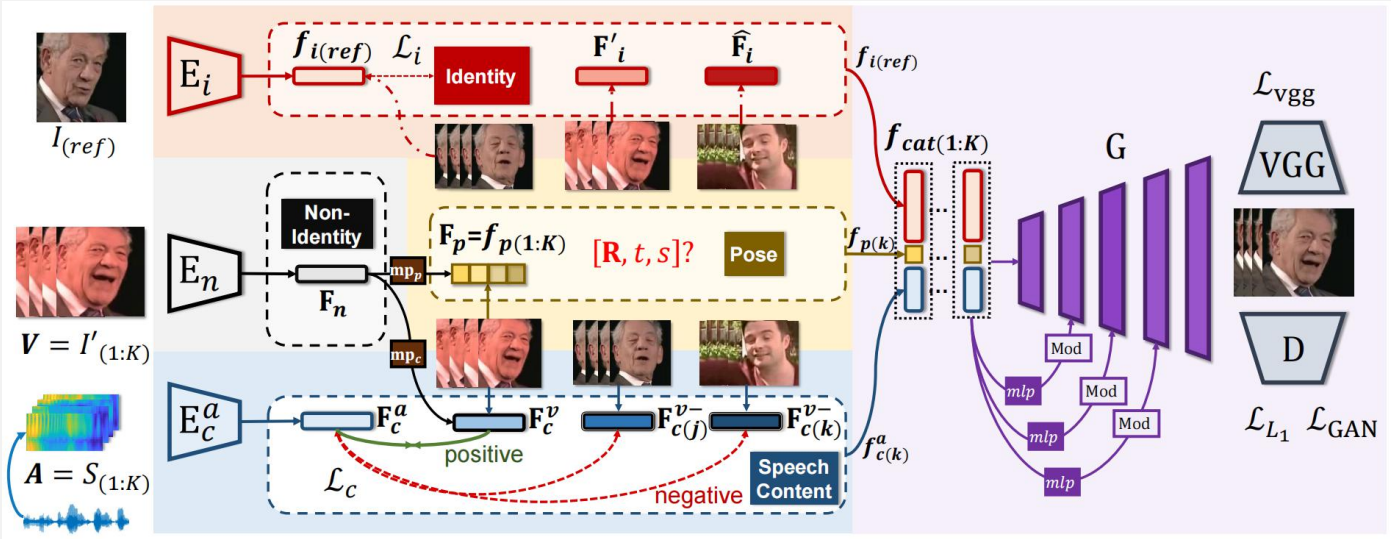
➢ PCAVS

Final stage:

$$\mathcal{L}_{\text{GAN}} = \min_{\text{G}} \max_{\text{D}} \sum_{n=1}^{N_D} (\mathbb{E}_{I_{(k)}}[\log \text{D}_n(I_{(k)})] + \mathbb{E}_{f_{cat(k)}}[\log(1 - \text{D}_n(\text{G}(f_{cat(k)})]),$$

$$\mathcal{L}_{L_1} = \sum_{i=1}^{N_D} \|\text{D}_n(\text{I}_{(k)}) - \text{D}_n(\text{G}(f_{cat(k)}))\|_1,$$

$$\mathcal{L}_{\text{vgg}} = \sum_{i=1}^{N_P} \|\text{VGG}_n(I_{(k)}) - \text{VGG}_n(\text{G}(f_{cat(k)}))\|_1.$$

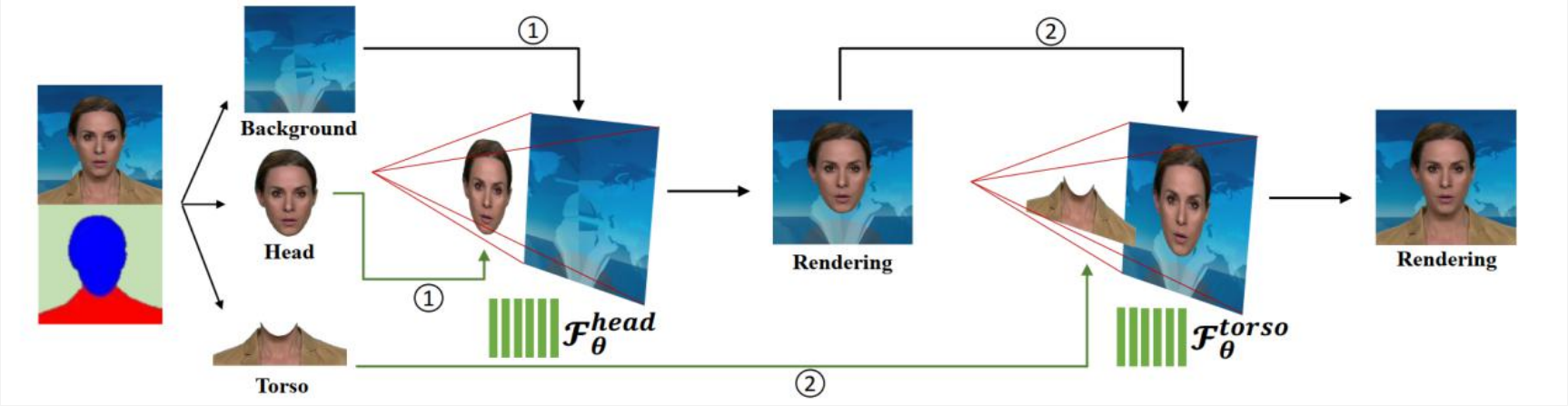


$$\mathcal{L}_{total} = \mathcal{L}_{\text{GAN}} + \lambda_1 \mathcal{L}_{L_1} + \lambda_v \mathcal{L}_{\text{vgg}} + \lambda_c \mathcal{L}_c + \lambda_i \mathcal{L}_i$$

[1]: Hang Zhou et al. “Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation..” Computer Vision and Pattern Recognition (2021).

