Latent Image Animator: Learning to Animate Images via Latent Space Navigation

ICLR 2022

Yaohui Wang, Di Yang, Francois Bremond and Antitza Dantcheva

Inria

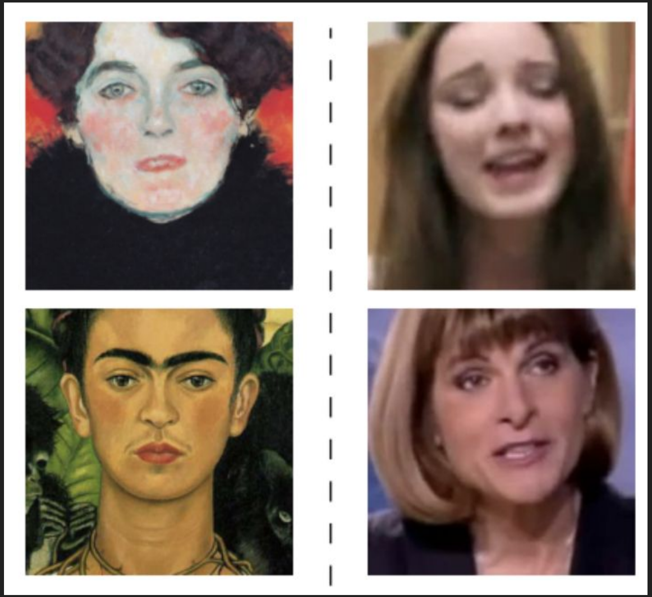
Presenter: Jaeseong Lee

Computer Vision Seminar
21 Feb 2022

Contents

Preliminary(cont'd)

Neural Talking Head: Reenacting driving images(or a video)' **Facial Expression & Head Rotation** to a source image



source

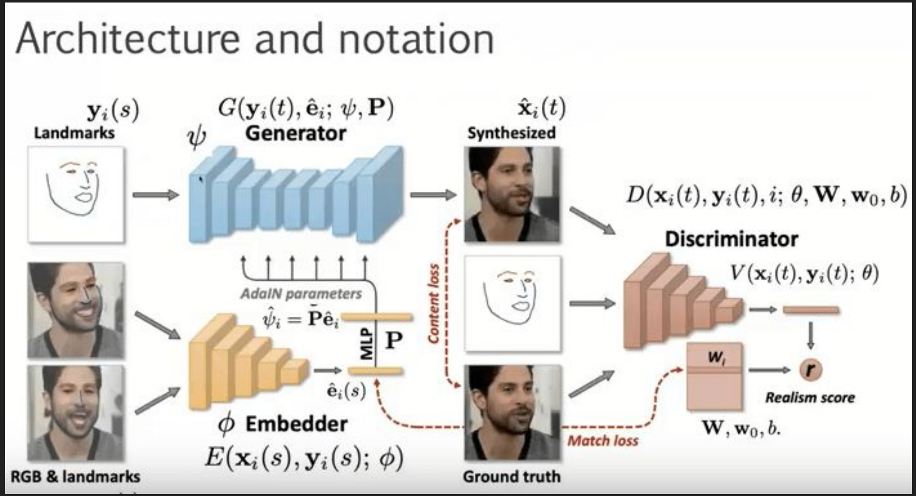
driving

reenactment
(retargeting)

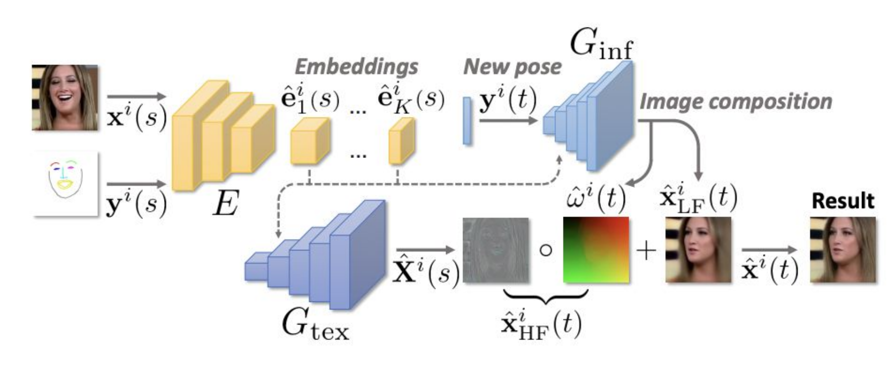
result

Preliminary(cont’d)

