JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

High-fidelity Generalized Emotional Talking Face Generation with Multi-modal Emotion Space Learning

Poster Session: TUE-PM-237

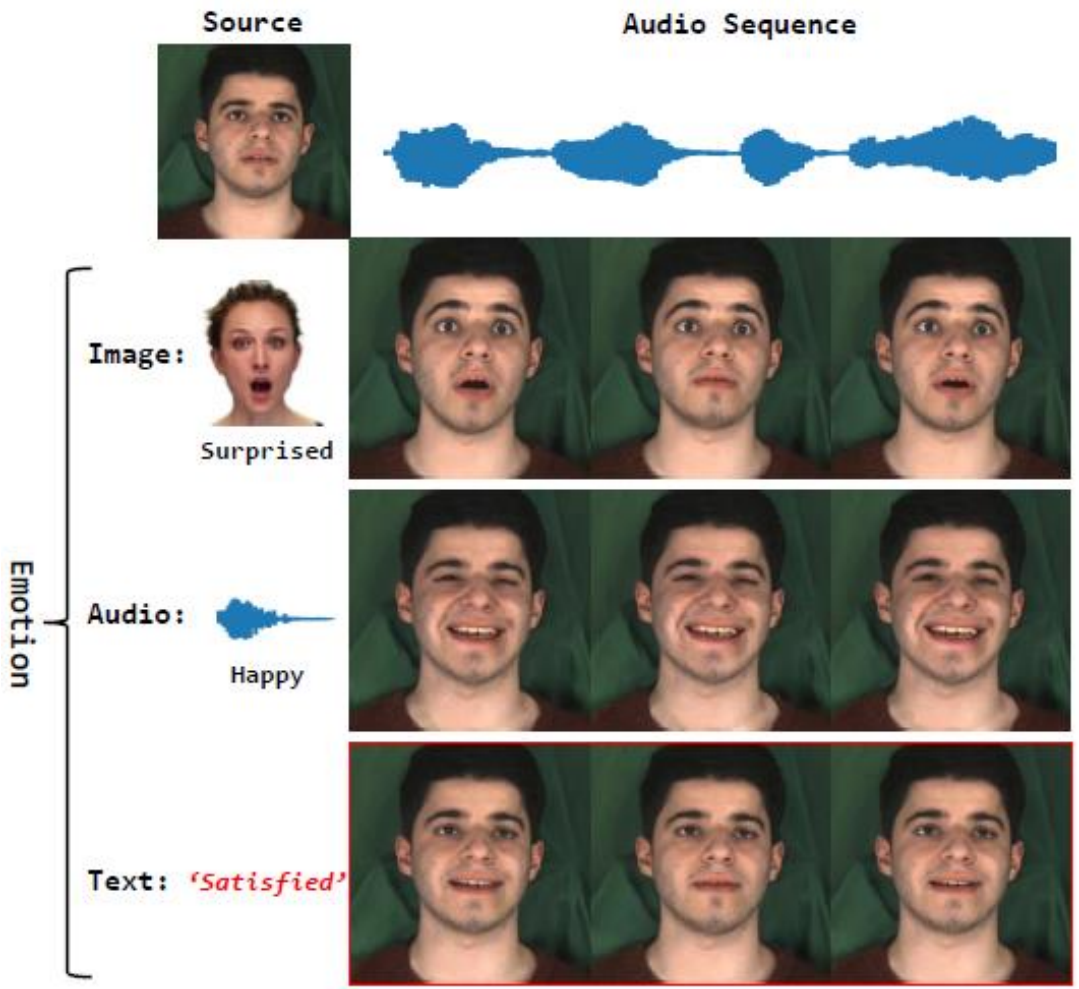
Chao Xu[1] Junwei Zhu[2] Jiangning Zhang[2] Yue Han[1] Wenqing Chu[2]
Ying Tai[2] Chengjie Wang[2,4] Zhifeng Xie[3] Yong Liu[1]

[1]Zhejiang University [2]Youtu Lab, Tencent
[3]Shanghai University [4]Shanghai Jiao Tong University

