IEEE ICASSP 2022

# Meta Talk: Learning to Data-Efficiently Generate Audio-Driven Lip-Synchronized Talking Face with High Definition

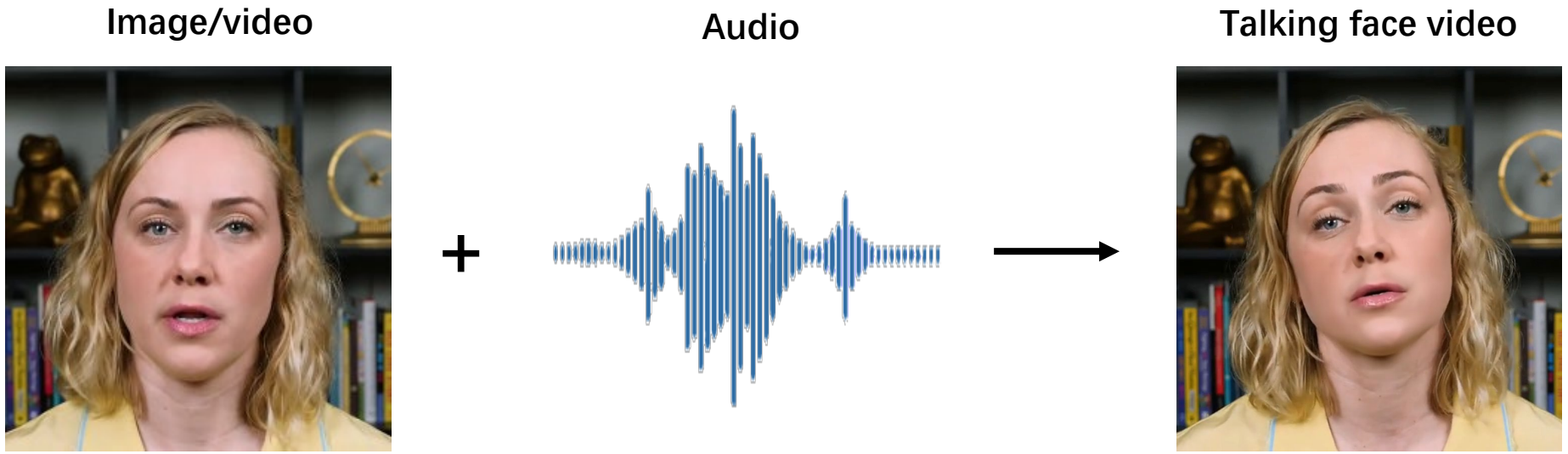
Yuhan Zhang[1]*, Weihua He[1] *, Minglei Li[2], Kun Tian[1], Ziyang Zhang[1#], Jie Cheng[1], Yaoyuan Wang[1#], Jianxing Liao[1]

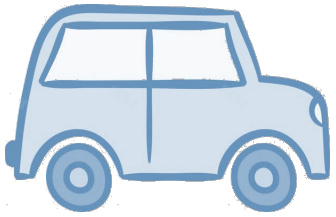
[1] Advanced Computing and Storage Lab, Huawei Technologies Co. Ltd., China

