Microsoft

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

High-Fidelity and Freely Controllable Talking Head Video Generation

Yue Gao, Yuan Zhou, Jinglu Wang, Xiao Li, Xiang Ming, Yan Lu
Microsoft Research

https://yuegao.me/PECHead/

TUE-PM-141

Summary

- Talking head video generation
  - Synthesize a talking head video with a given source identity and target motion.

Source identity

Targe video

Output

Summary - challenges

- The generated face obtained from existing method often has unexpected deformation and severe distortions.

Frontalization

Siarohin et.al. First order motion model for image animation. Advances in Neural Information Processing Systems, 32, 2019.

Summary - challenges

- The generated face obtained from existing method often has unexpected deformation and severe distortions.
- The movement-relevant information is not explicitly disentangled, which restricts the manipulation of different attributes during generation.

Summary - challenges

- The generated face obtained from existing method often has unexpected deformation and severe distortions.
- The movement-relevant information is not explicitly disentangled, which restricts the manipulation of different attributes during generation.
- The generated videos tend to have flickering artifacts due to the the sensitivity and inconsistency of the extracted landmarks.

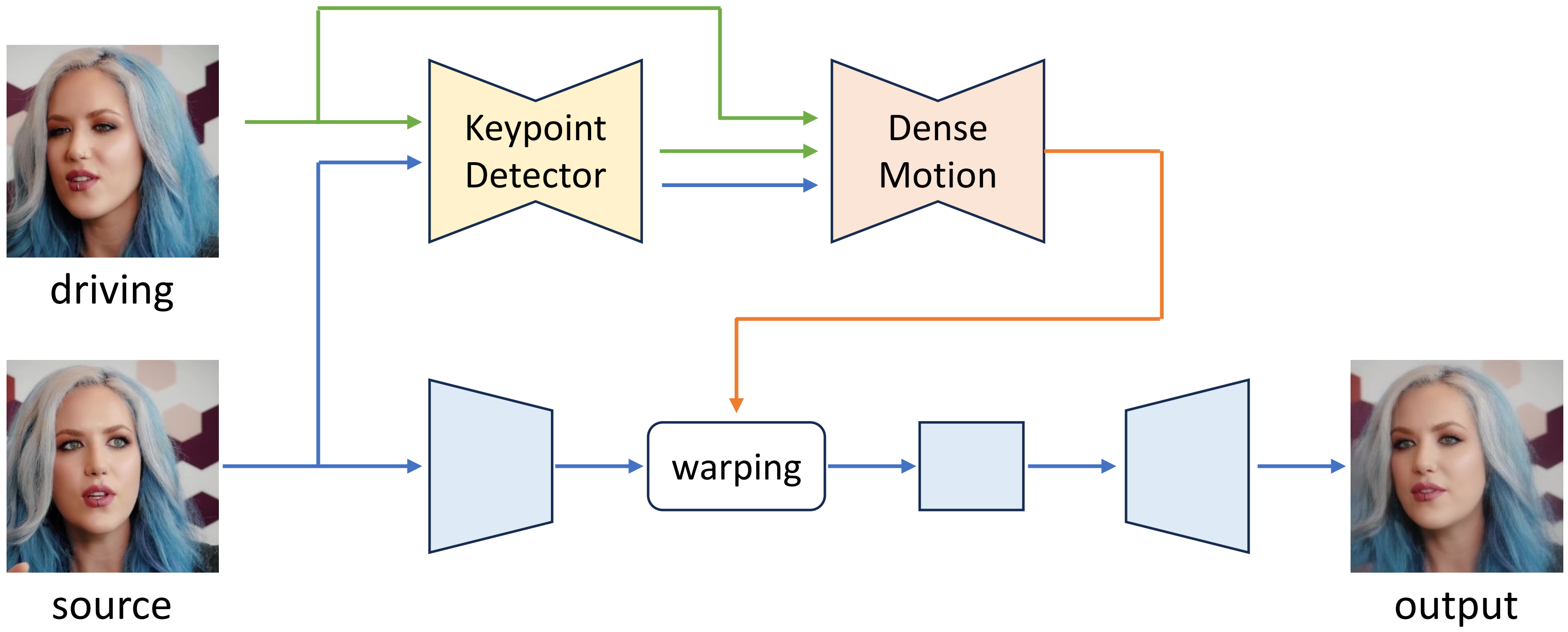
Summary

- Our method, PECHead, generates high-fidelity talking head videos enabling free control over the head pose and expression.

Introduction

Introduction

- Mainstream works follow the self-supervised learning pipeline



Introduction

- Three challenges for existing methods
  - Unexpected face distortions
  - Difficult to decouple and manipulate the movement information
  - Unnatural and flickering videos

Introduction

- Challenges - Unexpected face distortions
  - The learned landmarks-based approaches utilize the 2D learned landmarks without face shape constraints.
  - The predefined landmarks-based methods model the motion using only the predefined facial landmarks, leading to the non-facial parts of the head (such as the hair and neck) are not well handled.

Introduction

- Challenges - Difficult to semantically manipulate the movement
  - All the movement information needs to be obtained via one single driving image.
  - Hard to change the head poses or facial expressions of the source identity alone.

Introduction

- Challenges - Unnatural and flickering videos
  - Prior methods typically incorporate techniques to smoothen the extracted landmarks.
  - However, the sensitivity and inconsistency of the extracted landmarks pose a challenge in achieving smoothness.

