Expressive Talking Head Video Encoding in StyleGAN2 Latent Space

Trevine Oorloff and Yaser Yacoob
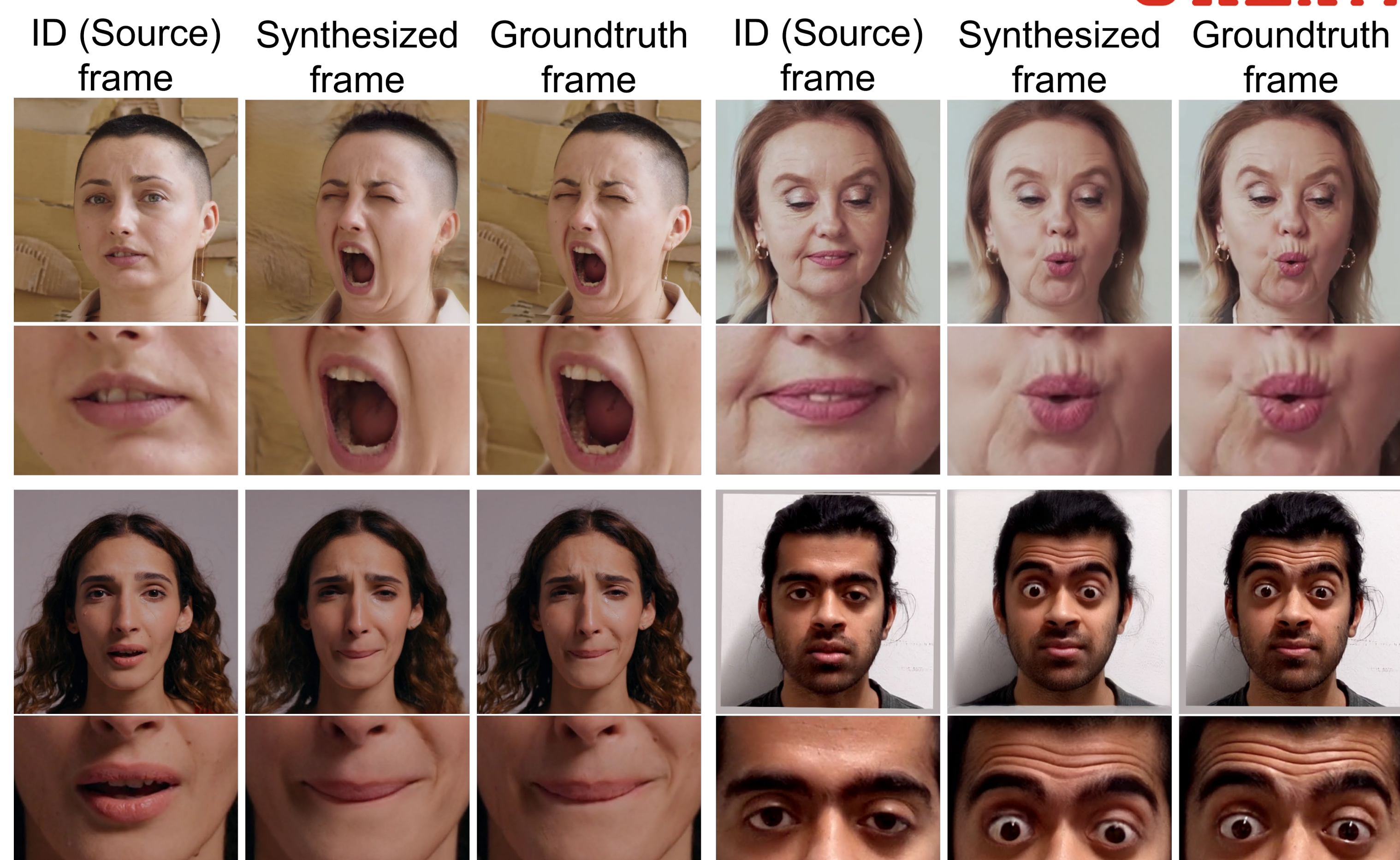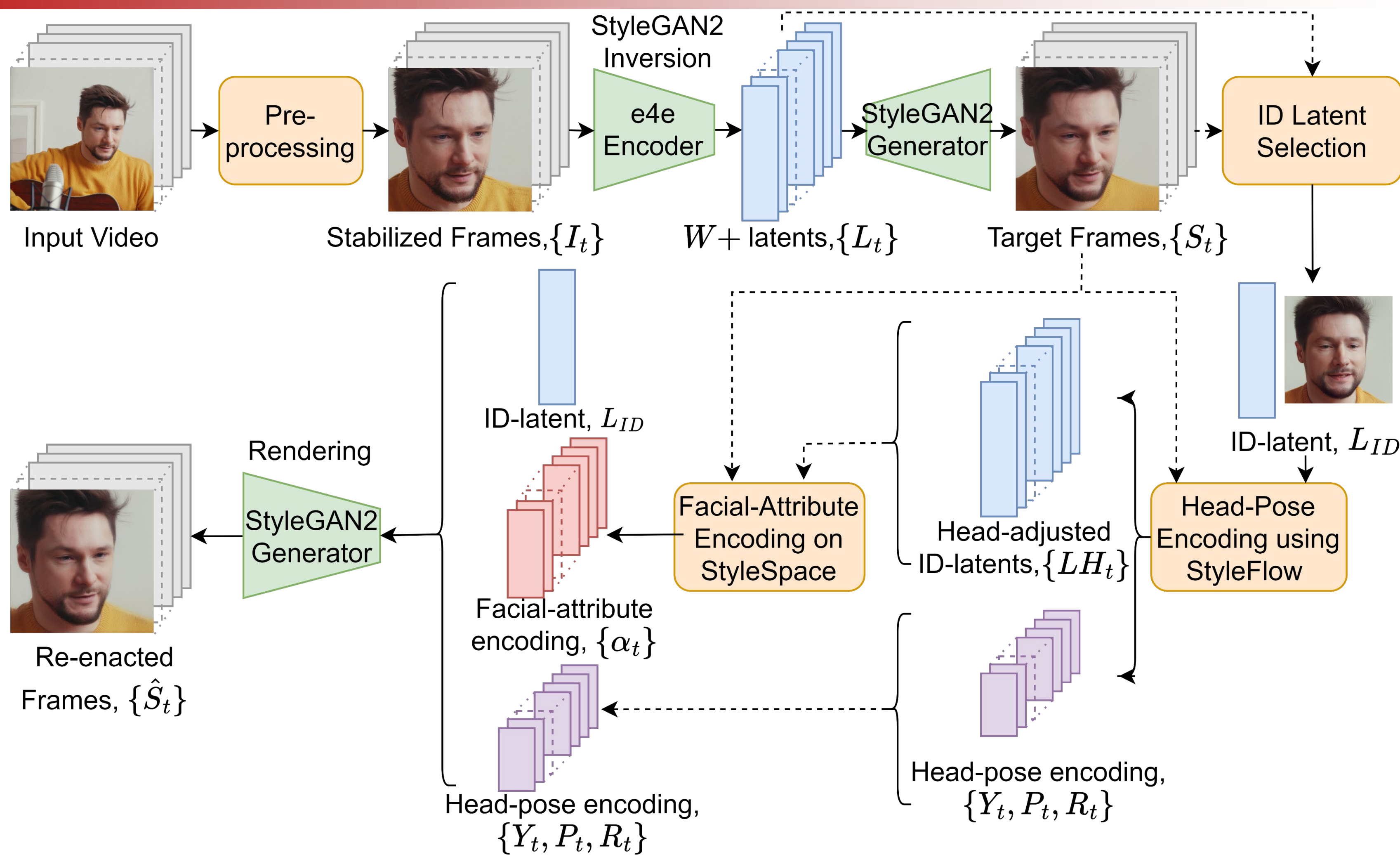University of Maryland, College Park, Maryland, USA