off-the-shelf
keypoints-based



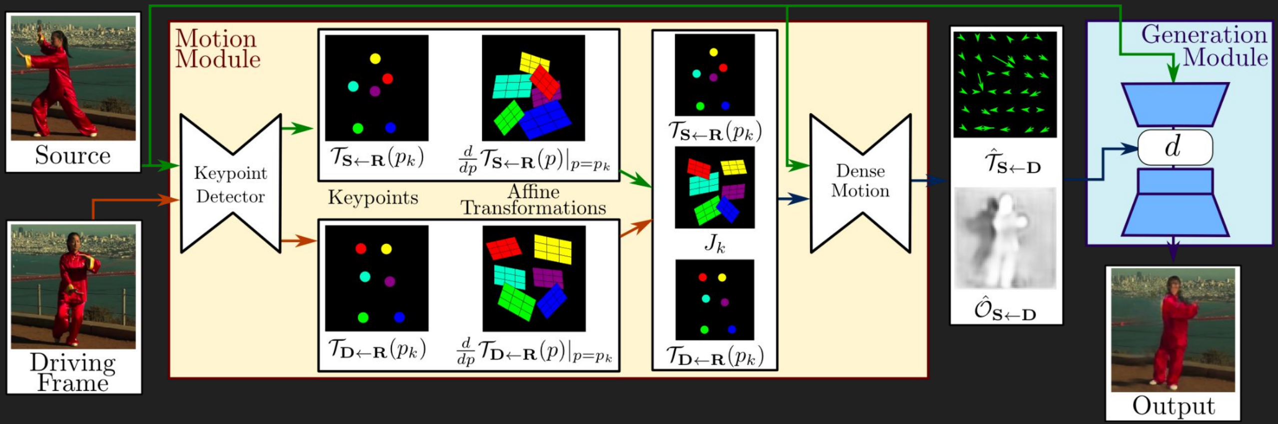
Few-Shot Adversarial Learning of
Realistic Neural Talking Head Models
(ICCV 2019)
Samsung AI, Moscow



Fast Bi-layer Neural Synthesis of
One-Shot Realistic Head Avatars
(ECCV 2020)
Samsung AI, Moscow

Preliminary(cont'd)