腾讯优图

APRIL机器人智能感知与学习实验室
Advanced Perception on Robotics and Intelligent Learning Lab

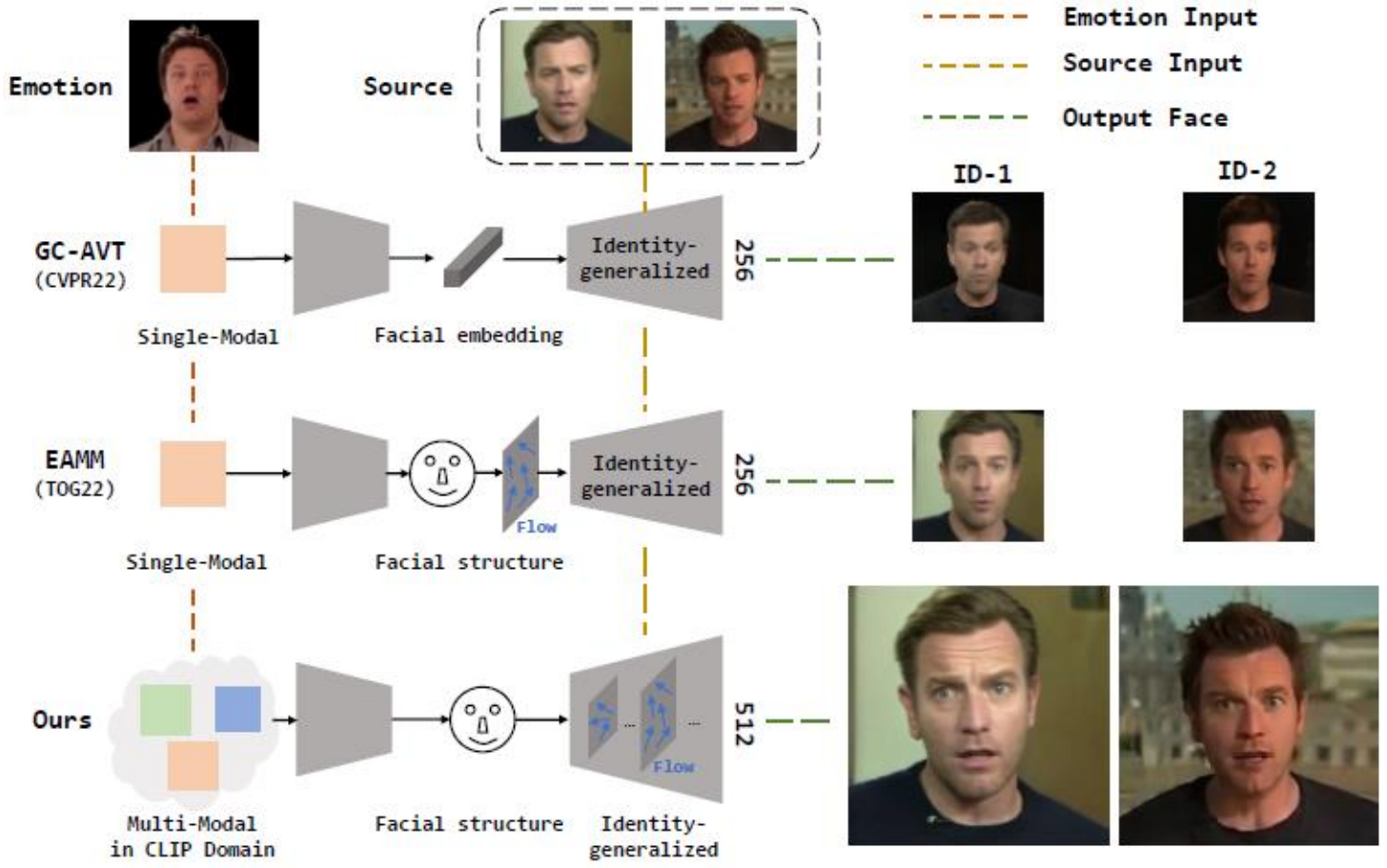
0. Summary



Contributions:

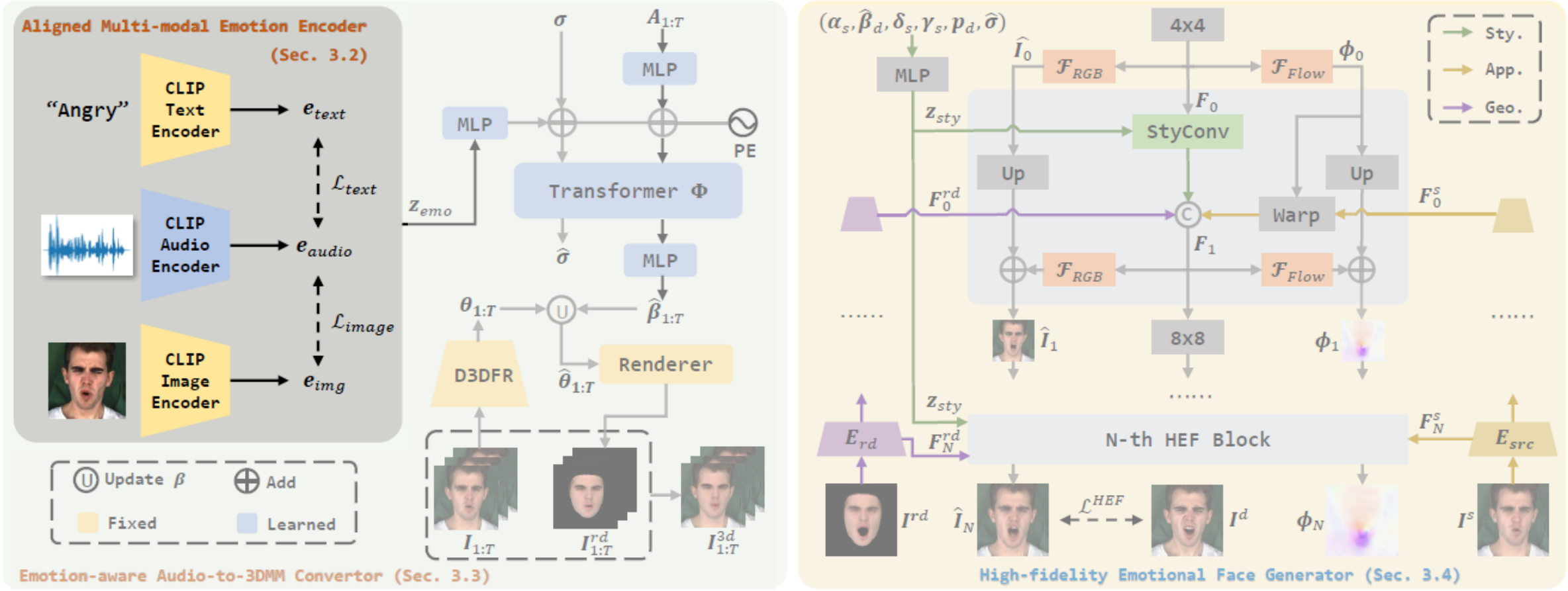
- We propose a novel AME that provides a unified multimodal semantic-rich emotion space, allowing flexible emotion control and unseen emotion generalization, which is the first attempt in this field.
- We propose a novel HEF to hierarchically learn the facial deformation by sufficiently modeling the interaction among emotion, source appearance, and drive geometry for the high-resolution one-shot generation.
- Abundant experiments are conducted to demonstrate the superiority of our method for flexible and generalized emotion control, and high-resolution one-shot talking face animation over SOTA methods.