➢ AD-NeRF

Target: Generate the talking face of a specific person in any pose.

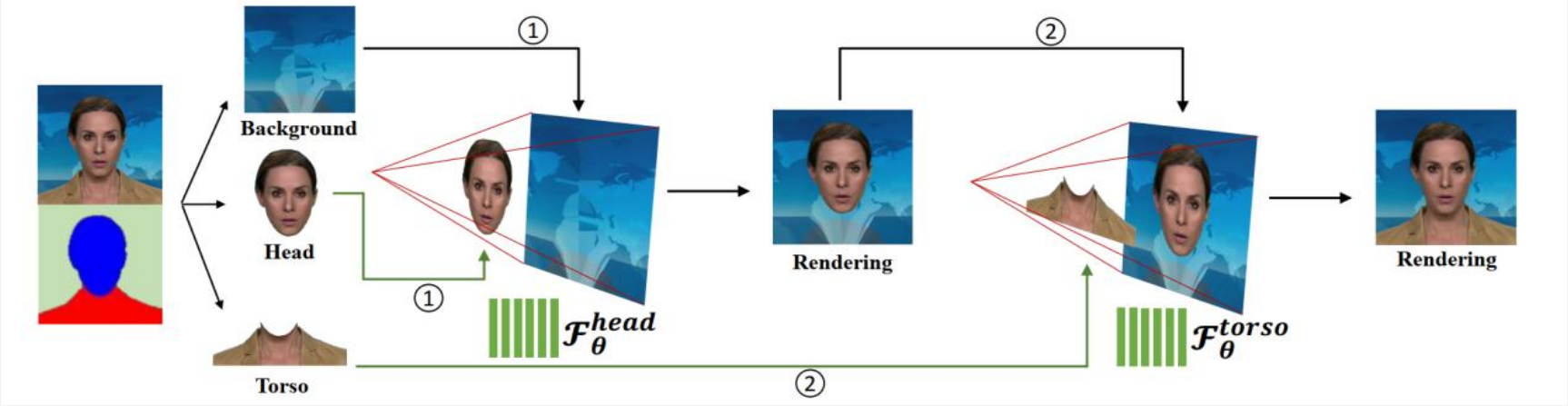


Basic idea and pipeline

[1]: Yudong Guo et al. “AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis.” arXiv: Computer Vision and Pattern Recognition (2021): n. pag.

➢ AD-NeRF

Target: Generate the talking face of a specific person in any pose.



Basic idea and pipeline

Compare with NeRF (color values c accompanied with densities σ):

$$\text{NerF: } \mathcal{F}(d, x) \rightarrow (c, \sigma)$$

$$\text{AD-NerF: } \mathcal{F}(a, d, x) \rightarrow (c, \sigma)$$

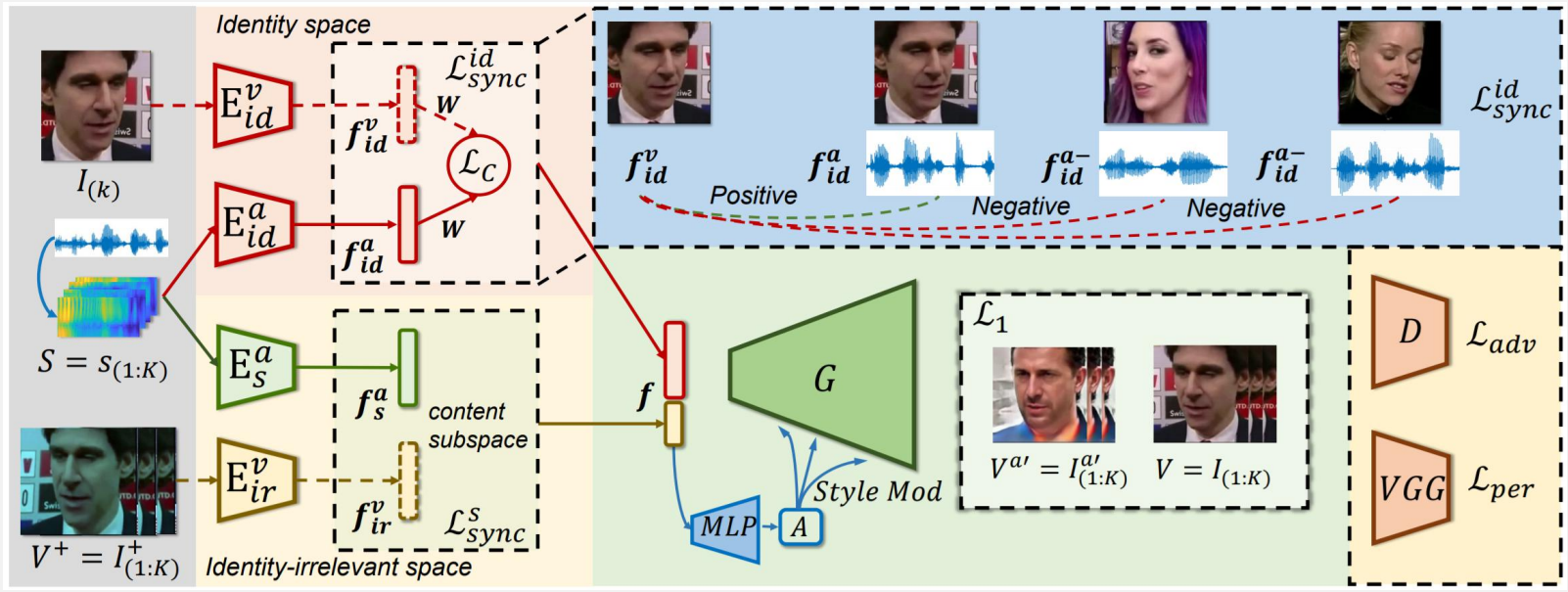
$$\longrightarrow \quad \mathcal{C}(\mathbf{r}; \theta, \Pi, \mathbf{a}) = \int_{t_n}^{t_f} \sigma_\theta(\mathbf{r}(t)) \cdot \mathbf{c}_\theta(\mathbf{r}(t), \mathbf{d}) \cdot T(t) dt$$

[1]: Yudong Guo et al. “AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis.” arXiv: Computer Vision and Pattern Recognition (2021): n. pag.