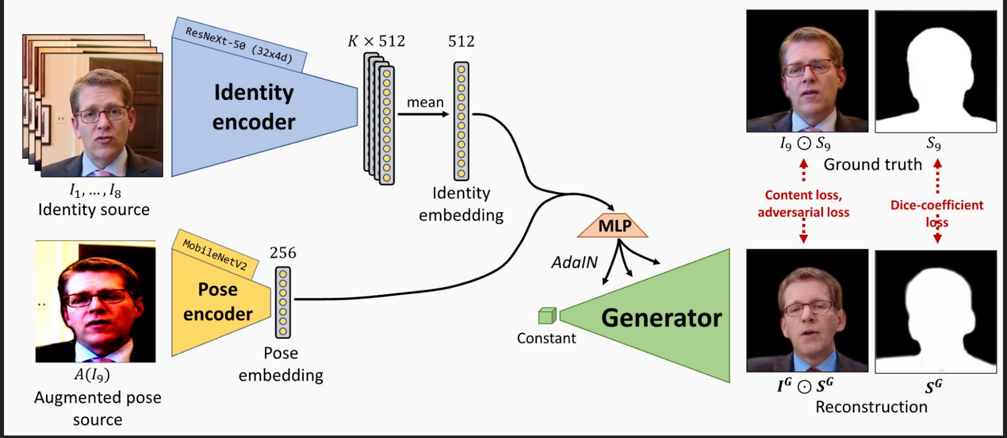
self-sup
keypoints-based



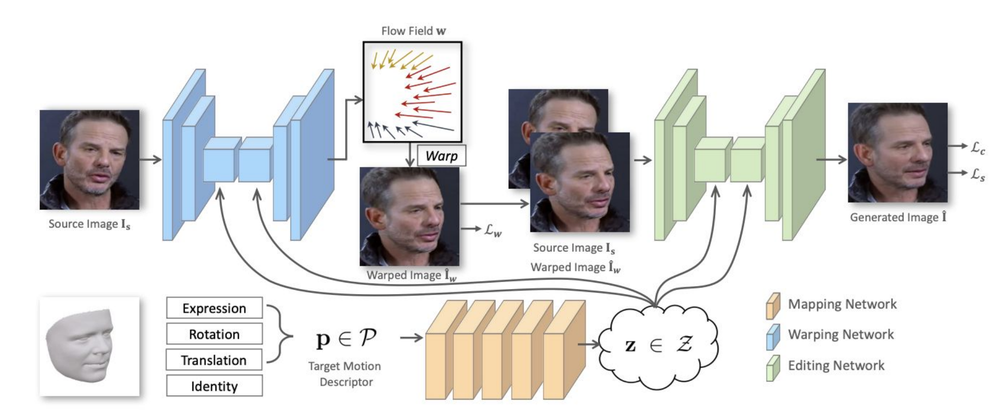
First Order Motion Model for Image Animation
(NeurIPS 2019)
University of Trento

Preliminary

pose
descriptor-based
(AdaIN)



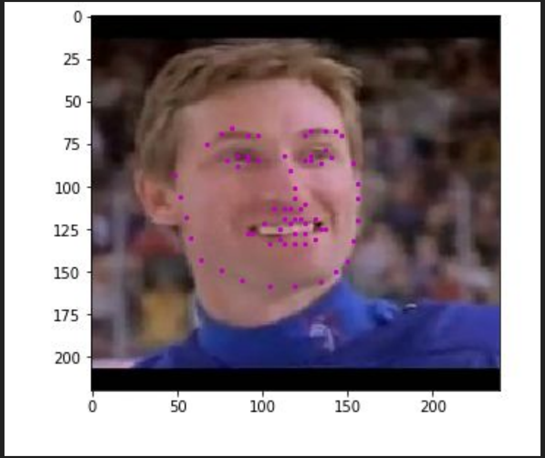
Neural Head Reenactment with Latent Pose Descriptors
(CVPR 2020)
Samsung AI, Moscow



PIRenderer: Controllable Portrait Image Generation via Semantic Neural Rendering
(ICCV 2021)
Peking University

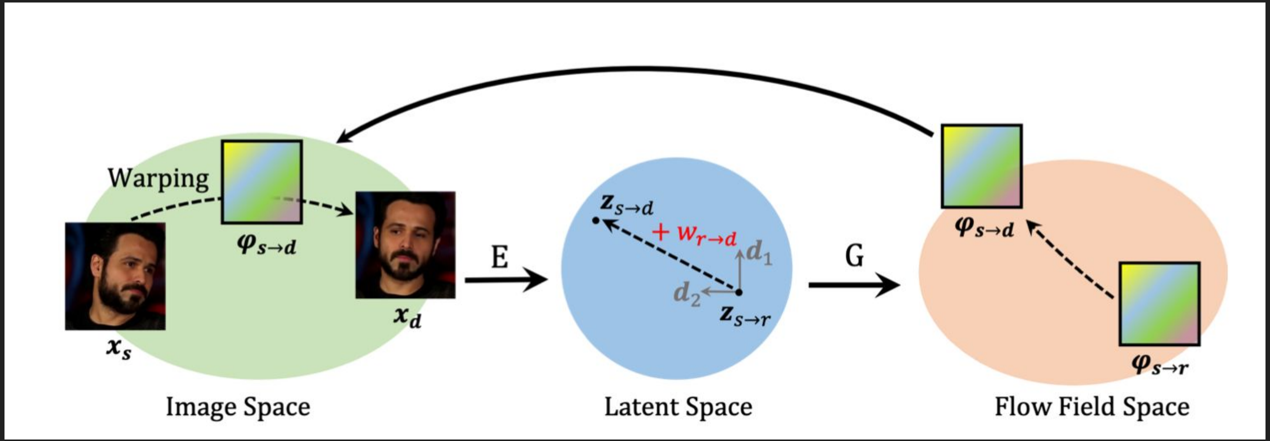
Introduction(cont'd)

- The poses constraints(e.g., keypoint, 3DMM parameters) on applications, where such representations of unseen testing images might be **fragmentary or missing**.