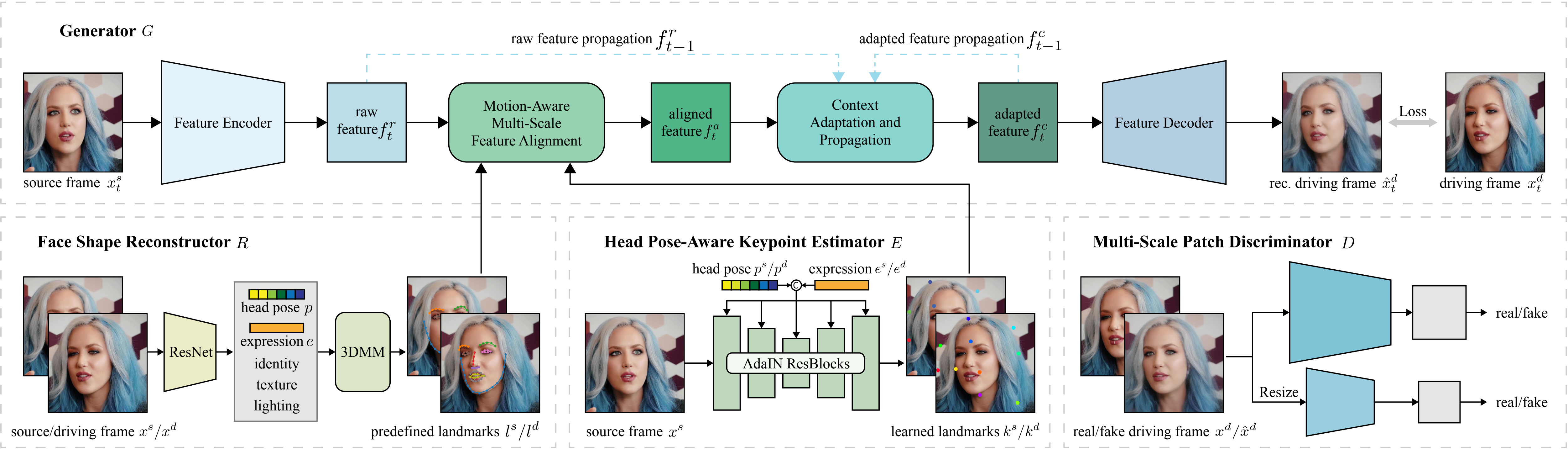
Introduction

- Contributions
  - We propose PECHead, that generates high-fidelity face reenactment results and talking head videos, enabling free control over the head pose and expression in talking head generation.
  - We incorporate the learned and predefined face landmarks for global and local motion estimation with the alignment module, which substantially enhances the quality of synthesized images.
  - We introduce a video-based pipeline with the adaptation and propagation module to further improve the smoothness and naturalness of the results.