Road Map

➢ Speech2Talking-Face

Contribution: The discrimination of identity features is strengthened; for the first time, face reconstruction and speech-driven face speech generation are completed in the same framework.



[1]: Yasheng Sun et al. “Speech2Talking-Face: Inferring and Driving a Face with Synchronized Audio-Visual Representation” International Joint Conference on Artificial Intelligence (2021).

Road Map

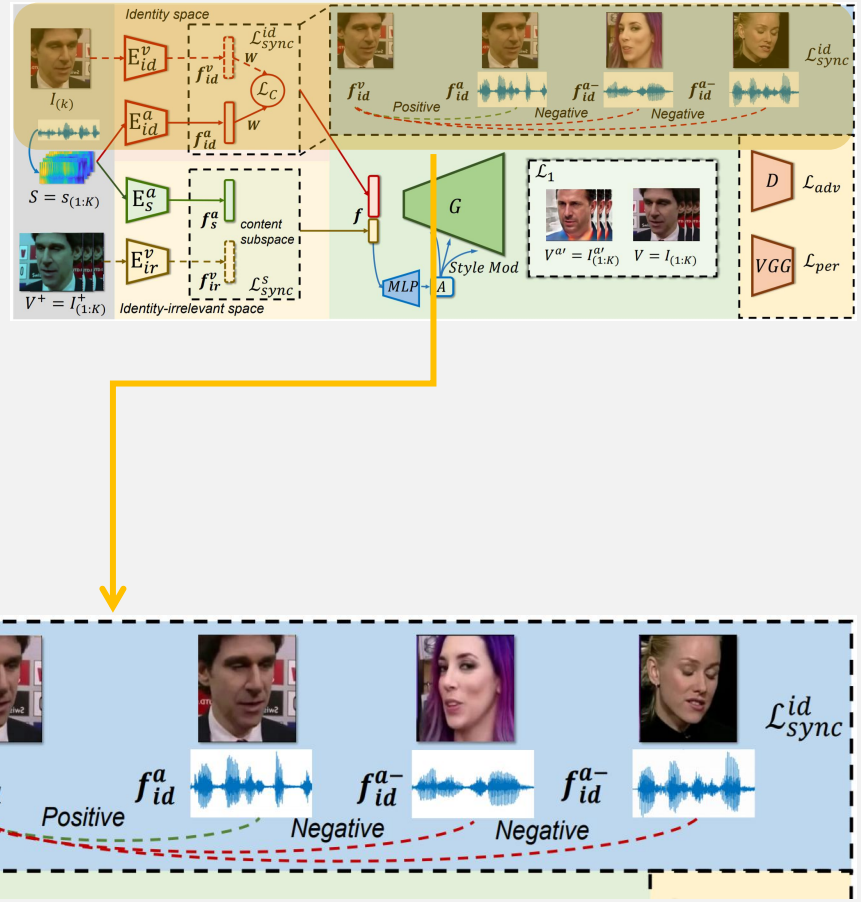
➢ Speech2Talking-Face

Identity space:

$$\mathcal{L}_{sync}^{id} = -\log\left[\frac{e^{\cos\theta(f_{id}^{v}, f_{id}^{a})}}{\underbrace{e^{\cos\theta(f_{id}^{v}, f_{id}^{a})}}_{\text{positive}} + \underbrace{\sum_{i=1}^{M} e^{\cos\theta(f_{id}^{v-}(i), f_{id}^{a})}}_{\text{nagative}}}\right]$$

where $\cos\theta_{(f_1, f_2)} = \frac{f_1^{\mathbf{T}} * f_2}{|f_1| \cdot |f_2|}$

$$\mathcal{L}_c = -\sum_{i=1}^{M} \log\left[\frac{e^{s\cos(\theta_{(W_j, f_{id})}+m)}}{e^{s\cos(\theta_{(W_j, f_{id})}+m)} + \sum_{p\neq j}^{P} e^{s\cos\theta_{(W_p, f_{id})}}}\right]$$



[1]: Yasheng Sun et al. “Speech2Talking-Face: Inferring and Driving a Face with Synchronized Audio-Visual Representation” International Joint Conference on Artificial Intelligence (2021).

Road Map

➢ Speech2Talking-Face

Identity-irrelevant space:

$$\mathcal{L}_{sync}^{s} = -\log\left[\frac{e^{\cos\theta(\Delta\mathbf{f}_{ir}^{v},\mathbf{f}_{s}^{a})}}{e^{\cos\theta(\Delta\mathbf{f}_{ir}^{v},\mathbf{f}_{s}^{a})} + \sum_{i=1}^{M} e^{\cos\theta(\Delta\mathbf{f}_{ir}^{v-}(i),\mathbf{f}_{s}^{a})}}\right]$$