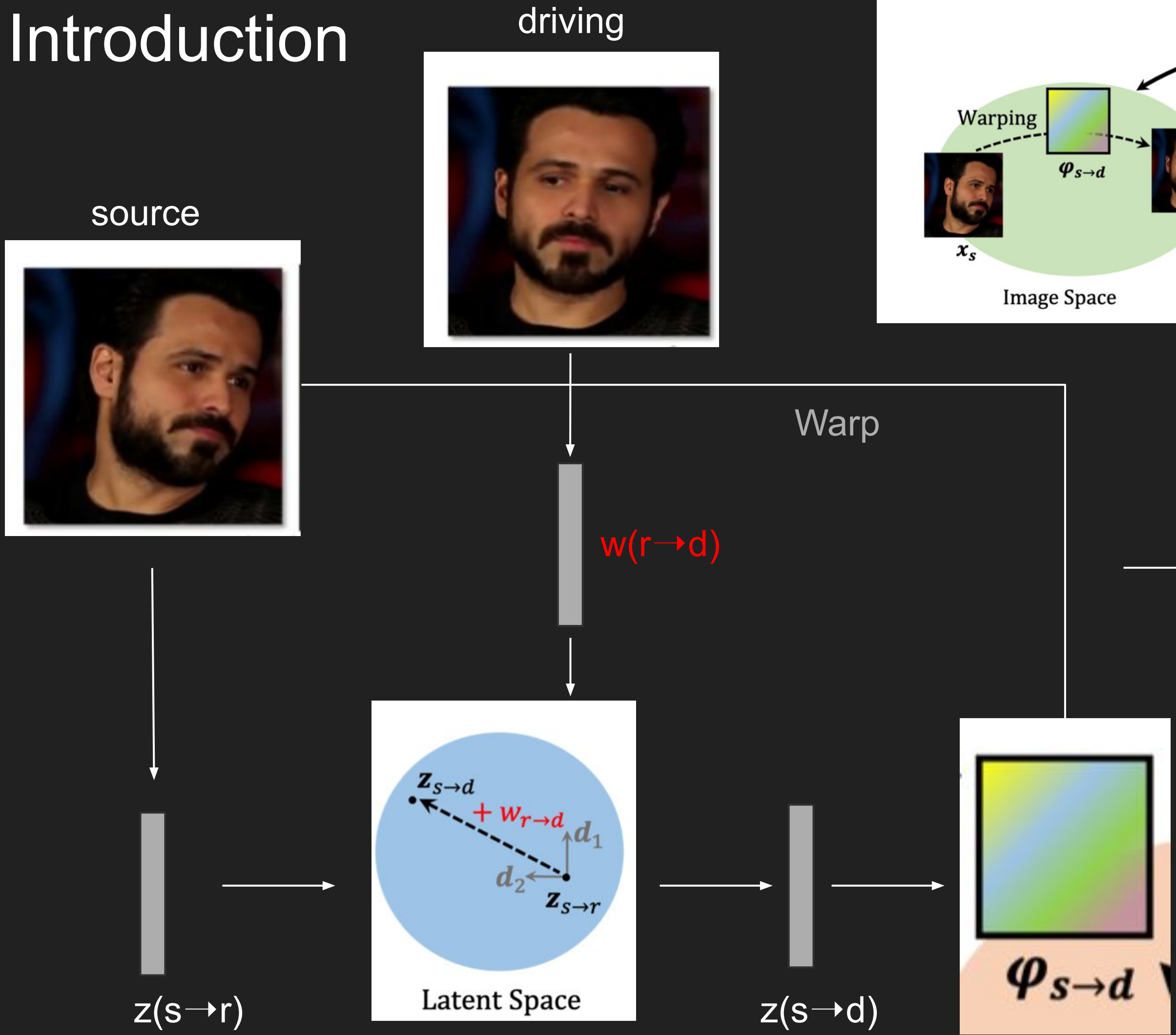
Introduction(cont'd)

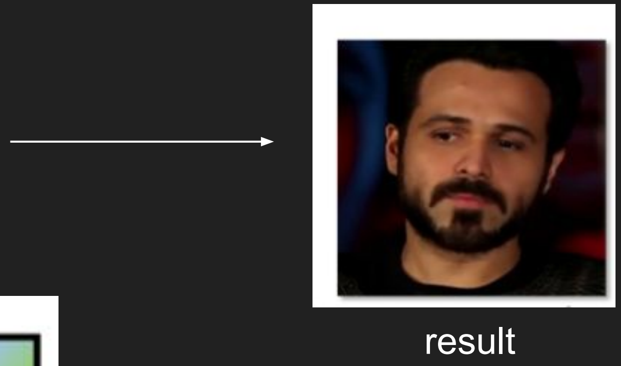
- Existing methods using structural information(e.g., keypoints or descriptors) are hard to perform well when the source and the driving have **large appearance variation**. ⏩ **Latent Space Navigating**