1. Introduction

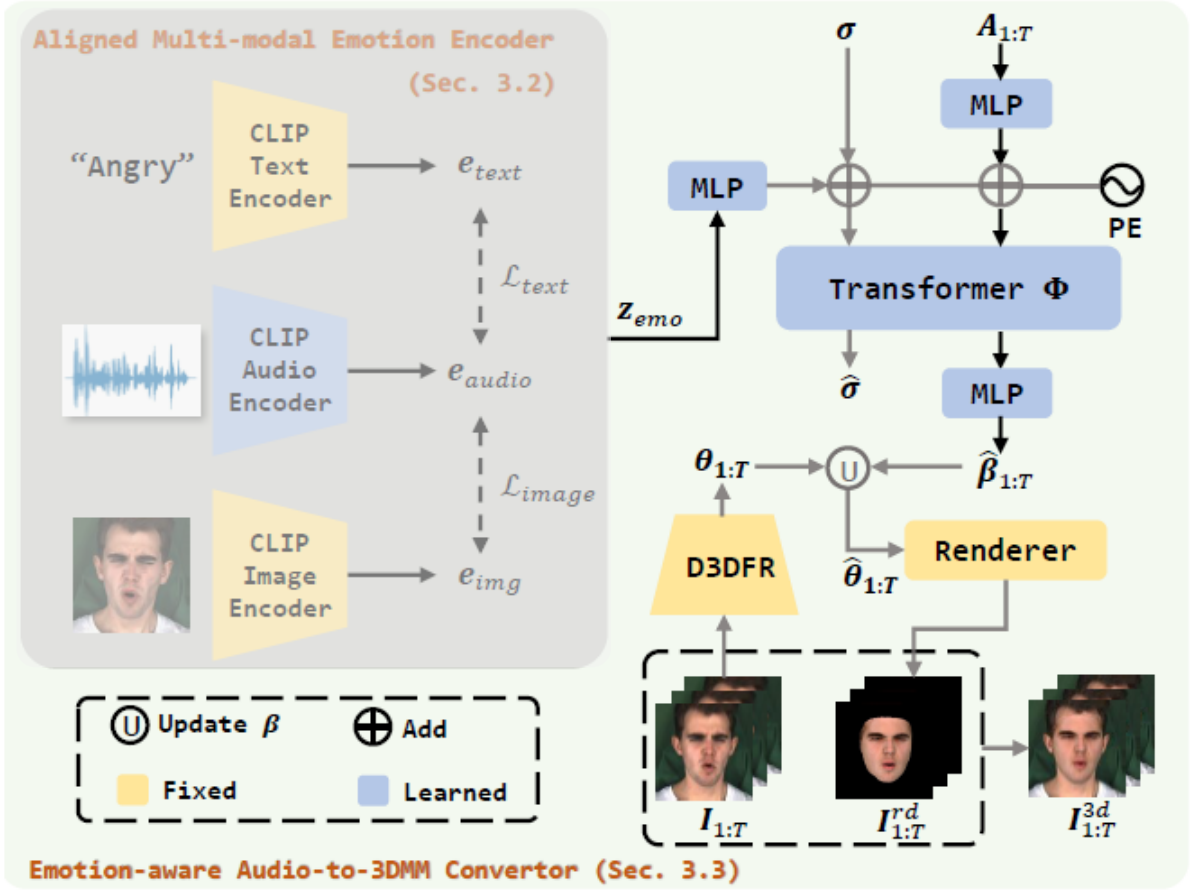


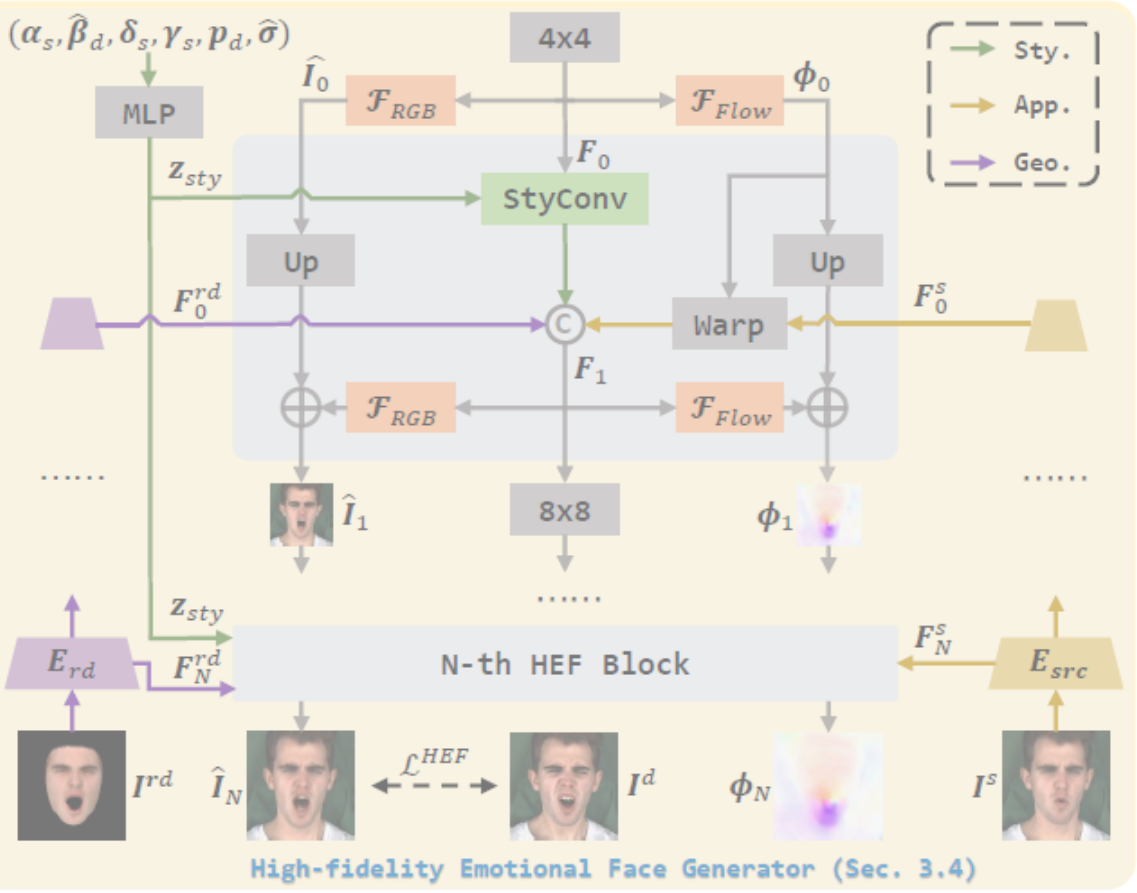
- How to explore a more semantic emotion embedding to achieve better generalization for unseen emotions.
- Could we construct multi-modal emotion sources into a unified feature space to allow a more flexible and user-friendly emotion control
- How to design a high-resolution identity-generalized generator