positive　　nagative

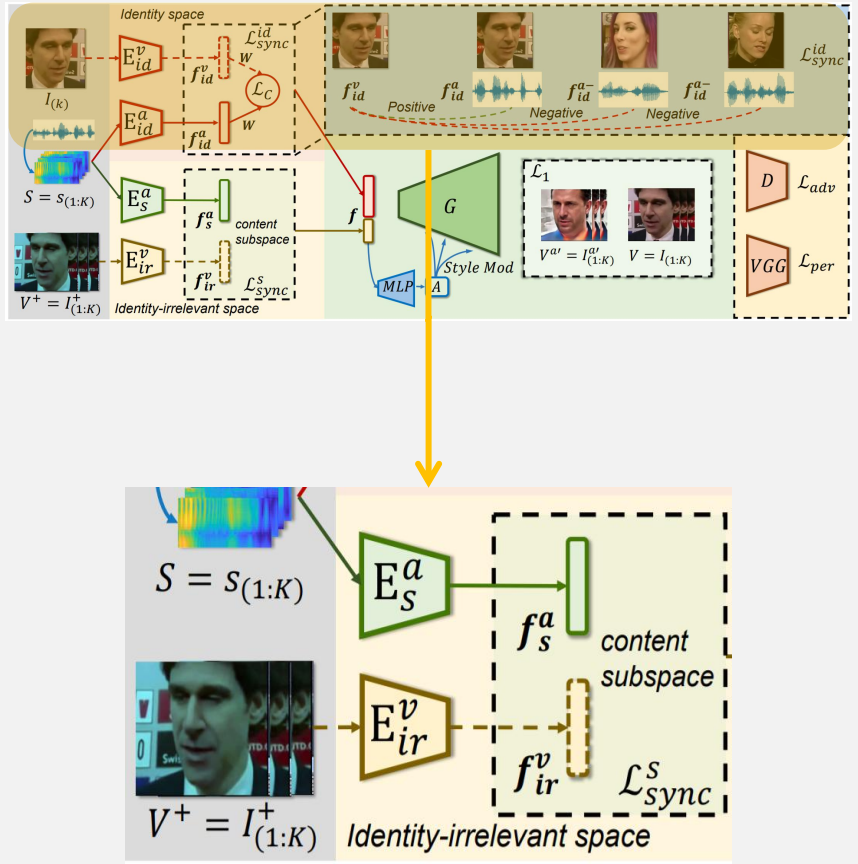
where $\Delta\mathbf{f}_{ir}^{v} = \{f_{ir}^{v_2} - f_{ir}^{v_1}, \ldots, f_{ir}^{v_K} - f_{ir}^{v_1}\}$

$$\mathbf{f}_{s}^{a} = \{f_{s}^{a_2}, \ldots, f_{s}^{a_K}\}$$

Final stage:

$$\mathcal{L}_{all} = \mathcal{L}_{\text{adv}} + \lambda_r\mathcal{L}_{\text{rec}} + \lambda_c\mathcal{L}_c + \lambda_{sync}^{id}\mathcal{L}_{sync}^{id} + \lambda_{sync}^{s}\mathcal{L}_{sync}^{s}$$

Adversarial loss, Reconstruction loss, Several contrastive loss



[1]: Yasheng Sun et al. “Speech2Talking-Face: Inferring and Driving a Face with Synchronized Audio-Visual Representation” International Joint Conference on Artificial Intelligence (2021).

Road Map

- End-to-end, GAN-based
- 3D model based
- (2D/3D) landmark based

Road Map

- End-to-end, GAN-based
- 3D model based
- (2D/3D) landmark based

Road Map

➢ ADEVP

PROBLEM: How to decouple the emotion and the speech content?

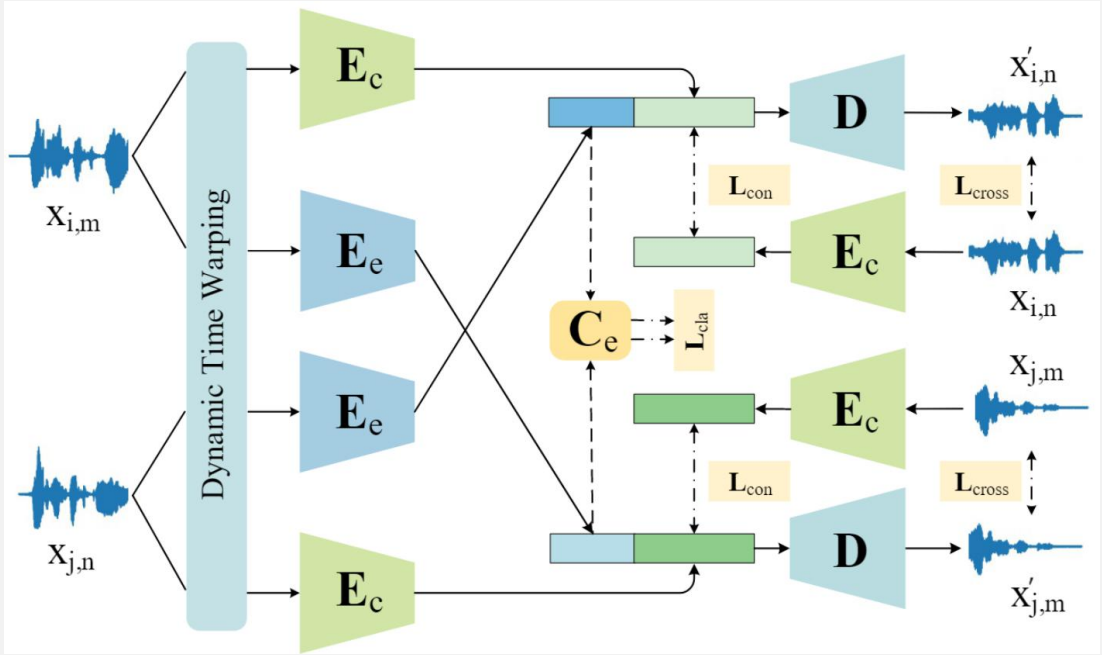
Q: How to length-align speech with uneven length?
A: DTW Algorithm.

We will train E_c, E_e, D in this stage.