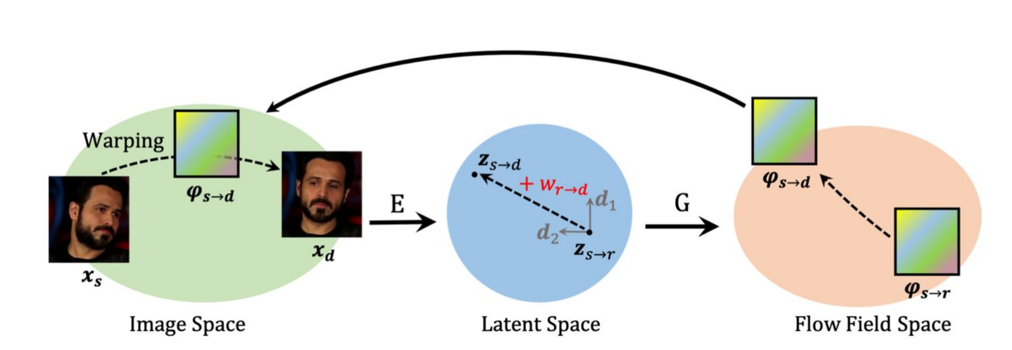


General pipeline of **LIA**

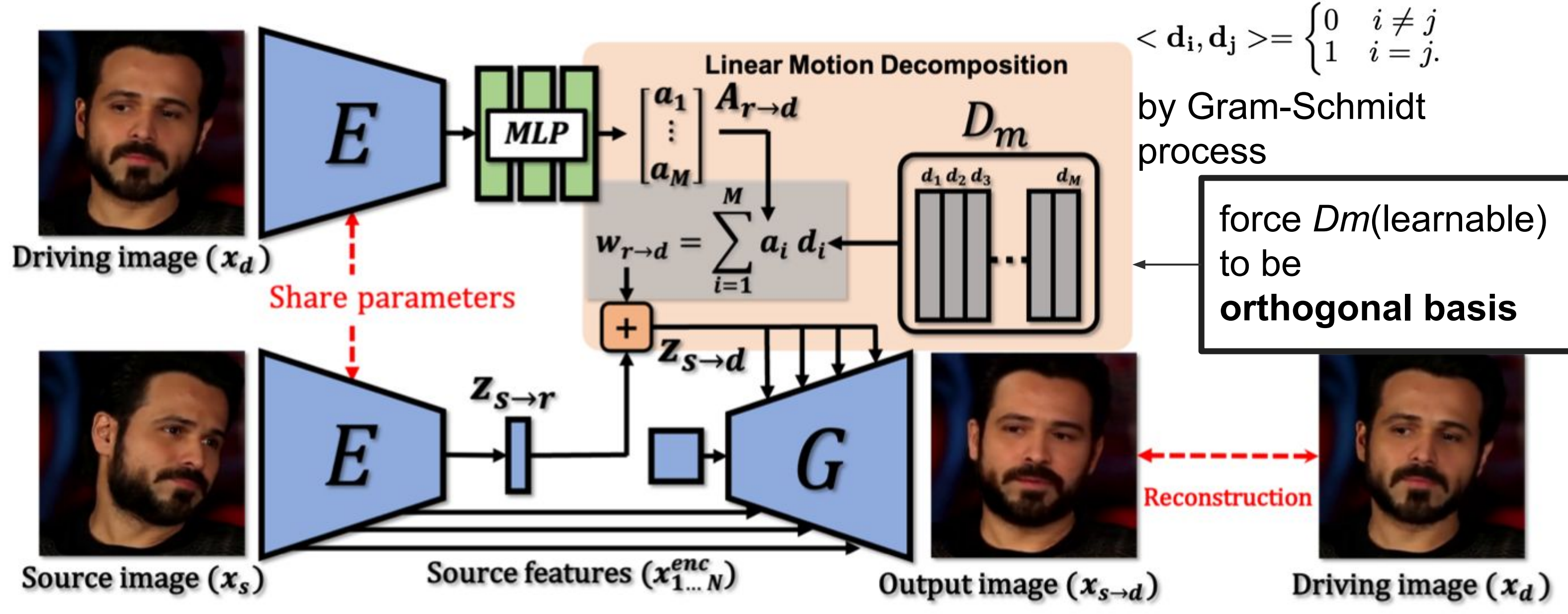
Introduction







Method(cont'd)