1. Method - AME

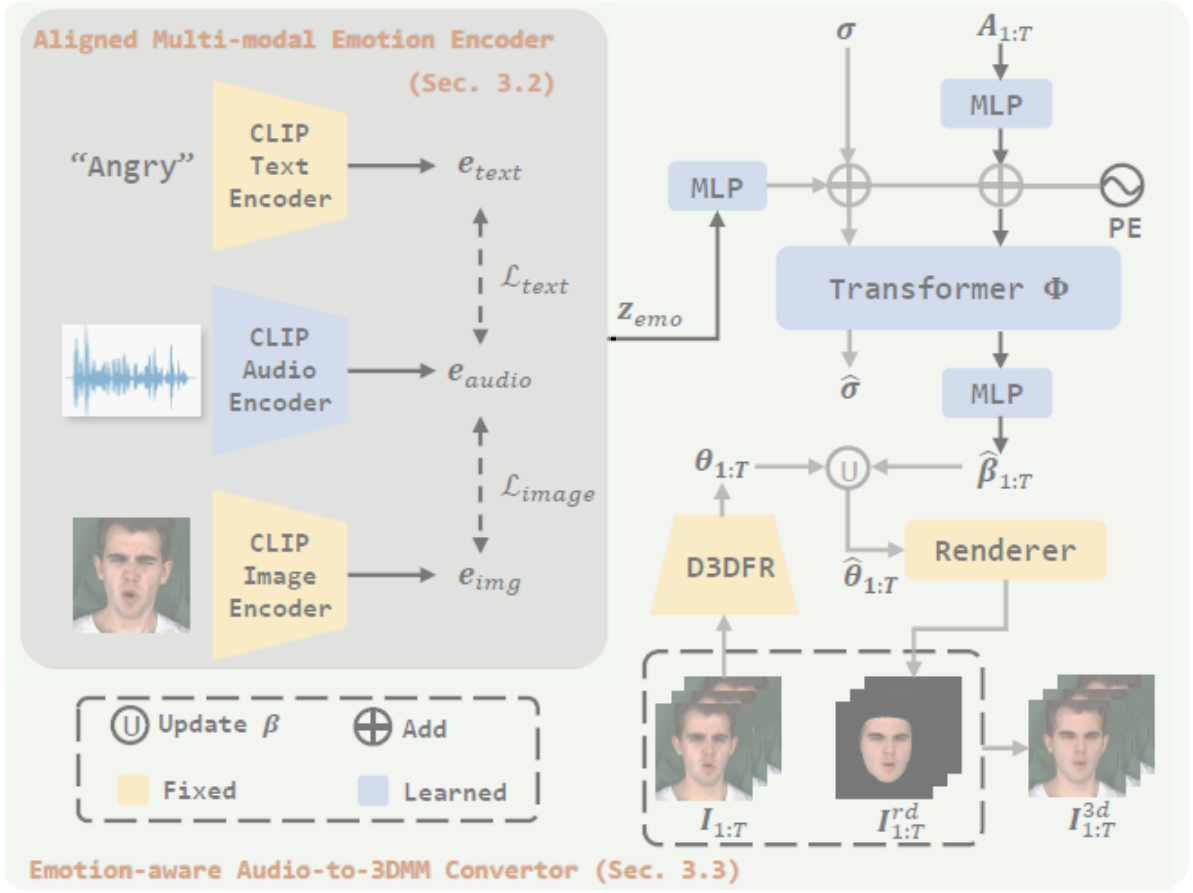

Aligned Multi-modal Emotion Encoder
(Sec. 3.2)
"Angry"
CLIP Text Encoder
CLIP Audio Encoder
CLIP Image Encoder
e_{text}
e_{audio}
e_{img}
$\mathcal{L}_{text}$
$\mathcal{L}_{image}$
σ
$A_{1:T}$
MLP
PE
Transformer Φ
z_{emo}
$\hat{\sigma}$
$\theta_{1:T}$
$\hat{\beta}_{1:T}$
D3DFR
$\hat{\theta}_{1:T}$
Renderer
Update β
Add
Fixed
Learned
$I_{1:T}$
$I^{rd}_{1:T}$
$I^{3d}_{1:T}$
Emotion-aware Audio-to-3DMM Convertor (Sec. 3.3)
$(\alpha_s, \hat{\beta}_d, \delta_s, \gamma_s, p_d, \hat{\sigma})$
MLP
z_{sty}
$\hat{I}_0$
$\mathcal{F}_{RGB}$
4x4
$\mathcal{F}_{Flow}$
ϕ_0
F_0
StyConv
Up
Up
F^{rd}_0
Warp
F^s_0
F_1
$\hat{I}_1$
8x8
ϕ_1
Sty.
App.
Geo.
E_{rd}
F^{rd}_N
N-th HEF Block
F^s_N
E_{src}
I^{rd}
$\hat{I}_N$
$\mathcal{L}^{HEF}$
I^d
ϕ_N
I^s
High-fidelity Emotional Face Generator (Sec. 3.4)

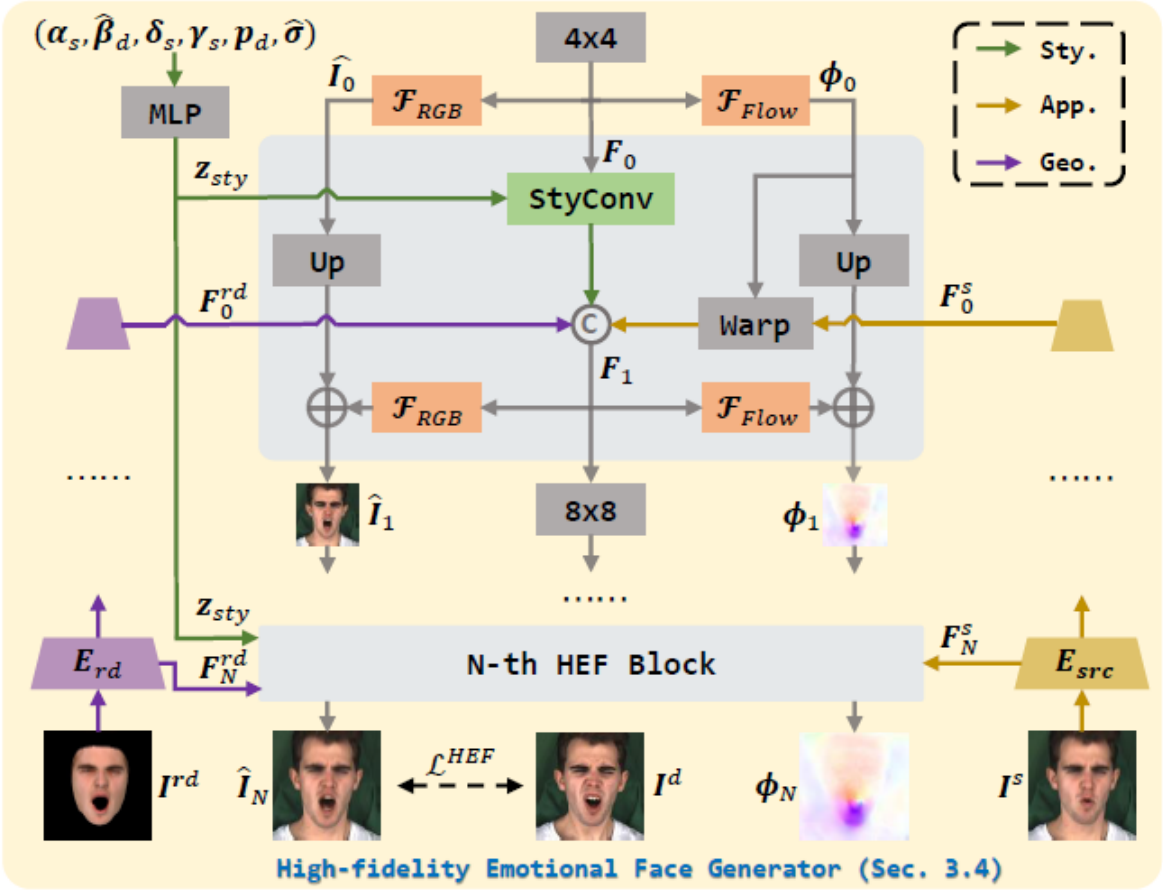
1. Method - EAC


Aligned Multi-modal Emotion Encoder
(Sec. 3.2)
"Angry"
CLIP Text Encoder
CLIP Audio Encoder
CLIP Image Encoder
e_{text}
e_{audio}
e_{img}
$\mathcal{L}_{text}$
$\mathcal{L}_{image}$
σ
$A_{1:T}$
MLP
MLP
PE
Transformer Φ
z_{emo}
$\hat{\sigma}$
MLP
$\theta_{1:T}$
U
$\hat{\beta}_{1:T}$
D3DFR
$\hat{\theta}_{1:T}$
Renderer
Update β
Add
Fixed
Learned
$I_{1:T}$
$I^{rd}_{1:T}$
$I^{3d}_{1:T}$
Emotion-aware Audio-to-3DMM Convertor (Sec. 3.3)