[2] Language & Speech Innovation Lab, Huawei Technologies Co. Ltd., China

# Introduction

**Audio-Driven Talking Face**



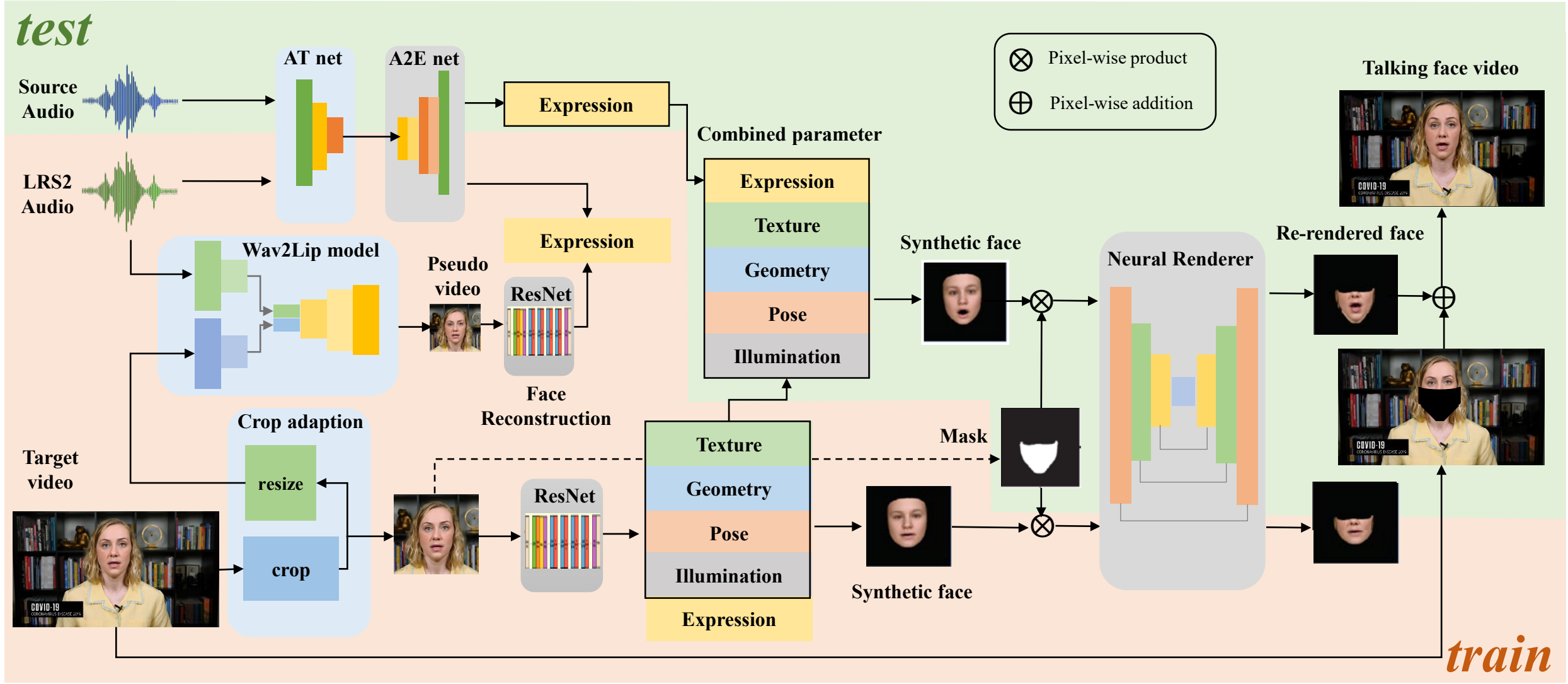
# Motivation

- AudioDVP [1]: an audio-driven 3D talking face video generation model
- Wav2Lip [2]: a 2D speech-to-lip model

| Method | Training data | Definition of generated face | Lip-sync of generated video |
|---|---|---|---|
| AudioDVP | 3 min high-definition video | Limited-definition (256x256) | limited |
| AudioDVP | 5∼6h high-definition video | Limited-definition (256x256) | ✓ |
| Wav2Lip | Over 30h low-definition video | Low-definition (96x96) | ✓ |
| ★ Our method | 3 min high-definition video | High-definition (1024x1024) | ✓ |

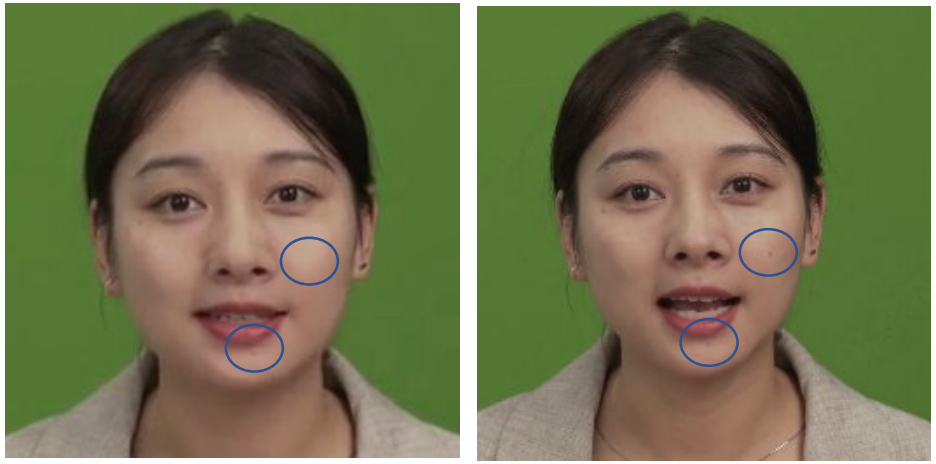
[1] XinWen, MiaoWang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu, "Photorealistic audio-driven video portraits," IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 12, pp. 3457–3466, 2020.
[2] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in Proceedings of the 28th ACM International Conference on Multimedia, New York, NY, USA, 2020, MM '20, p. 484–492, Association for Computing Machinery.