$$< \mathbf{d_i}, \mathbf{d_j} >= \begin{cases} 0 & i \neq j \\ 1 & i = j. \end{cases}$$

by Gram-Schmidt process

force *Dm*(learnable) to be **orthogonal basis**

Objectives:

$$\mathcal{L}_{recon}(x_{s\to d}, x_d) = \mathbb{E}[\|x_d - x_{s\to d}\|_1].$$

$$\mathcal{L}_{vgg}(x_{s\to d}, x_d) = \mathbb{E}[\sum_{n}^{N} \|F_n(x_d) - F_n(x_{s\to d})\|_1],$$

$$\mathcal{L}_{adv}(x_{s\to d}) = \mathbb{E}_{x_{s\to d}\sim p_{rec}}[-log(D(x_{s\to d})],$$

Method

Goal

$$z_{s\to t} = (z_{s\to r} + w_{r\to s}) + (w_{r\to t} - w_{r\to 1})$$

when,s!=1

@ Inference time → 'relative transfer'

$$= z_{s\to s} + (w_{r\to t} - w_{r\to 1}),\ t \in \{1, ..., T\}.$$

- z(s→s): reconstruction
- w(r→t) - w(r→1): motion from 1 to t

* The more similar btw xs & x1, the more replication

when, s=1

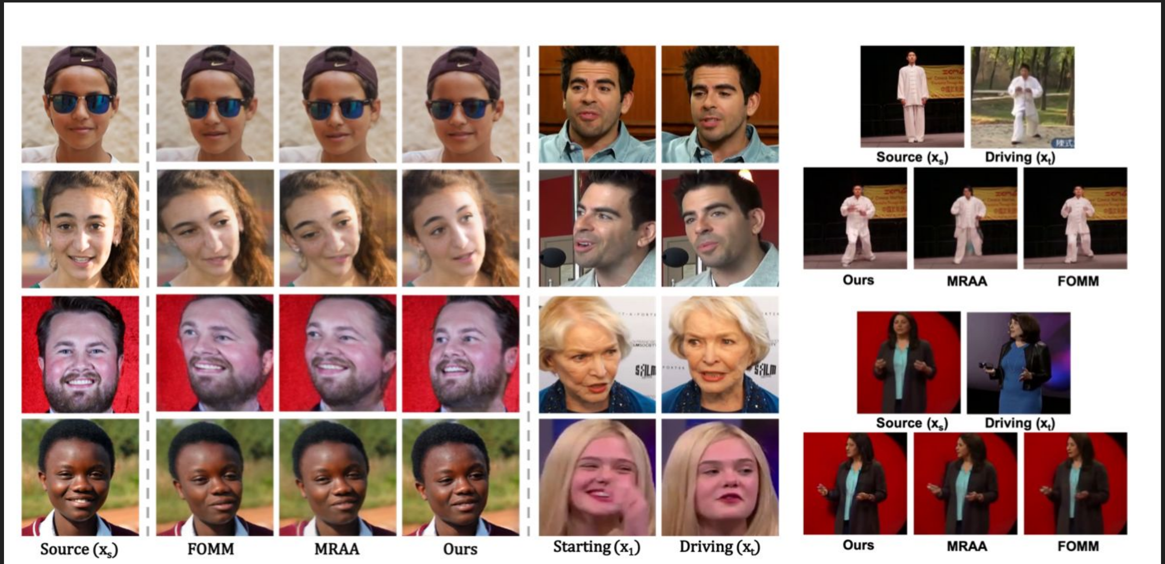
cf) @ Training time → 'absolute transfer'

$$= z_{s\to r} + w_{r\to t},\ t \in \{1, ..., T\}.$$

Experiments

same-identity

Method	VoxCeleb $\mathcal{L}_1$	AKD	AED	LPIPS
X2Face	0.078	7.687	0.405	-
Monkey-Net	0.049	1.878	0.199	-
FOMM	0.046	1.395	0.141	0.136
MRAA *w/o* bg	0.043	**1.307**	0.140	0.127
Ours	**0.041**	1.353	**0.138**	**0.123**