$(\alpha_s, \hat{\beta}_d, \delta_s, \gamma_s, p_d, \hat{\sigma})$
MLP
z_{sty}
4x4
$\hat{I}_0$
$\mathcal{F}_{RGB}$
$\mathcal{F}_{Flow}$
ϕ_0
F_0
StyConv
Sty.
App.
Geo.
Up
Up
F^{rd}_0
F^s_0
C
Warp
F_1
$\mathcal{F}_{RGB}$
$\mathcal{F}_{Flow}$
.....
$\hat{I}_1$
8x8
ϕ_1
.....
z_{sty}
E_{rd}
F^{rd}_N
N-th HEF Block
F^s_N
E_{src}
I^{rd}
$\hat{I}_N$
$\mathcal{L}^{HEF}$
I^d
ϕ_N
I^s
High-fidelity Emotional Face Generator (Sec. 3.4)

1. Method - HEF


Aligned Multi-modal Emotion Encoder
(Sec. 3.2)
"Angry"
CLIP Text Encoder
CLIP Audio Encoder
CLIP Image Encoder
e_{text}
e_{audio}
e_{img}
$\mathcal{L}_{text}$
$\mathcal{L}_{image}$
σ
$A_{1:T}$
MLP
PE
Transformer Φ
z_{emo}
$\hat{\sigma}$
$\theta_{1:T}$
$\hat{\beta}_{1:T}$
D3DFR
$\hat{\theta}_{1:T}$
Renderer
Update β
Add
Fixed
Learned
$I_{1:T}$
$I^{rd}_{1:T}$
$I^{3d}_{1:T}$
Emotion-aware Audio-to-3DMM Convertor (Sec. 3.3)


$(\alpha_s, \hat{\beta}_d, \delta_s, \gamma_s, p_d, \hat{\sigma})$
MLP
4x4
$\hat{I}_0$
$\mathcal{F}_{RGB}$
$\mathcal{F}_{Flow}$
ϕ_0
Sty.
App.
Geo.
z_{sty}
F_0
StyConv
Up
Up
F^{rd}_0
Warp
F^s_0
F_1
$\hat{I}_1$
8x8
ϕ_1
E_{rd}
F^{rd}_N
N-th HEF Block
F^s_N
E_{src}
I^{rd}
$\hat{I}_N$
$\mathcal{L}^{HEF}$
I^d
ϕ_N
I^s
High-fidelity Emotional Face Generator (Sec. 3.4)

2. Experiments – Qualitative Results

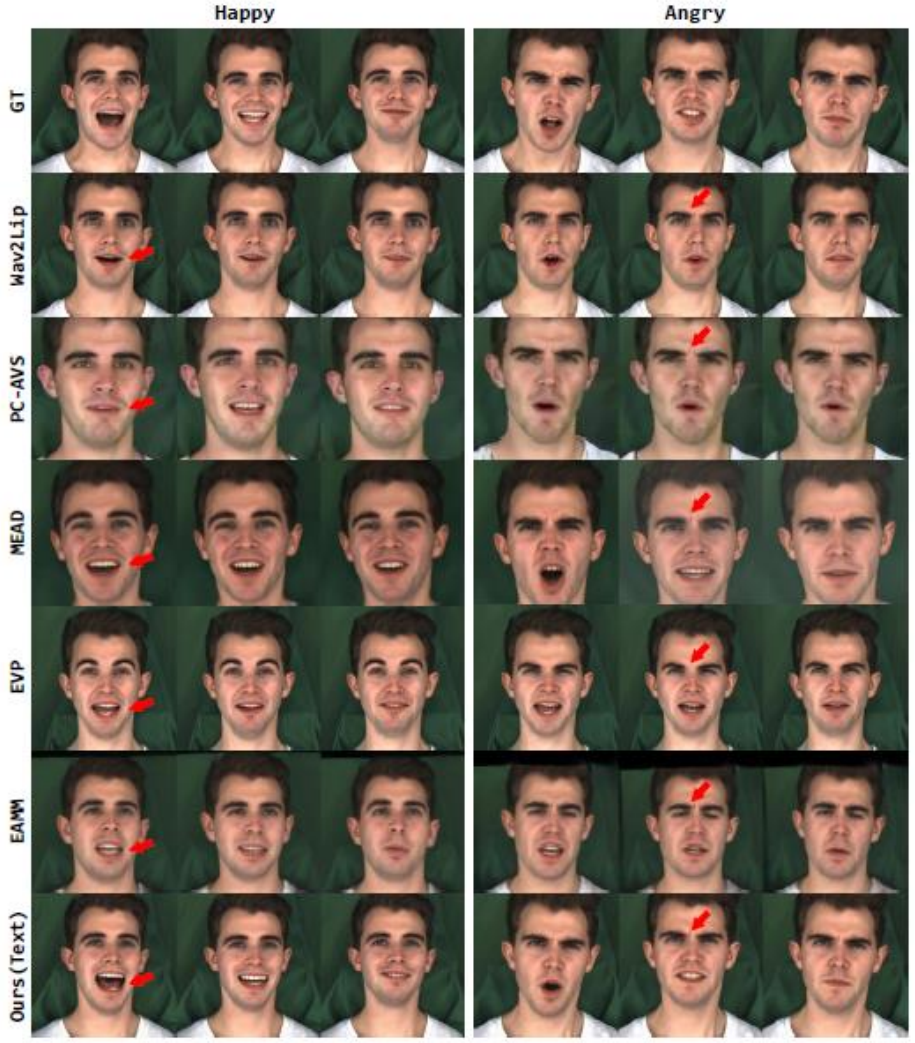


Figure 3. Qualitative results on MEAD dataset. Different columns mean several sampled timestamps (*same as the following figures*). Images are from officially released codes for fair comparisons.