Method

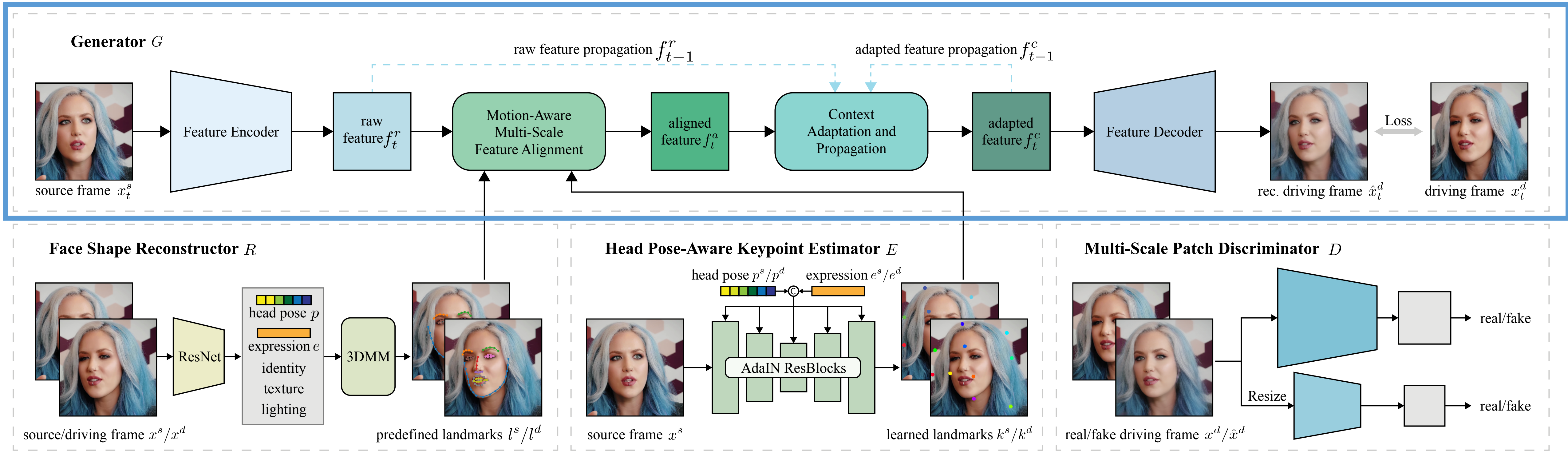
Method

- Overview



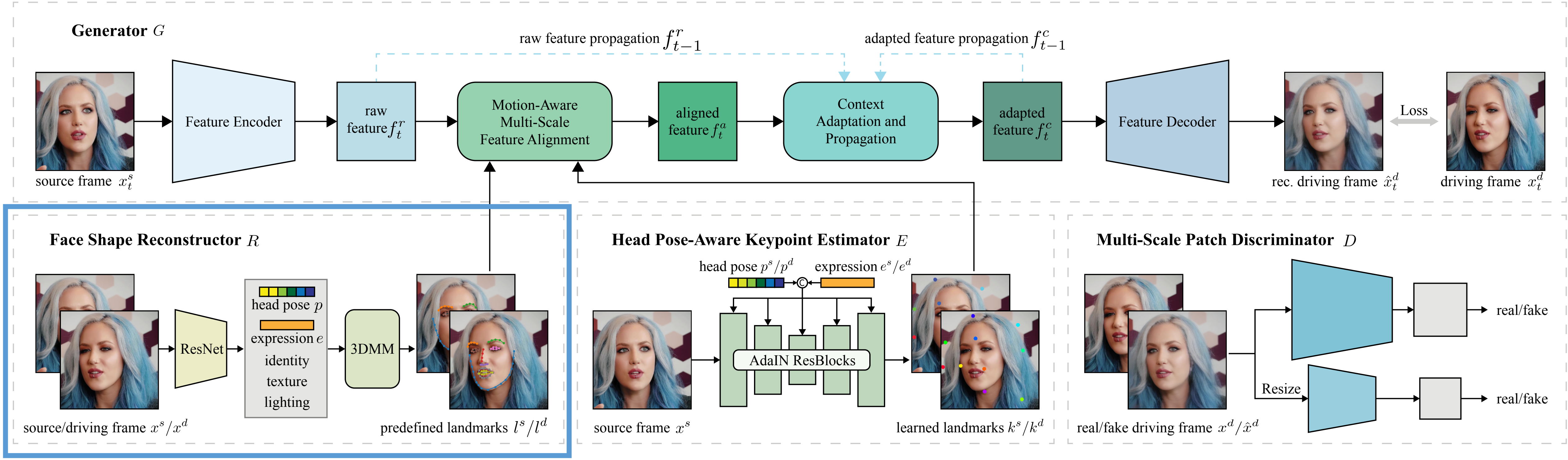
Method

- Overview



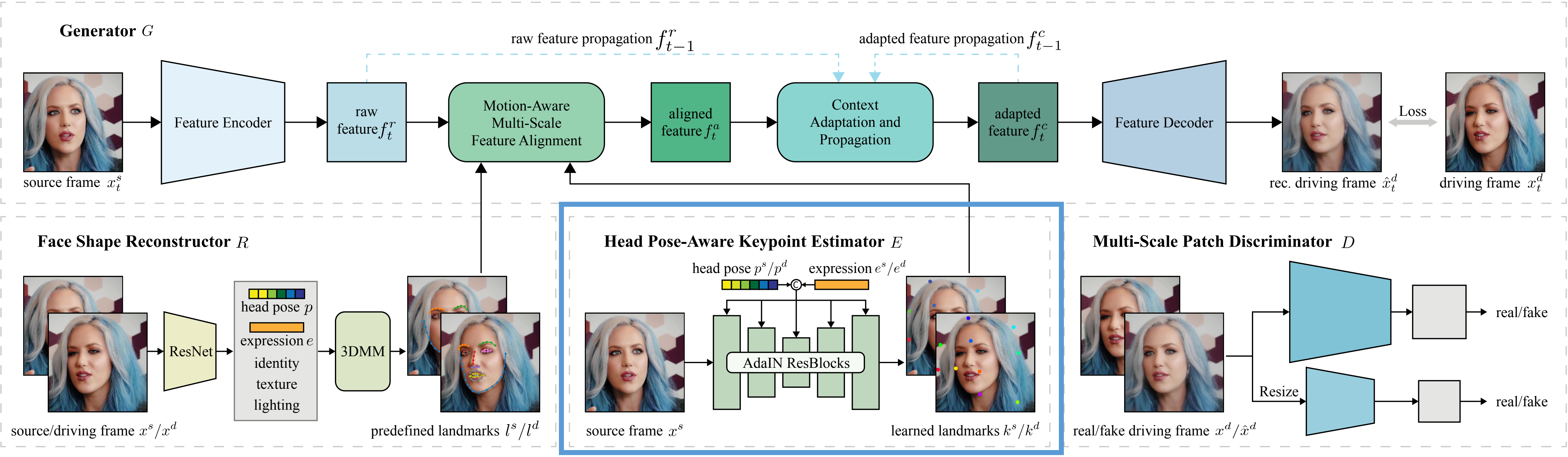
Method

- Overview



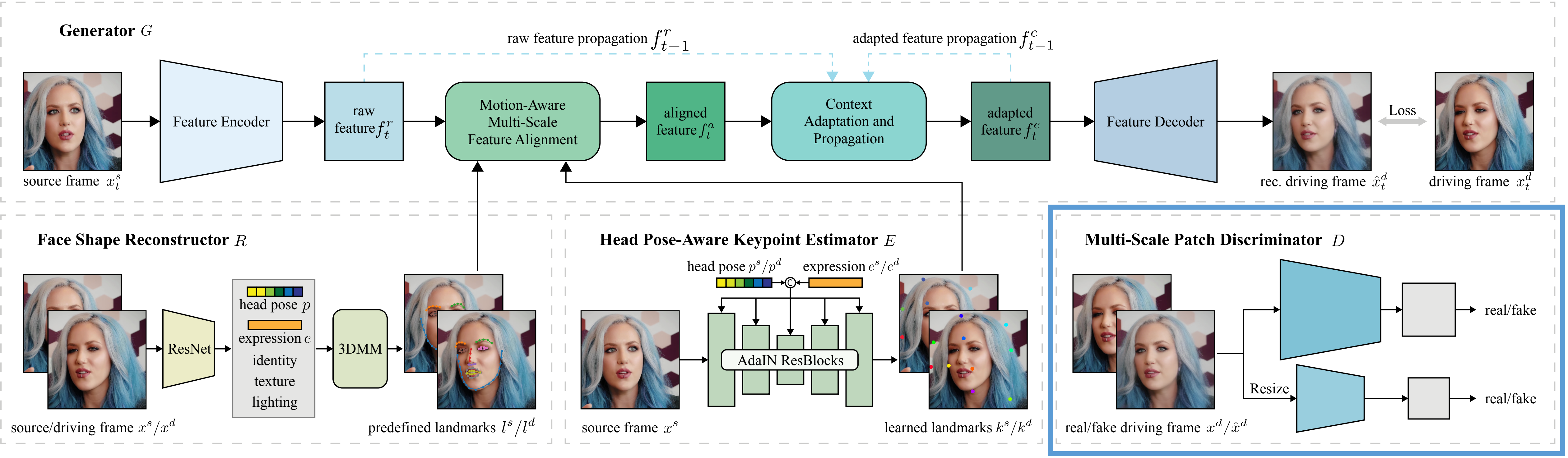
Method

- Overview