Loss:

$$L_{cross} = \|D(E_c(x_{i,m}), E_e(x_{j,n})) - x_{i,n}\|_2 + \|D(E_c(x_{j,n}), E_e(x_{i,m})) - x_{j,m}\|_2$$

$$L_{self} = \|D(E_c(x_{i,m}), E_e(x_{i,m})) - x_{i,m}\|_2 + \|D(E_c(x_{j,n}), E_e(x_{j,n})) - x_{j,n}\|_2.$$



Cross-reconstruction for disentanglemen

[1]: Ji, Xinya, et al. "Audio-driven emotional video portraits." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

Road Map

➢ ADEVP

PROBLEM: How to decouple the emotion and the speech content?

Loss:

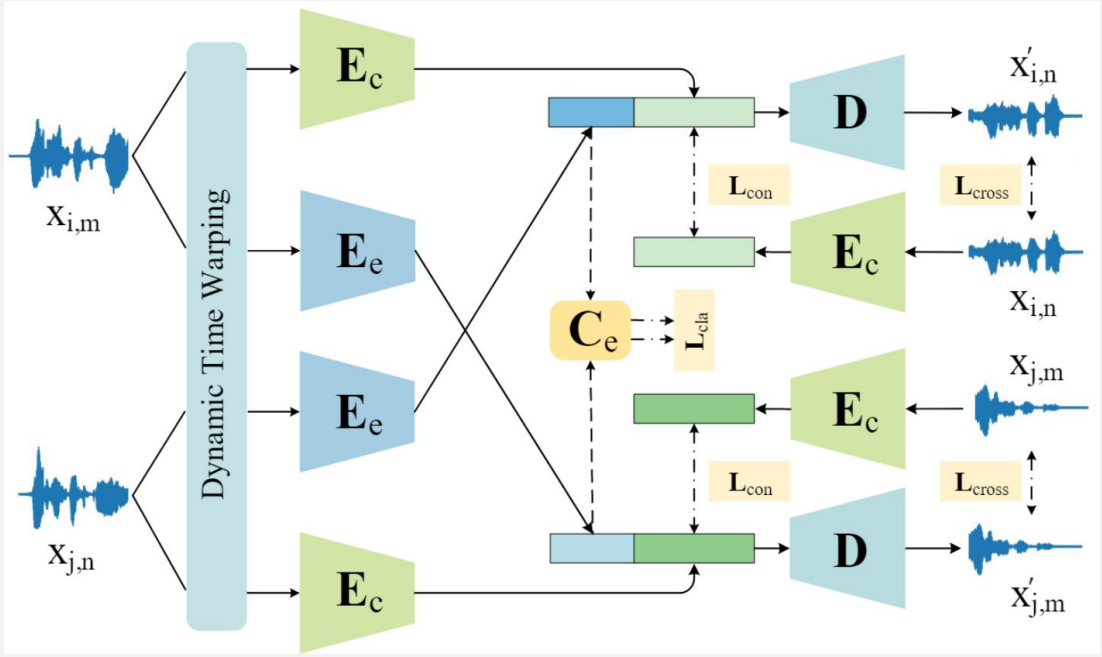
$$L_{cross} = \|D(E_c(x_{i,m}), E_e(x_{j,n})) - x_{i,n}\|_2 + \|D(E_c(x_{j,n}), E_e(x_{i,m})) - x_{j,m}\|_2$$

$$L_{self} = \|D(E_c(x_{i,m}), E_e(x_{i,m})) - x_{i,m}\|_2 + \|D(E_c(x_{j,n}), E_e(x_{j,n})) - x_{j,n}\|_2.$$

$$L_{cla} = -\sum_{k=1}^{N}(p_k * \log q_k)$$

$$L_{con} = \sum_{k=i,j} \|E_c(x_{k,m}) - E_c(x_{k,n})\|_1$$

$$L_{dis} = L_{cross} + L_{self} + \lambda_{cla} L_{cla} + \lambda_{con} L_{con}$$

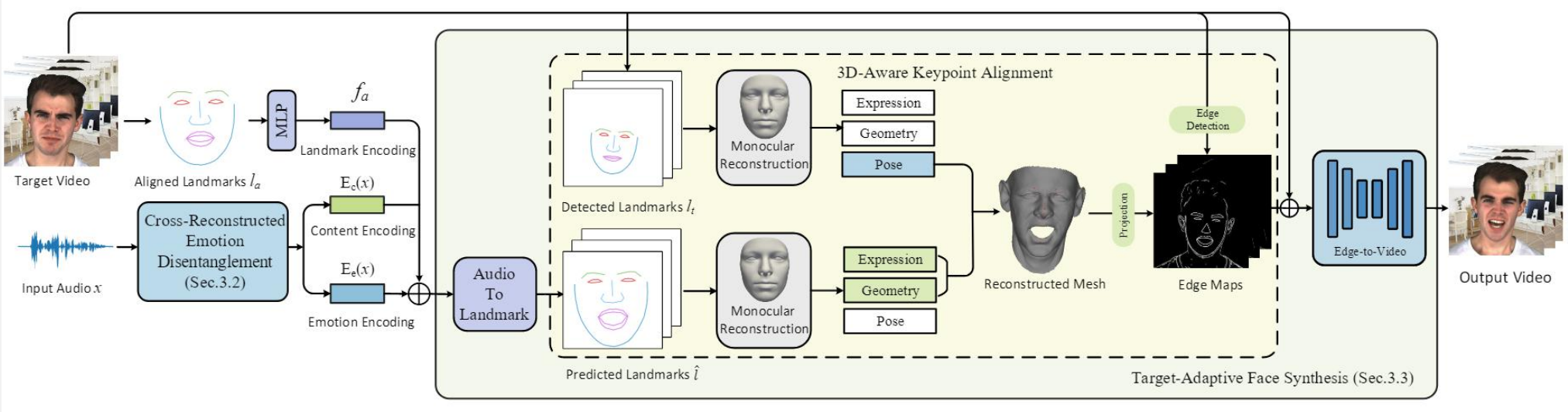


Cross-reconstruction for disentanglemen

[1]: Ji, Xinya, et al. "Audio-driven emotional video portraits." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

➢ ADEVP

PROBLEM: How to decouple the emotion and the speech content?