# Framework

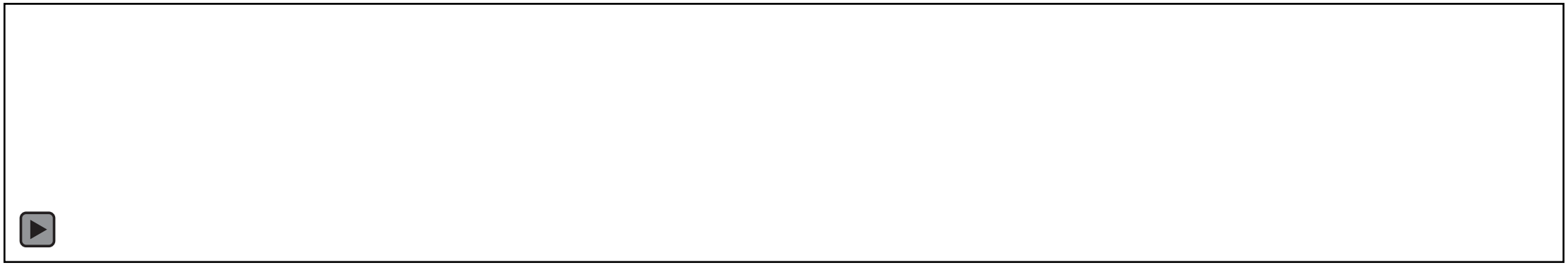

test
Source Audio
LRS2 Audio
AT net
A2E net
Expression
Expression
Wav2Lip model
Pseudo video
ResNet
Face Reconstruction
Combined parameter
Expression
Texture
Geometry
Pose
Illumination
Synthetic face
Pixel-wise product
Pixel-wise addition
Talking face video
Neural Renderer
Re-rendered face
Mask
Crop adaption
resize
crop
Target video
ResNet
Texture
Geometry
Pose
Illumination
Expression
Synthetic face
train

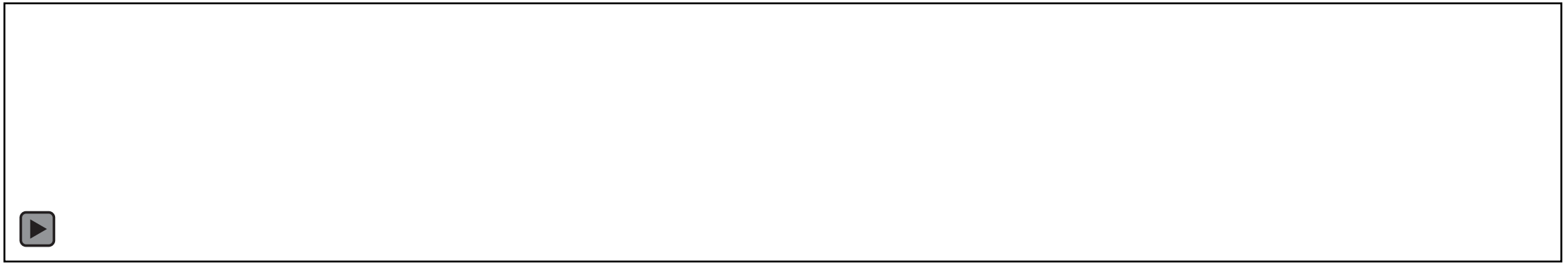
# Experiment


using audio in the wild to compare our method with ATVG, Wav2lip, AudioDVP and MakeIttalk



AudioDVP Ours





# Experiment

**Table 2.** Quantitative evaluation on the test sets of videos. For LSE-D and FID the lower the better, and the higher the better for LSE-C and SSIM.