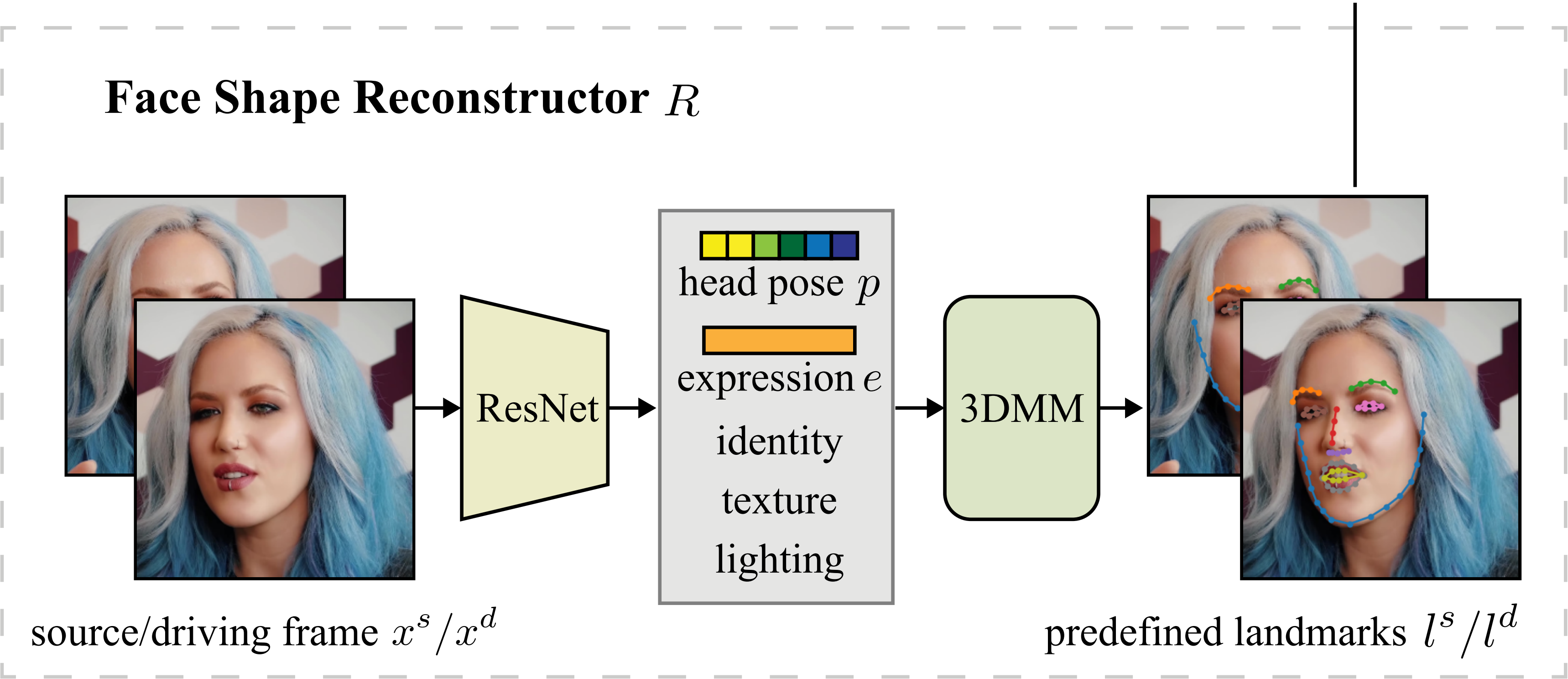
Method

- Overview



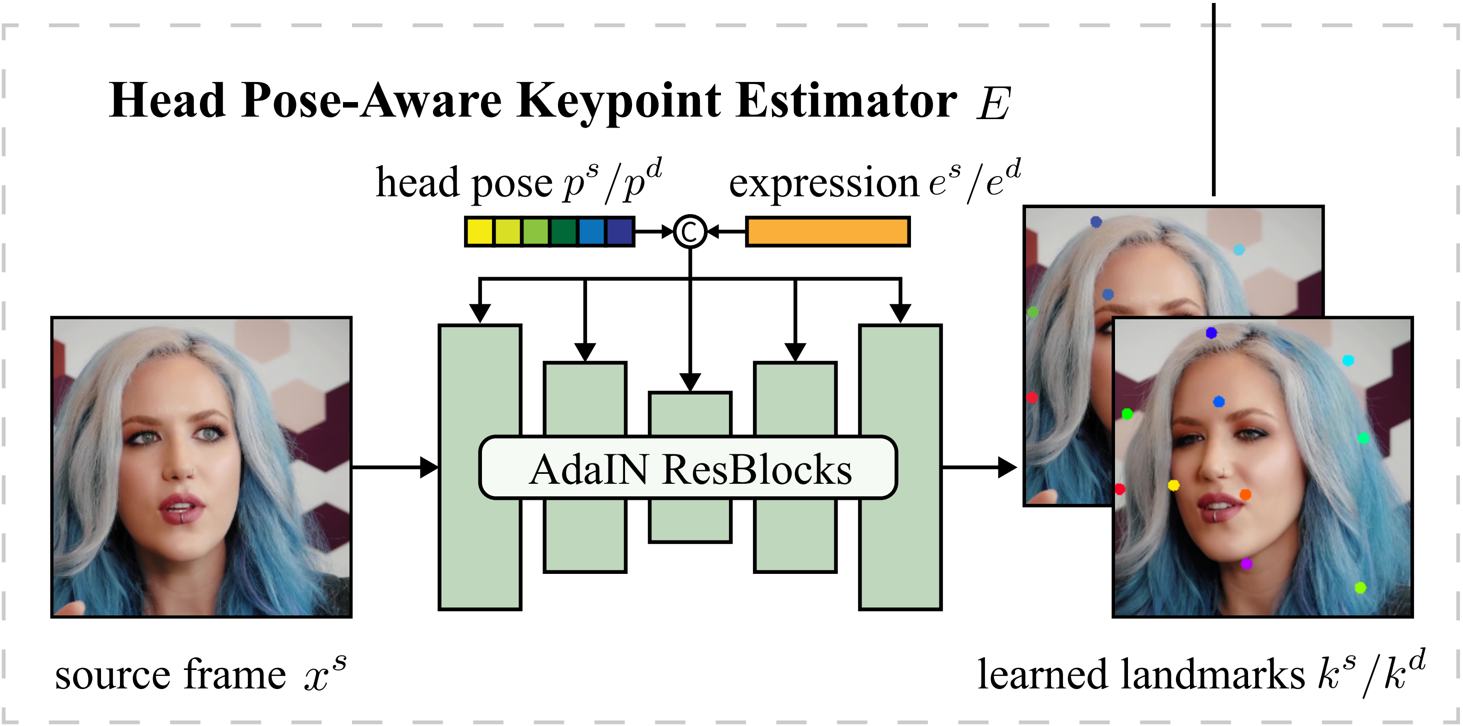
Method

- We first extract the face coefficients and predefined landmarks through R



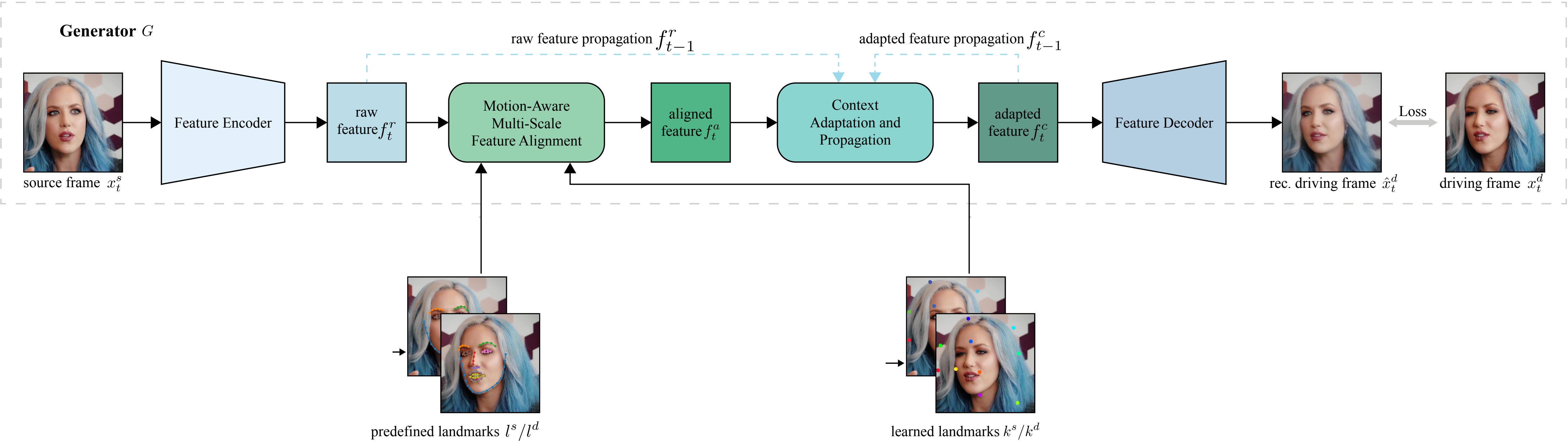
Method