cross-identity

	VoxCeleb	GermanAudio
FOMM	0.323	0.456
MRAA	0.308	0.454
Ours	**0.161**	**0.406**

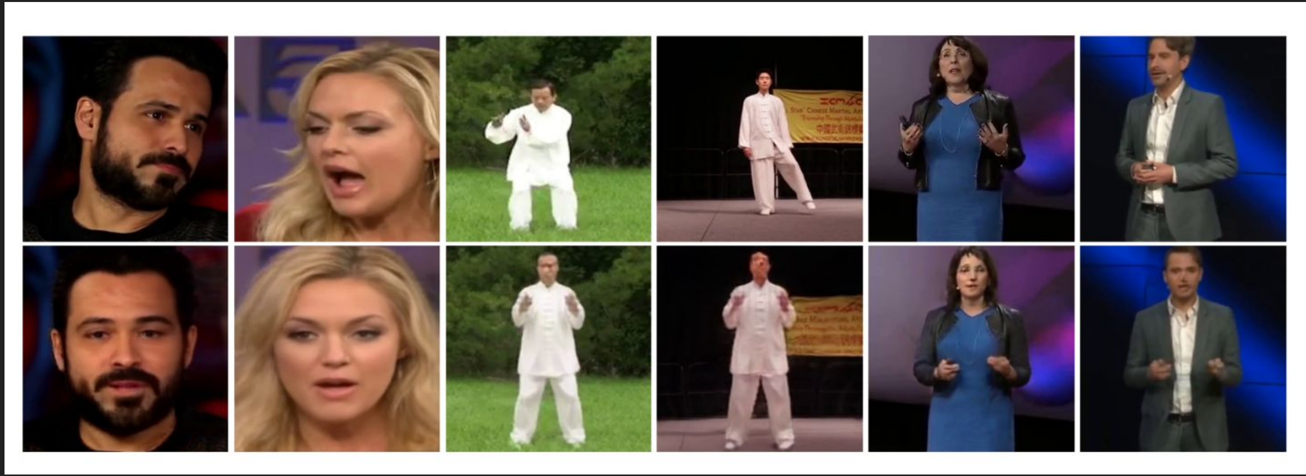
out-of-domain source images(FFHQ) inference

Additional analysis(cont'd)

1. Literally, xr represents what?
   A) Canonical(frontal/neutral) pose of xs, **regardless of original poses.**

source

reference



Additional analysis(cont’d)

2. Manipulation on Dm

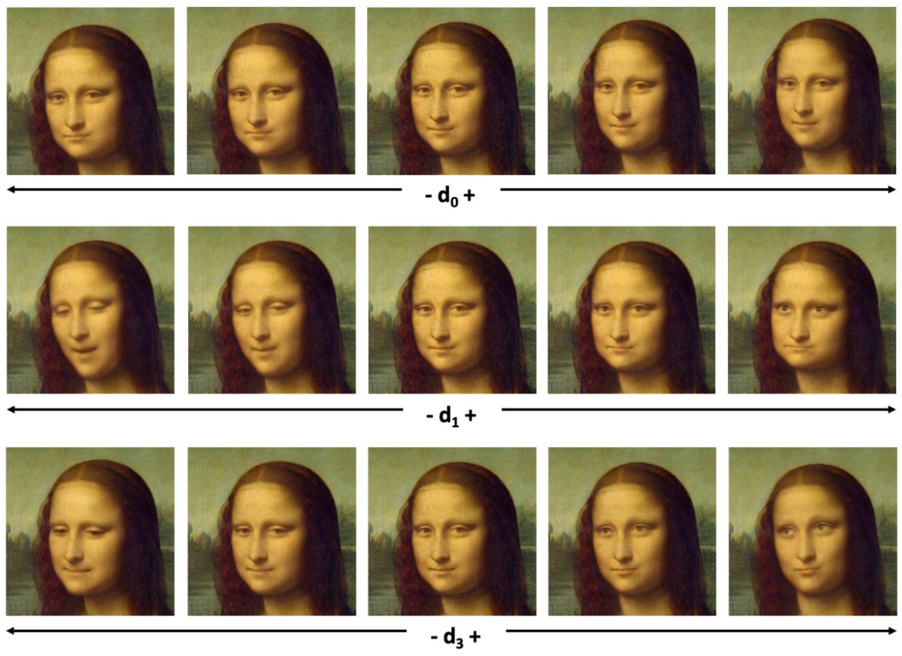


Figure 10: **Manipulation of motion dictionary.**

Conclusion

‘LIA opens a new door in design of **interpretable** generative models for video generation.’