| | Methods | ATVG | Wav2Lip | AudioDVP | MakeIttalk | Ours |
|---|---|---|---|---|---|---|
| A | LSE-D↓ | 9.114 | 7.756 | 10.195 | 9.977 | 8.636 |
| | LSE-C↑ | 5.653 | 7.555 | 4.138 | 4.716 | 6.060 |
| | FID↓ | 21.572 | 11.847 | 9.437 | 23.158 | 6.734 |
| | SSIM↑ | 0.5298 | 0.6072 | 0.9490 | 0.5526 | 0.9832 |
| B | LSE-D↓ | 10.400 | 7.540 | 14.978 | 11.911 | 9.878 |
| | LSE-C↑ | 5.234 | 6.066 | 0.238 | 2.135 | 4.963 |
| | FID↓ | 19.983 | 13.120 | 10.234 | 19.315 | 7.065 |
| | SSIM↑ | 0.6721 | 0.6049 | 0.9645 | 0.6238 | 0.9896 |
| C | LSE-D↓ | 10.581 | 6.637 | 11.712 | 16.170 | 9.530 |
| | LSE-C↑ | 5.122 | 8.951 | 3.322 | 0.06 | 6.141 |
| | FID↓ | 19.311 | 11.154 | 9.677 | 19.46 | 6.498 |
| | SSIM↑ | 0.350 | 0.5647 | 0.9316 | 0.4781 | 0.9744 |
| D | LSE-D↓ | 10.005 | 6.546 | 11.804 | 11.444 | 9.091 |
| | LSE-C↑ | 5.808 | 9.023 | 2.884 | 3.725 | 6.155 |
| | FID↓ | 17.969 | 12.485 | 9.076 | 19.93 | 6.881 |
| | SSIM↑ | 0.5682 | 0.5766 | 0.9439 | 0.4705 | 0.9875 |
| E | LSE-D↓ | 12.506 | 6.571 | 11.713 | 14.339 | 9.831 |
| | LSE-C↑ | 2.734 | 8.989 | 3.063 | 1.018 | 5.493 |
| | FID↓ | 18.697 | 11.185 | 9.523 | 19.502 | 7.131 |
| | SSIM↑ | 0.5716 | 0.6482 | 0.9237 | 0.4872 | 0.9867 |
| F | LSE-D↓ | 9.567 | 6.343 | 9.953 | 11.167 | 8.817 |
| | LSE-C↑ | 5.803 | 9.314 | 4.794 | 4.277 | 5.841 |
| | FID↓ | 20.839 | 13.457 | 11.348 | 17.746 | 6.775 |
| | SSIM↑ | 0.5321 | 0.7015 | 0.9577 | 0.4764 | 0.9894 |
| G | LSE-D↓ | 9.687 | 6.013 | 13.332 | 8.831 | 8.880 |
| | LSE-C↑ | 7.261 | 10.237 | 2.074 | 6.795 | 7.539 |
| | FID↓ | 21.894 | 14.94 | 10.795 | 23.880 | 6.579 |
| | SSIM↑ | 0.5794 | 0.6102 | 0.9102 | 0.7736 | 0.9866 |


**Table 3.** Ablation study on the test sets of videos.

| Video | LSE-D | LSE-C | FID | SSIM |
|---|---|---|---|---|
| Baseline | 11.955 | 2.930 | 10.007 | 0.940 |
| Our A2E | 9.578 | 6.343 | 9.264 | 0.952 |
| High definition | 11.834 | 3.427 | 7.449 | 0.976 |
| Ours | 9.237 | 6.027 | 6.804 | 0.985 |

**Table 4.** User study results

| Method | Average | → Rating of realistic and definition → | | | | | synchronization |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 'sync' |
| ATVG | 1.98 | 32.1% | 42.9% | 19.6% | 5.4% | 0.0% | 73.2% |
| Wav2Lip | 3.23 | 0.0% | 26.8% | 30.4% | 35.7% | 7.1% | 76.7% |
| AudioDVP | 3.16 | 1.8% | 16.1% | 51.7% | 25.0% | 5.4% | 26.7% |
| MakeIttalk | 2.07 | 33.9% | 30.4% | 30.4% | 5.3% | 0.0% | 32.1% |
| Ours | **4.45** | 0.0% | 0.0% | 12.5% | 30.4% | 57.1% | 83.9% |

# Conclusion

## Contribution:

- The low-definition pseudo video predicted by Wav2Lip with the target video and LRS2 audio is introduced to enhance the audio-driven identity-disentangled ability of talking face generation.
- We train a modified audio-to-expression (A2E) network to guarantee the accurate lip motion driven by arbitrary audio, which makes our method possess an powerful audio-driven performance comparable to Wav2Lip.
- A modified crop module is introduced for automatically adapting the size of the 3DMM synthetic face to the original face area, then enabled our framework to meet the requirements of 4K-definition photo-realistic talking face video.

## Future work:

- talking face generation based on target identity disentanglement.

Thank you