- Then, we estimate the learned landmarks through *E* with the head pose and expression as conditions.



Method

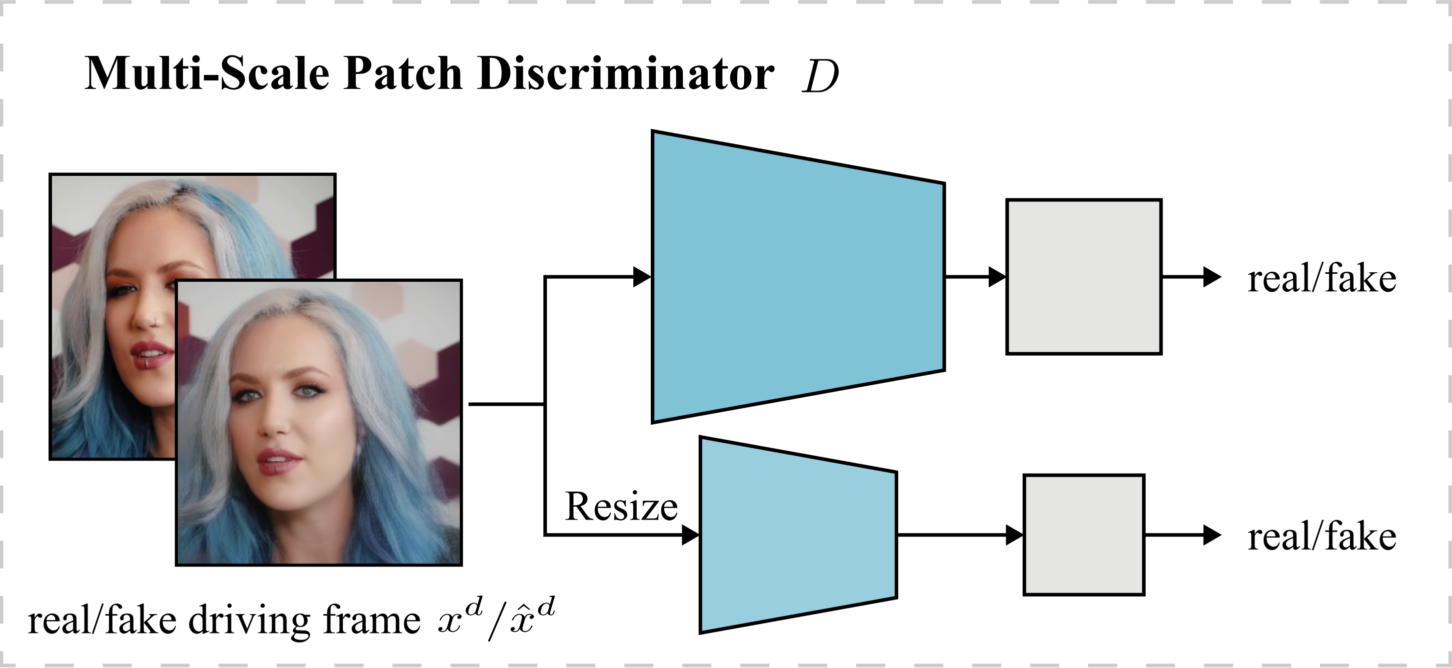
- The generator G takes the predefined and learned landmarks pairs to estimate the dense flow and generates the results.