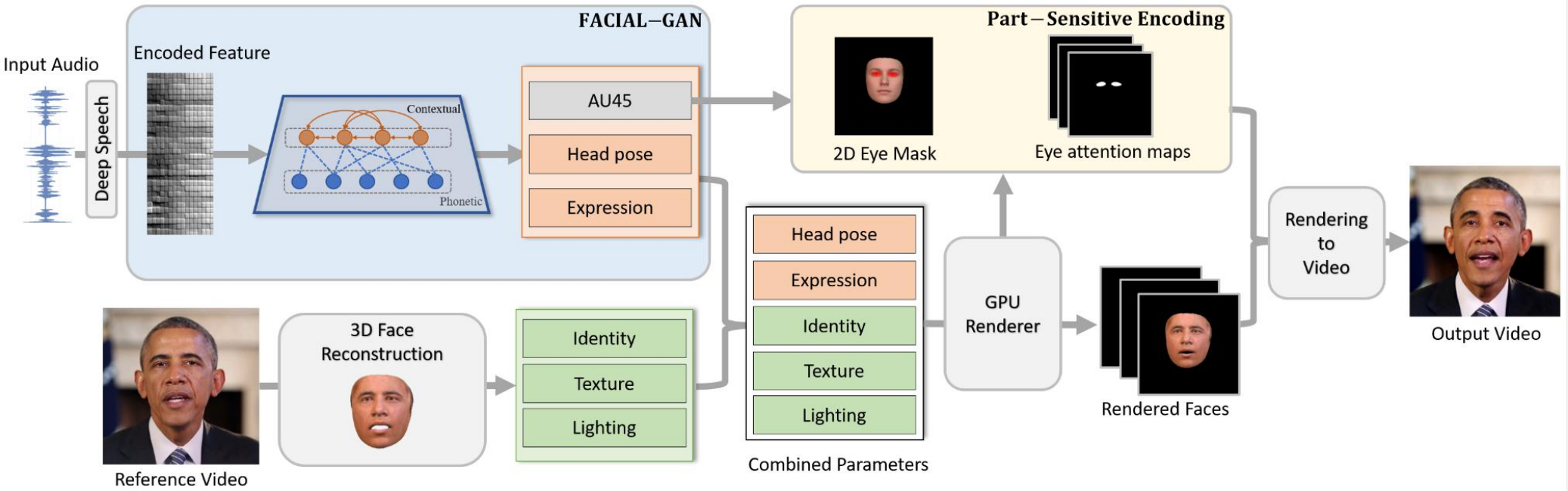
Overview of Emotional Video Portrait algorithm

[1]: Ji, Xinya, et al. "Audio-driven emotional video portraits." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

Road Map

➢ FACIAL

(1) How do explicit and implicit attributes potentially affect each other?
(2) How to model implicit properties?

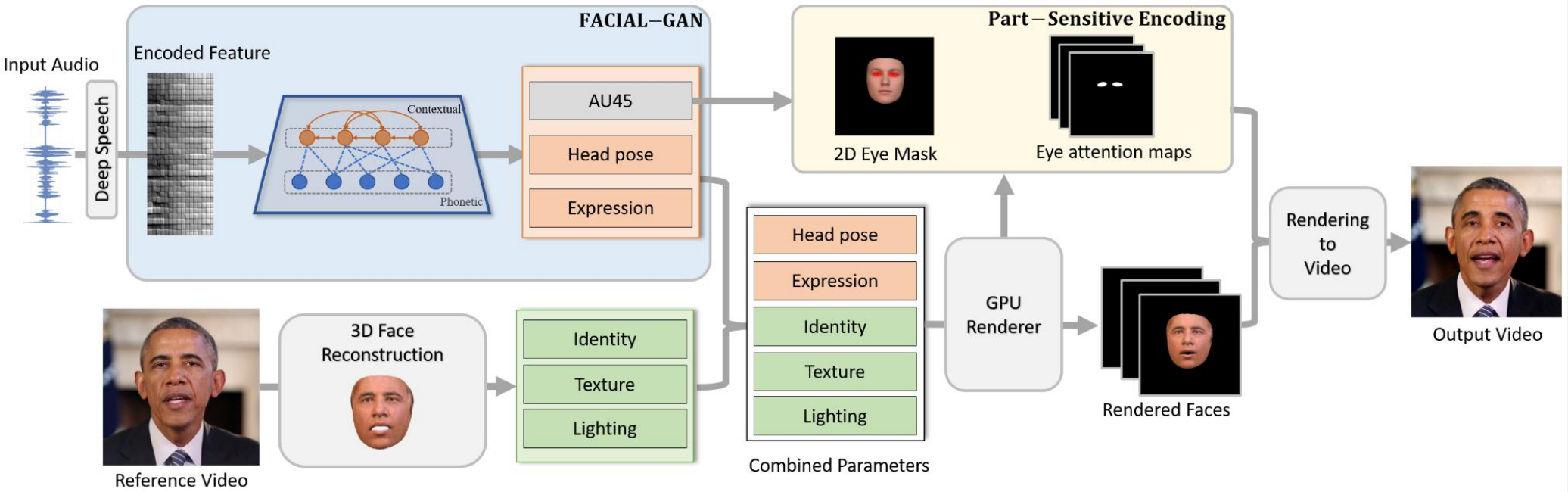


[1]: Chenxu Zhang et al. “FACIAL: Synthesizing Dynamic Talking Face with Implicit Attribute Learning” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

Road Map

➢ FACIAL

The generated videos have not only synchronized lip movements, but also natural head movements and eye blink information.