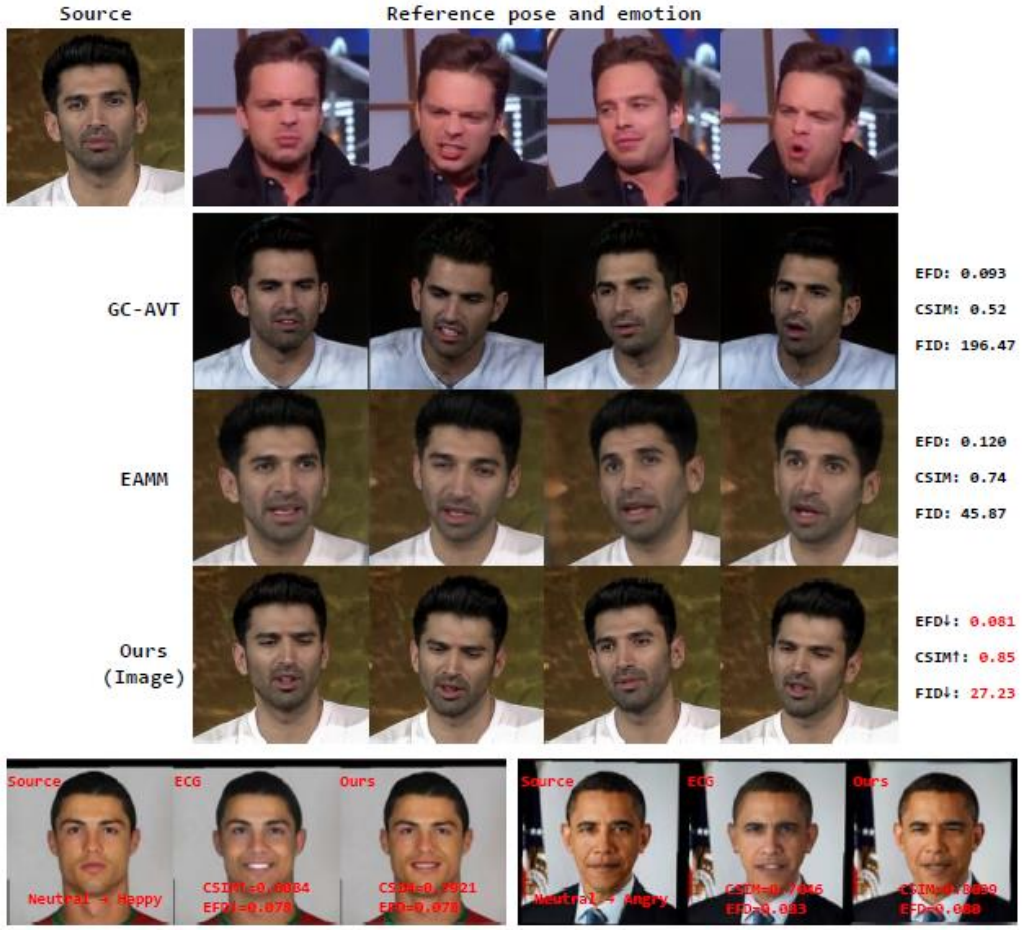


Figure 4. Qualitative comparison with GC-AVT, EAMM, and ECG. The top part is sampled from Fig. 3 of GC-AVT. The bottom part is sampled from Fig. 4 of ECG. Quantitative results of these cases are attached in the figure. We ignore the metric of mouth shape because the audios for these sequences are not available.

2. Experiments – Quantitative Results

Method	EFD ↓	LMD ↓	Sync ↑	CSIM ↑	FID ↓	PSNR ↑	SSIM ↑
Wav2Lip	0.112	2.59	3.26	0.82	20.15	29.22	0.70
PC-AVS	0.110	2.68	3.12	0.80	29.55	28.97	0.68
MEAD	0.084	2.62	3.09	0.81	30.69	28.48	0.65
EVP	0.106	2.54	3.21	0.70	**12.83**	29.67	0.73
EAMM	0.092	2.50	3.26	0.74	29.01	29.33	0.75
Ours-A	0.069	2.36	3.50	0.83	15.91	30.09	0.85
Ours-I	0.071	2.36	3.53	**0.84**	15.89	**30.10**	0.87
Ours-T	**0.065**	**2.31**	**3.57**	**0.84**	15.90	30.09	**0.88**

Table 1. Quantitative comparison on MEAD dataset. Ours-A, -I, and -T mean audio, image, and text, respectively.

2. Experiments – Analysis

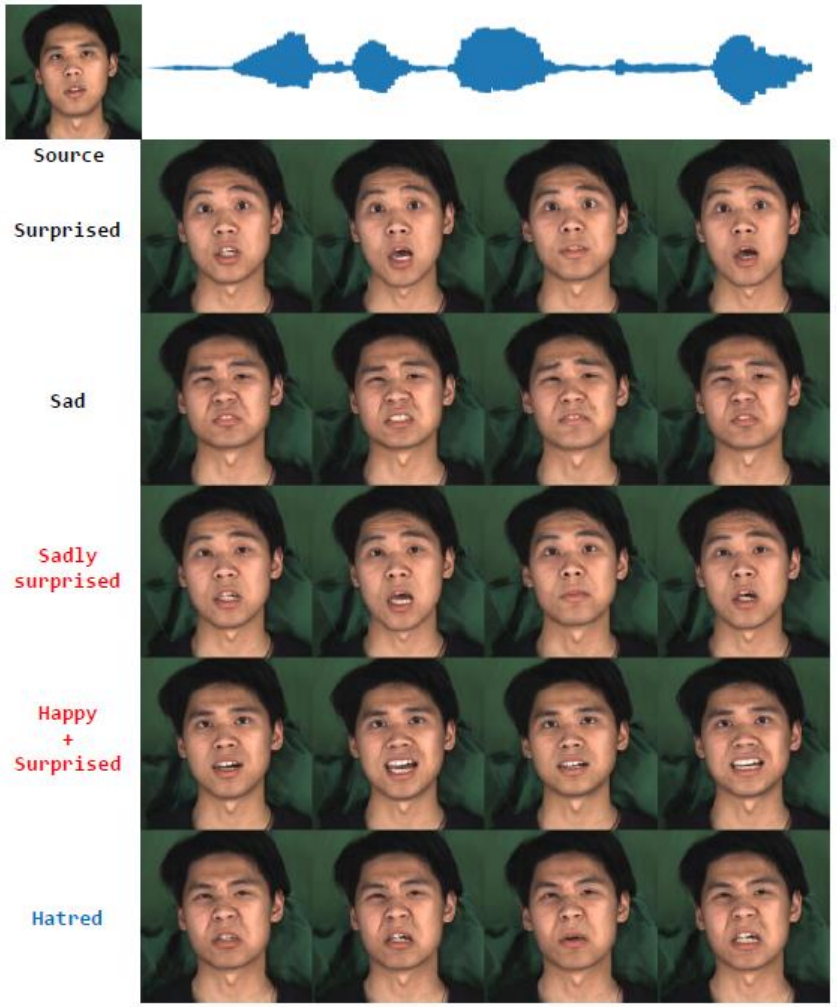


Figure 5. Results of unseen emotion styles. Rows 4 and 5 (in Red) are the compound styles, and row 6 (in Blue) is a totally new style.