Microsoft Research

Method

- The Multi-Scale Discriminator *D* is utilized to encourage the generator *G* produce more realistic frames.



Method

- Loss functions:
  - Pixel-wise loss $\mathcal{L}_p$ ensures the synthesis frames are similar to the driving frames.
  - Perceptual loss $\mathcal{L}_v$ guarantees consistency of high-level characteristics.
  - Learned landmarks loss $\mathcal{L}_k$ encourages the estimated learned landmarks to spread out across the whole frame.
  - Equivariance loss $\mathcal{L}_e$ constrains the consistency of E.
  - Warping loss $\mathcal{L}_w$ ensures the predicted deformations are reasonable.
  - GAN Loss $\mathcal{L}_G, \mathcal{L}_D$ improves the realism of the synthesized frames.

$$L_G = \lambda_p \mathcal{L}_p + \lambda_v \mathcal{L}_v + \lambda_k \mathcal{L}_k + \lambda_e \mathcal{L}_e + \lambda_w \mathcal{L}_w + \lambda_G \mathcal{L}_G,$$

$$L_D = \mathcal{L}_D$$

Experiments

Experiments

- **Datasets.** We evaluate our model on VoxCeleb2, TalkingHead-1KH, CelebV-HQ, and VFHQ.
- **Baselines.** We compare our approach with the recently proposed representative methods, FOMM, MRAA, OSFV, TPSMM, LIA, Face2Face$^{\rho}$ and DaGAN.
- **Metrics.** We evaluate a synthesis model on 1) reconstruction faithfulness using L_1 , MS-SSIM, PSNR, 2) output visual quality using FID, FVD, and 3) semantic consistency using average keypoint distance (AKD).