[1]: Chenxu Zhang et al. “FACIAL: Synthesizing Dynamic Talking Face with Implicit Attribute Learning” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

Road Map

- End-to-end, GAN-based
- 3D model based
- (2D/3D) landmark based

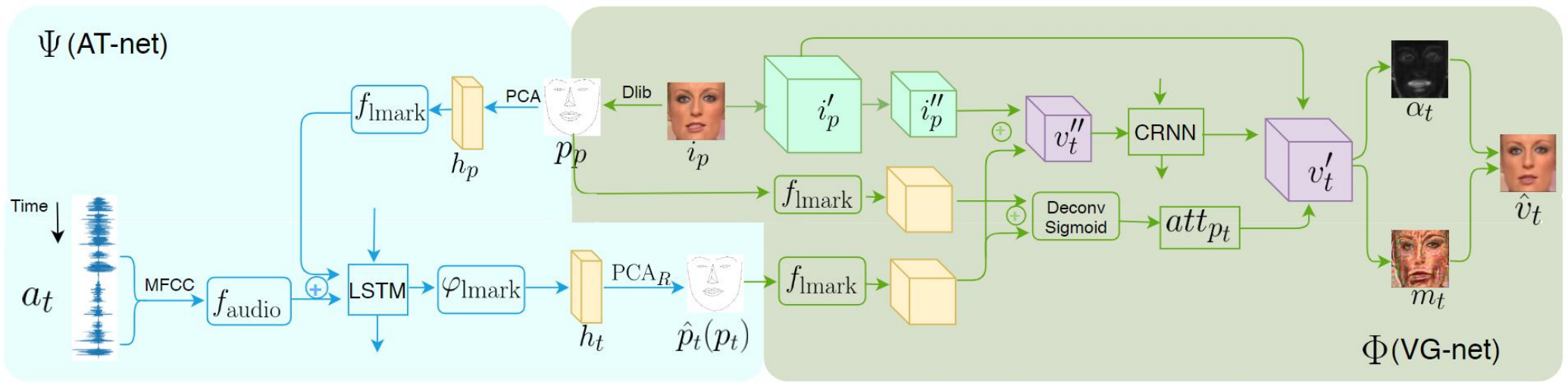
Road Map

- End-to-end, GAN-based
- 3D model based
- (2D/3D) landmark based

Road Map

➢ ATVG-Net

The generated videos have not only synchronized lip movements, but also natural head movements and eye blink information.

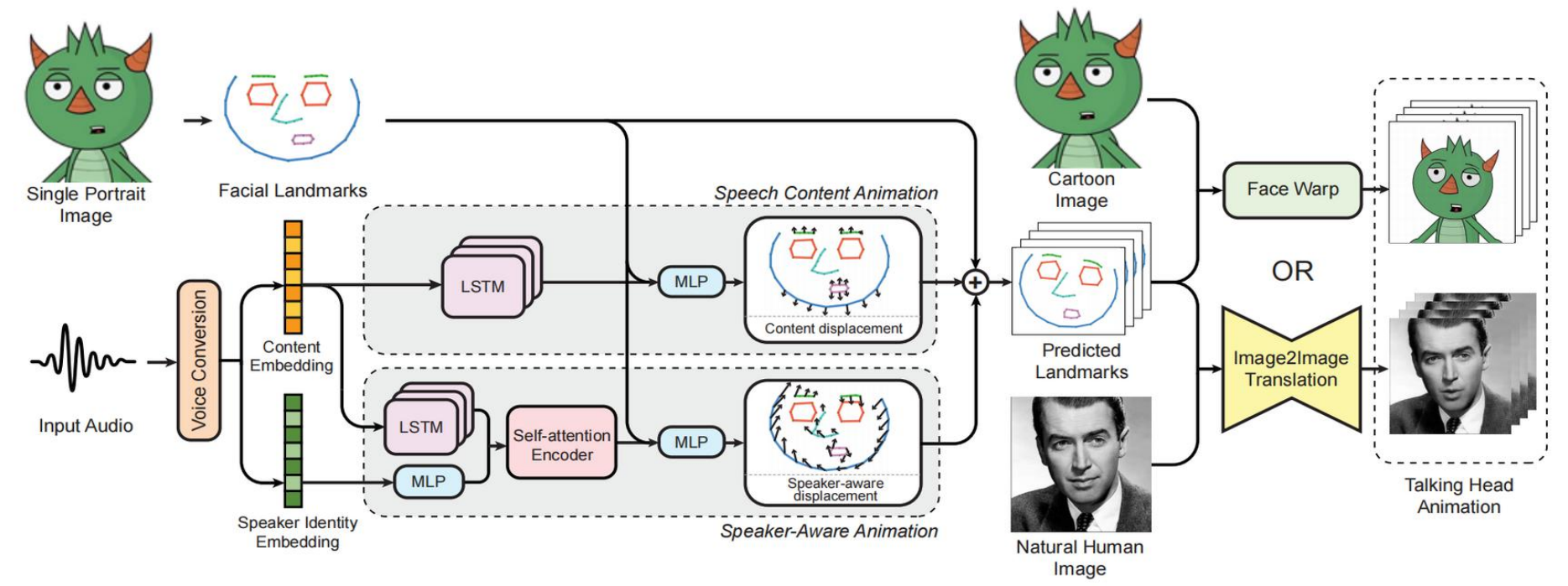


[1]: Chen, Lele, et al. "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

Road Map

➢ MakeItTalk

Alignment of lip movements, facial expressions, head movements.