Figure 6. Results of unseen identity. We visualize the rendered images, final outputs, and predicted flow fields. The color wheel of flow fields is attached on the top right for reference.

2. Experiments – Analysis

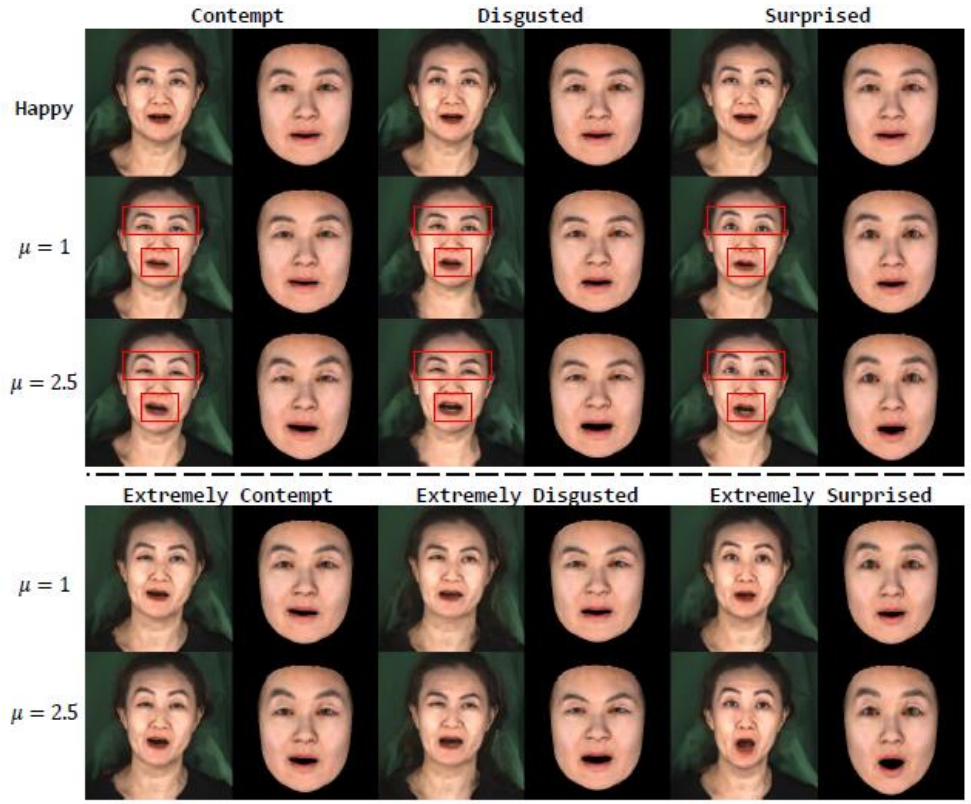


Figure 7. Results of different emotion styles and intensity levels. The top part shows the manipulation from the happy to three distinct emotion styles. The bottom part shows the results of style semantics that already encode intensity.

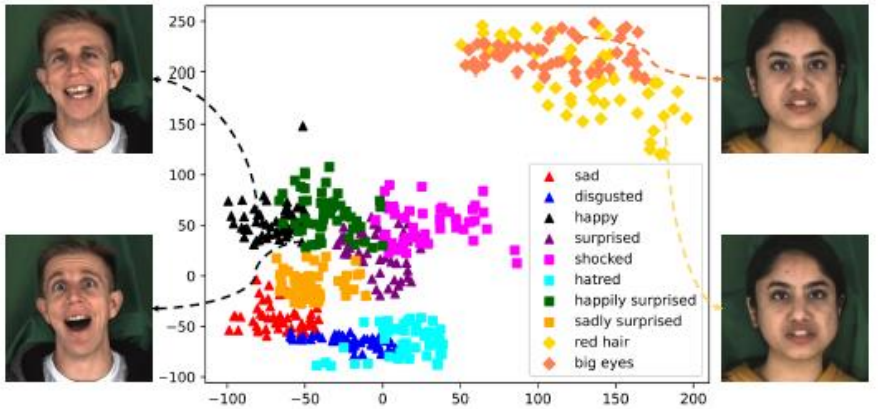


Figure 8. Clusters of the intensity token with the emotion and emotion-unrelated text descriptions. Markers △, □, ◊ mean basic emotion, unseen emotion, and emotion-unrelated text prompts.