Experiments

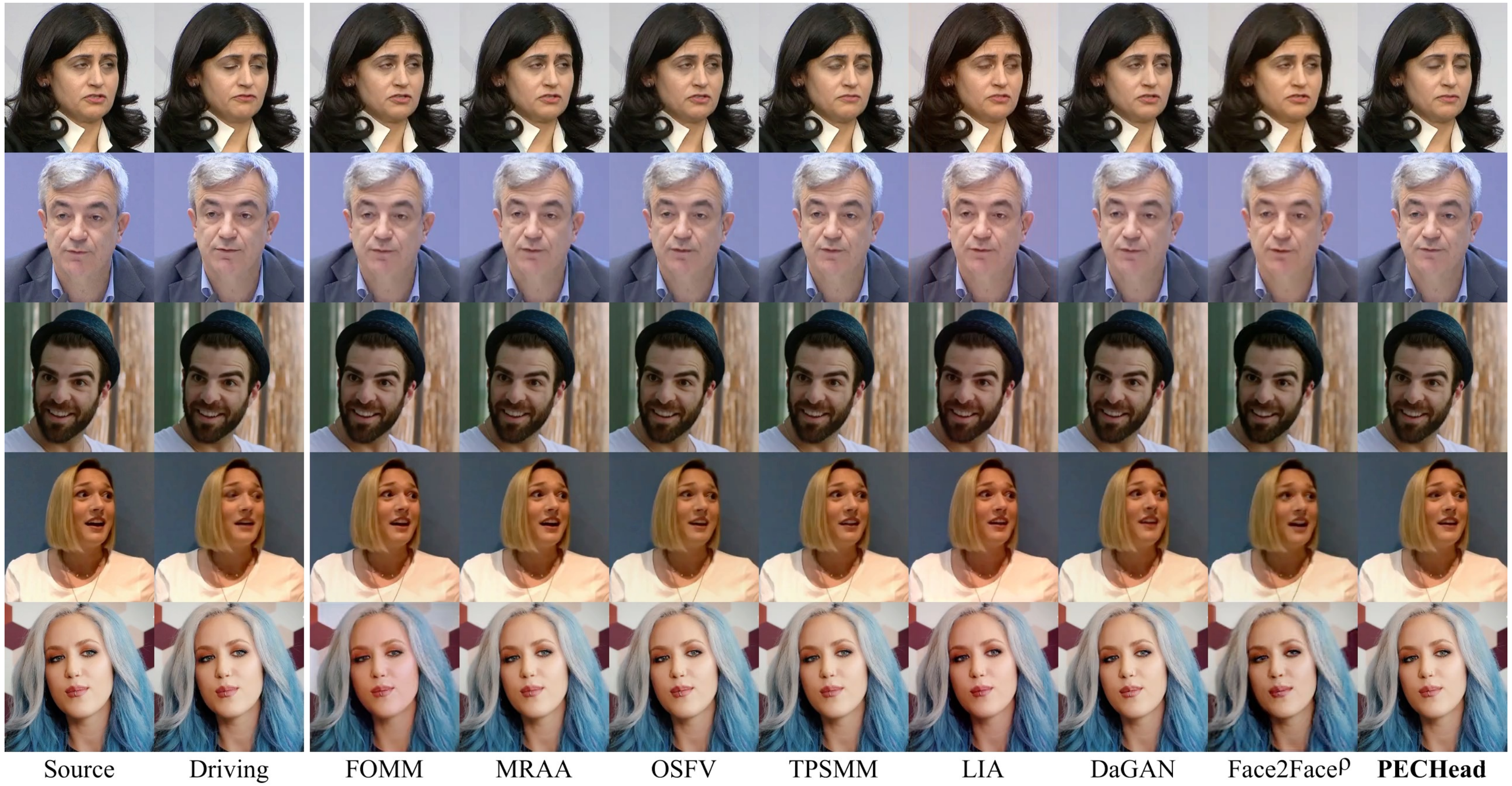
- Same-identity Video Reconstruction

Table 1. Quantitative results of different methods on four datasets for the same-identity video reconstruction.

Methods	VoxCeleb2					TalkHead-1KH					CelebV-HQ					VFHQ				
	L_1	MS-SSIM	PSNR	FID	AKD	L_1	MS-SSIM	PSNR	FID	AKD	L_1	MS-SSIM	PSNR	FID	AKD	L_1	MS-SSIM	PSNR	FID	AKD
FOMM [46]	0.0481	0.838	23.02	25.90	1.219	0.0431	0.821	23.28	33.22	2.905	0.0602	0.769	21.85	62.84	3.453	0.0526	0.780	21.76	47.82	2.868
MRAA [47]	0.0353	0.881	25.94	26.23	0.929	0.0361	0.882	25.50	32.57	1.057	0.0568	0.777	22.33	64.23	2.863	0.0454	0.812	22.60	40.17	2.123
OSFV [51]	0.0403	0.865	25.66	30.21	1.279	0.0432	0.837	23.59	35.12	3.100	0.0589	0.746	21.56	67.40	2.432	0.0491	0.804	21.79	41.95	1.730
TPSMM [67]	0.0318	0.902	26.88	24.39	0.709	0.0359	0.886	25.53	32.77	0.983	0.0615	0.757	22.05	64.89	3.714	0.0516	0.780	22.10	40.84	2.254
LIA [52]	0.0538	0.846	22.29	30.23	1.049	0.0477	0.879	24.43	38.89	0.932	0.0654	0.754	20.75	65.15	2.287	0.0537	0.815	21.47	42.27	1.502
DaGAN [22]	0.0359	0.881	25.64	24.92	0.844	0.0413	0.846	23.95	34.35	2.405	0.0637	0.739	21.32	68.04	4.800	0.0453	0.826	22.56	37.36	1.523
Face2Face$^{\rho}$ [59]	0.0507	0.816	20.83	31.71	1.332	0.0466	0.832	22.45	37.64	1.772	0.0709	0.710	19.94	71.87	3.754	0.0649	0.764	19.55	84.57	1.863
PECHead	**0.0304**	**0.905**	**26.96**	**23.05**	**0.626**	**0.0357**	**0.903**	**26.76**	**30.10**	**0.746**	**0.0552**	**0.803**	**24.29**	**56.68**	**1.215**	**0.0435**	**0.859**	**23.03**	**31.20**	**0.839**

Experiments

- Same-identity Video Reconstruction



Experiments

- Cross-identity Video Face Reenactment

Table 2. Quantitative results for the cross-identity reenactment.

Methods	CelebV-HQ				VFHQ			
	CSIM	ARD	AUH	FVD	CSIM	ARD	AUH	FVD
FOMM [46]	0.687	2.76	0.174	202.5	0.675	2.18	0.174	211.7
MRAA [47]	0.670	2.65	0.145	219.1	0.662	2.07	0.159	205.9
OSFV [51]	0.706	3.21	0.171	207.3	0.754	4.11	0.205	213.4
TPSMM [67]	0.673	1.85	0.125	220.2	0.674	1.84	0.143	207.8
LIA [52]	0.713	2.68	0.143	199.5	0.712	2.48	0.170	213.8
DaGAN [22]	0.716	2.66	0.154	205.9	0.684	1.91	0.143	217.6
Face2Face$^{\rho}$ [59]	0.535	9.91	0.251	232.5	0.673	2.13	0.170	206.4
PECHead	**0.733**	**0.85**	**0.118**	**192.2**	**0.789**	**0.81**	**0.104**	**201.6**

Experiments

- Cross-identity Video Face Reenactment

Experiments

- Head Pose and Expression Editing

Table 3. Quantitative results of pose and expression editing.