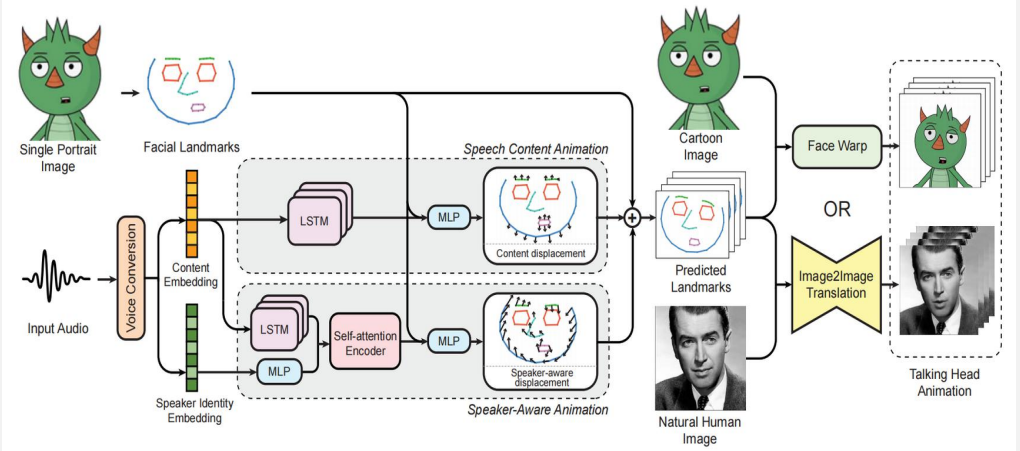
[1]: Zhou, Yang, et al. "MakeItTalk: speaker-aware talking-head animation." ACM Transactions on Graphics (TOG) 39.6 (2020): 1-15.

Road Map

➢ MakeItTalk

Alignment of lip movements, facial expressions, head movements.

- Learning active dynamics: $\Delta \boldsymbol{p}$.
- Get dynamic landmarks of driven object.
- The GLSL-based method is used to stretch the landmarks for cartoon images to achieve the purpose; the images are generated using the Image2Image architecture.



[1]: Zhou, Yang, et al. "MakeItTalk: speaker-aware talking-head animation." ACM Transactions on Graphics (TOG) 39.6 (2020): 1-15.

Summary

Methods	GAN-based	Landmark-based	3D-model-based	Input	Journal/Conference
Adversarial TFG	☑	☐	☐	Audio Referenced Video	AAAI-19
ATVG-Net	☑	☑	☐	Audio Image	CVPR-19
Wav2Lip	☑	☐	☐	Audio Referenced Video	ACM MM-20
MakeItTalk	☑	☑	☐	Audio Image	ACM TOG-21
ADEVP	☑	☑	☑	Audio Image	CVPR-21
PCAVS	☑	☐	☐	Audio Referenced Video Image	CVPR-21
FICIAL	☑	☐	☑	Audio Image	CVPR-21
Speech2Talking-Face	☑	☐	☐	Audio Referenced Video Image	IJCAI-21
Imitating ADTF-Sync	☐	☐	☑	Audio Referenced Video Image	ACM MM-21
Info-Bottleneck TFG	☑	☑	☑	Audio Image	CIKM-21(short)
AD-NeRF	☑	☑	☐	Audio Image	ICCV-21

Challenges

- Jitter between video frames.
- How to make posture and expression controllable and editable?
- How to better decouple the information between face ID and lip movements, blinks, and facial expression?
- Robustness of the model and the sensitivity of noise.
- Generated faces are not enough yet.

Reference

[1]: Zhou, Yang, et al. "MakeItTalk: speaker-aware talking-head animation." ACM Transactions on Graphics (TOG) 39.6 (2020): 1-15.
[2]: Zhou, Hang, et al. "Talking face generation by adversarially disentangled audio-visual representation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.
[3]: Prajwal, K. R., et al. "A lip sync expert is all you need for speech to lip generation in the wild." Proceedings of the 28th ACM International Conference on Multimedia. 2020.
[4]: Hang Zhou et al. “Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation..” Computer Vision and Pattern Recognition (2021).
[5]: Burkov, Egor, et al. "Neural head reenactment with latent pose descriptors." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[6]: Yudong Guo et al. “AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis.” arXiv: Computer Vision and Pattern Recognition (2021): n. pag.
[7]: Yasheng Sun et al. “Speech2Talking-Face: Inferring and Driving a Face with Synchronized Audio-Visual Representation” International Joint Conference on Artificial Intelligence (2021).
[8]: Ji, Xinya, et al. "Audio-driven emotional video portraits." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[9]: Chenxu Zhang et al. “FACIAL: Synthesizing Dynamic Talking Face with Implicit Attribute Learning” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[10]: Chen, Lele, et al. "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
[11]: Ben Mildenhall et al. “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis” European Conference on Computer Vision (2020).
[12]: Wu, Haozhe, et al. "Imitating Arbitrary Talking Style for Realistic Audio-Driven Talking Face Synthesis." Proceedings of the 29th ACM International Conference on Multimedia. 2021.
[13]: Tang, Jie, et al. "Talking Face Generation Based on Information Bottleneck and Complementary Representations." Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021.
[14]: Doukas, Michail Christos, Stefanos Zafeiriou, and Viktoriia Sharmanska. "HeadGAN: One-shot Neural Head Synthesis and Editing." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[15]: Huang, Po-Hsiang, Fu-En Yang, and Yu-Chiang Frank Wang. "Learning Identity-Invariant Motion Representations for Cross-ID Face Reenactment." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

THANK YOU

Ling-Hao CHEN (陈 凌灏)
chenlinghao@idea.edu.cn
School of Computer Science and Technology
Xidian University
G-343, 266 Xinglong Section of Xifeng Road, Xi'an, ShaanXi, P.R. China.

www.idea.edu.cn