2. Experiments – Ablation Study

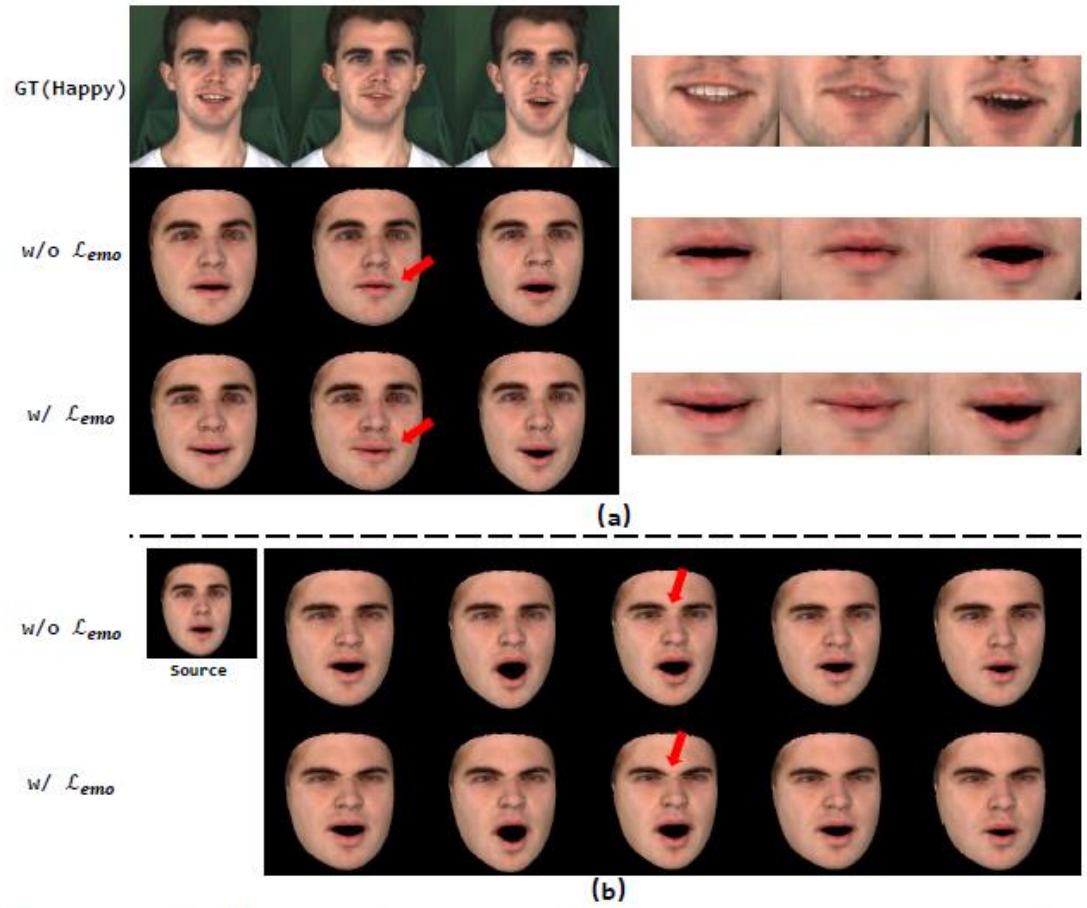


Figure 9. Qualitative ablation study for emotion consistency loss of EAC. (a) shows the effect on 3D face reconstruction of *happy* emotion and (b) illustrates manipulation by *angry* emotion.

2. Experiments – Ablation Study

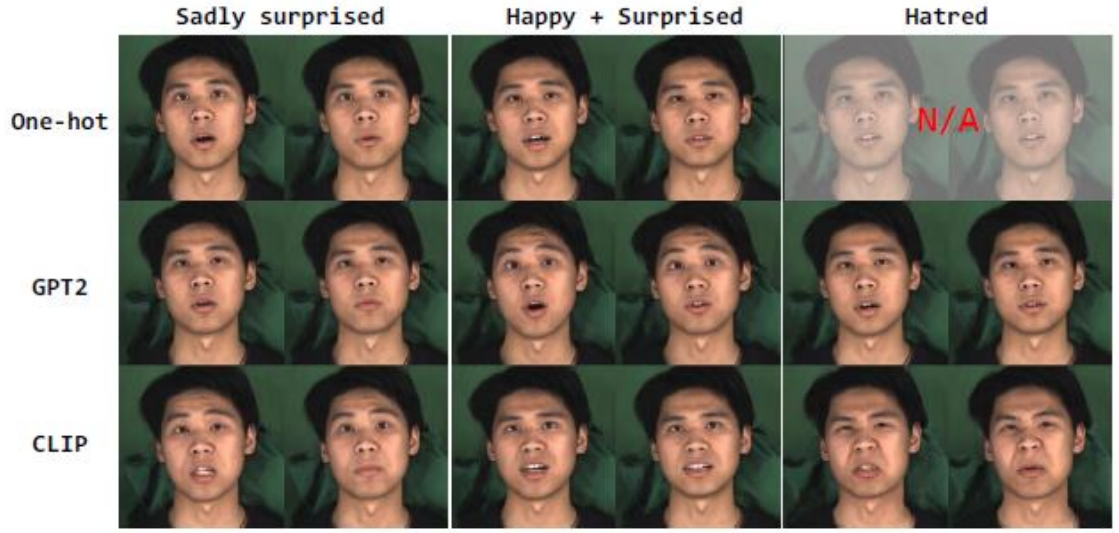


Figure 10. Qualitative ablation study of EAC with different emotion encodings on *unseen styles*. This case is sampled from Fig. 5.

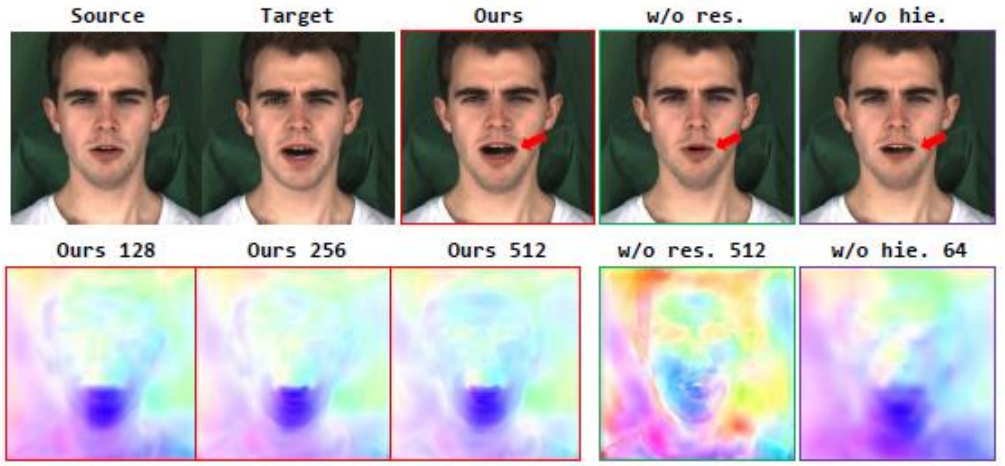


Figure 11. Qualitative ablation study of HEF with different flow estimation variants. We visualize the flow fields of w/o res. at scale 512 and the fixed 64×64 flow fields of w/o hie., both fail to model the precise movement, while our method gradually refines the high-resolution flow fields by hierarchical residual learning.

2. Experiments – Ablation Study

Method	EFD ↓	LMD ↓	Sync ↑
w/o $\mathcal{L}_{emo}$	0.096	2.40	3.53
w/ $\mathcal{L}_{emo}$	**0.065**	**2.31**	**3.57**
One-hot	0.070	2.33	3.53
GPT2	0.067	2.33	3.56
CLIP	**0.065**	**2.31**	**3.57**
MLPs	0.122	3.54	2.23
GRUs	0.088	2.47	3.19
Transformers	**0.065**	**2.31**	**3.57**
w/o res.	0.082	2.46	3.21
w/o hie.	0.076	2.42	3.25
Ours	**0.065**	**2.31**	**3.57**

Table 3. Quantitative ablation study with different losses and components, conducted on MEAD with *basic styles* by default.

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Thanks for your listening!

Website: april.zju.edu.cn

APRIL机器人智能感知与学习实验室
Advanced Perception on Robotics and Intelligent Learning Lab