Methods	TalkHead-1KH			VFHQ		
	ARE	FID	AUH	ARE	FID	AUH
OSFV [51]	4.89	**40.96**	0.136	3.46	**53.21**	0.158
Face2Face$^{\rho}$ [59]	2.44	88.71	0.121	2.11	125.72	0.141
PECHead	**1.15**	42.04	**0.075**	**0.93**	56.16	**0.080**

Experiments

- Head Pose and Expression Editing - Frontalization

Experiments

- Head Pose and Expression Editing - Expression

Experiments

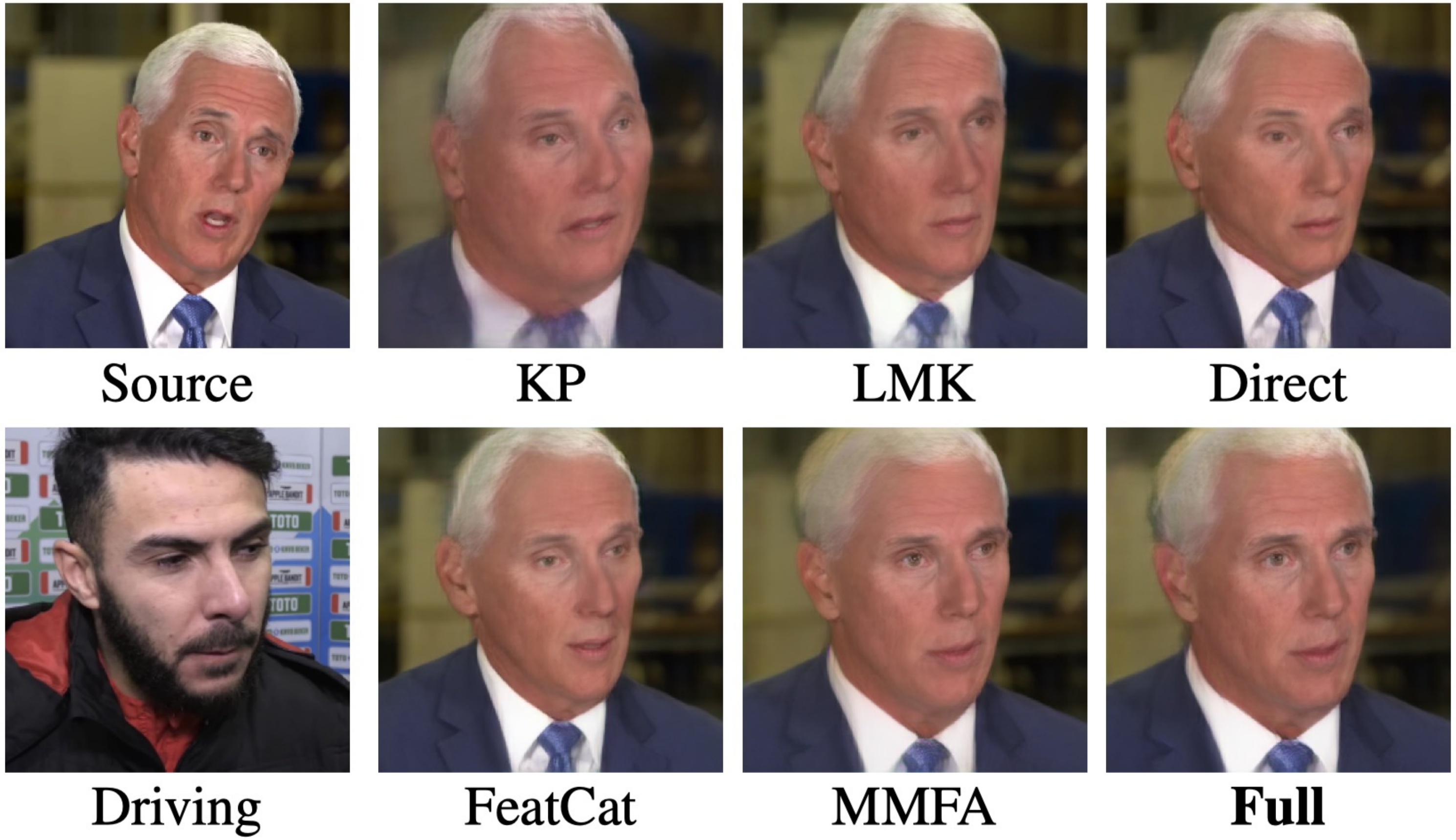
- Ablation Studies
  - Evaluate the performance of using both self-supervised learned and predefined facial landmarks.
  - Assess the performance of the proposed MMFA module.
  - Evaluate the performance of the proposed video-based framework involving the CAP module.

Table 4. Quantitative results for ablation studies.

Settings	TalkHead-1KH					VFHQ				
	L_1	FID	CSIM	ARD	FVD	L_1	FID	CSIM	ARD	FVD
KP	0.0446	35.82	0.726	1.41	215.8	0.0491	37.8	0.712	1.40	218.5
LMK	0.0426	37.30	0.717	1.29	213.9	0.0485	36.6	0.709	1.37	217.9
Direct	0.0439	35.58	0.730	1.37	212.7	0.0474	32.9	0.724	1.33	217.8
FeatCat	0.0430	34.96	0.732	1.34	208.2	0.0462	32.0	0.733	1.09	213.9
MMFA	0.0375	31.27	0.764	0.81	206.8	0.0448	**31.0**	0.782	0.85	209.9
Full	**0.0357**	**30.10**	**0.779**	**0.79**	**199.6**	**0.0435**	31.2	**0.789**	**0.84**	**201.6**

Experiments

- Ablation Studies



Experiments

- Wild Identities - Face Reenactment

Experiments

- Wild Identities - Free Editing

Input Yaw Pitch Roll Field of view

Microsoft Research

Conclusion

Conclusion

- We present a novel method, PECHead, which generates high-fidelity face reenactment results and talking head videos.
- Leveraging both learned and predefined landmarks, we introduce a motion-aware multi-scale feature alignment module to model global and local movements simultaneously.
- Furthermore, to improve the smoothness and naturalness of video synthesis, we introduce a context adaptation and propagation module that adapts the context of previous frames.
- Our method outperforms existing approaches in face reenactment and controllable talking head generation.

Thanks for Your Attention

Yue Gao

https://yuegao.me