ICCV23 PARIS
CVEU

Motivation

☐ Existing face video re-enactment approaches fall short in capturing the fine, detailed, and expressive facial features (*e.g., lip-pressing, mouth puckering, mouth gaping, and wrinkles*) and/or are limited to low resolution

☐ Being able to capture such fine expressive facial features at high-resolution is crucial towards improving the realism of re-enacted face video

☐ On the other hand, existing StyleGAN2 latent-based editing techniques focus on simply generating plausible edits of static images

☐ We bridge the gap between high-fidelity static portrait image synthesis/manipulation and face re-enactment of intense expressions and speech

ID (Source) frame | Synthesized frame | Groundtruth frame | ID (Source) frame | Synthesized frame | Groundtruth frame

Pipeline

Input Video → Pre-processing → Stabilized Frames, $\{I_t\}$ → StyleGAN2 Inversion e4e Encoder → $W+$ latents, $\{L_t\}$ → StyleGAN2 Generator → Target Frames, $\{S_t\}$ → ID Latent Selection → ID-latent, L_{ID}

Rendering StyleGAN2 Generator → Re-enacted Frames, $\{\hat{S}_t\}$

ID-latent, L_{ID} → Facial-Attribute Encoding on StyleSpace → Facial-attribute encoding, $\{\alpha_t\}$

Head-adjusted ID-latents, $\{LH_t\}$

Head-Pose Encoding using StyleFlow → ID-latent, L_{ID}

Head-pose encoding, $\{Y_t, P_t, R_t\}$

Head-pose encoding, $\{Y_t, P_t, R_t\}$

Main Contributions

☐ A novel StyleGAN2-based algorithm for high-resolution (1024^2) face video encoding for re-synthesis and puppeteering with emphasis on precise reconstruction of both expressive and talking facial attributes in contrast to common models that do not focus on fine/complex facial details

☐ A novel approach that employs image inversion and sparse latent space editing to produce an extremely compact encoding scheme capturing head-pose and facial attribute deformations (*i.e., a single ID-latent and 35 parameters/frame which amounts to merely 0.38% of the StyleGAN2 latent space*)

☐ An approach that automates the latent-space editing process to capture facial attribute deformations in contrast to prevailing work on latent-space editing that simply illustrate plausible semantic visual results

☐ A novel method to find StyleSpace channels corresponding to facial attributes based on index sensitivity

Quantitative Results

Re-Synthesis

Method	res.	L1 ↓	LPIPS↓	L_{ID}↓	PSNR↑	SSIM↑	FID↓	FVD↓	ρ_{AU}↑	ρ_{gz}↑	ρ_{pose}↑
FOMM	256^2	3.07	0.036	0.174	31.0	0.932	28.7	140.3	0.710	0.755	0.648
LIA	256^2	3.24	0.042	0.164	30.0	0.929	30.2	162.9	0.546	0.693	0.619
FS-Vid2Vid	512^2	5.75	0.093	0.158	25.2	0.900	42.4	359.6	0.571	0.784	0.629
StyleHEAT	1024^2	4.13	0.097	0.134	27.6	0.933	25.1	281.9	0.673	0.701	0.763
Ours	1024^2	1.99	0.030	0.097	34.2	0.963	15.9	85.2	0.771	0.834	0.880
StyleVideoGAN*	1024^2	4.04	0.109	0.104	28.8	0.926	28.8	223.3	0.739	0.884	0.979
Ours*	1024^2	1.96	0.026	0.067	34.1	0.960	13.6	79.8	0.899	0.971	0.987

Puppeteering

Method	res.	L_{ID}↓	FID↓	FVD↓	FVD$_\pi$↓	ρ_{AU+gz}↑
FOMM	256^2	0.153	77.0	396.8	103.0	0.501
LIA	256^2	0.174	82.3	406.0	112.4	0.527
FS-Vid2Vid	512^2	0.202	73.6	445.1	112.7	0.640
StyleHEAT	1024^2	0.181	81.0	437.5	109.8	0.667
Ours	1024^2	0.095	63.9	386.5	82.3	0.708

Qualitative Results

Re-Synthesis

FS-Vid2Vid | FOMM | LIA | StyleHEAT | **OURS** | Groundtruth

Puppeteering

Puppeteer | Puppet