One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing

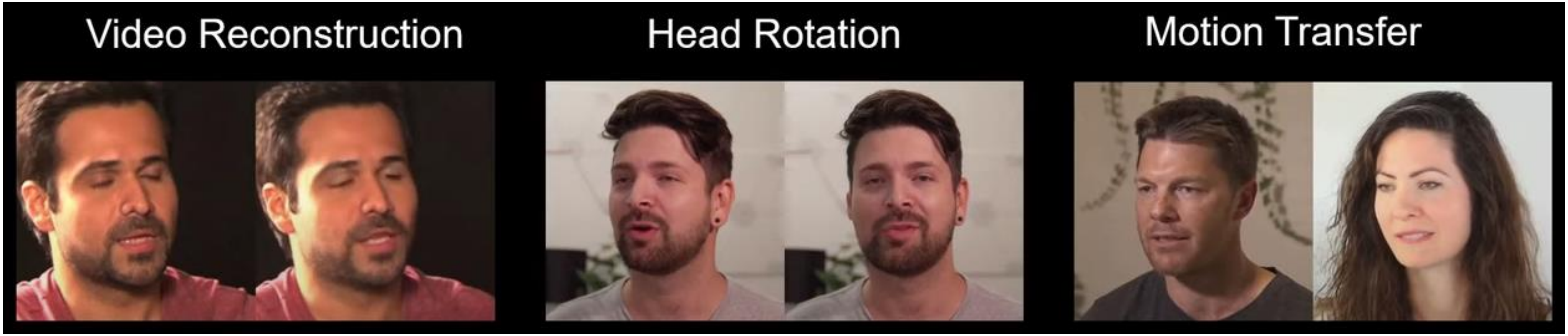
Ting-Chun Wang Arun Mallya Ming-Yu Liu

NVIDIA Corporation

CVPR 2021 (Oral)

Presentor : Taeu

Task : Neural Talking-Head Synthesis



https://nvlabs.github.io/face-vid2vid/

teaser_compression.gif (800×280) (nvlabs.github.io)   teaser_novelview.gif (600×323) (nvlabs.github.io)

https://www.youtube.com/watch?v=nLYg9Waw72U&t=70s

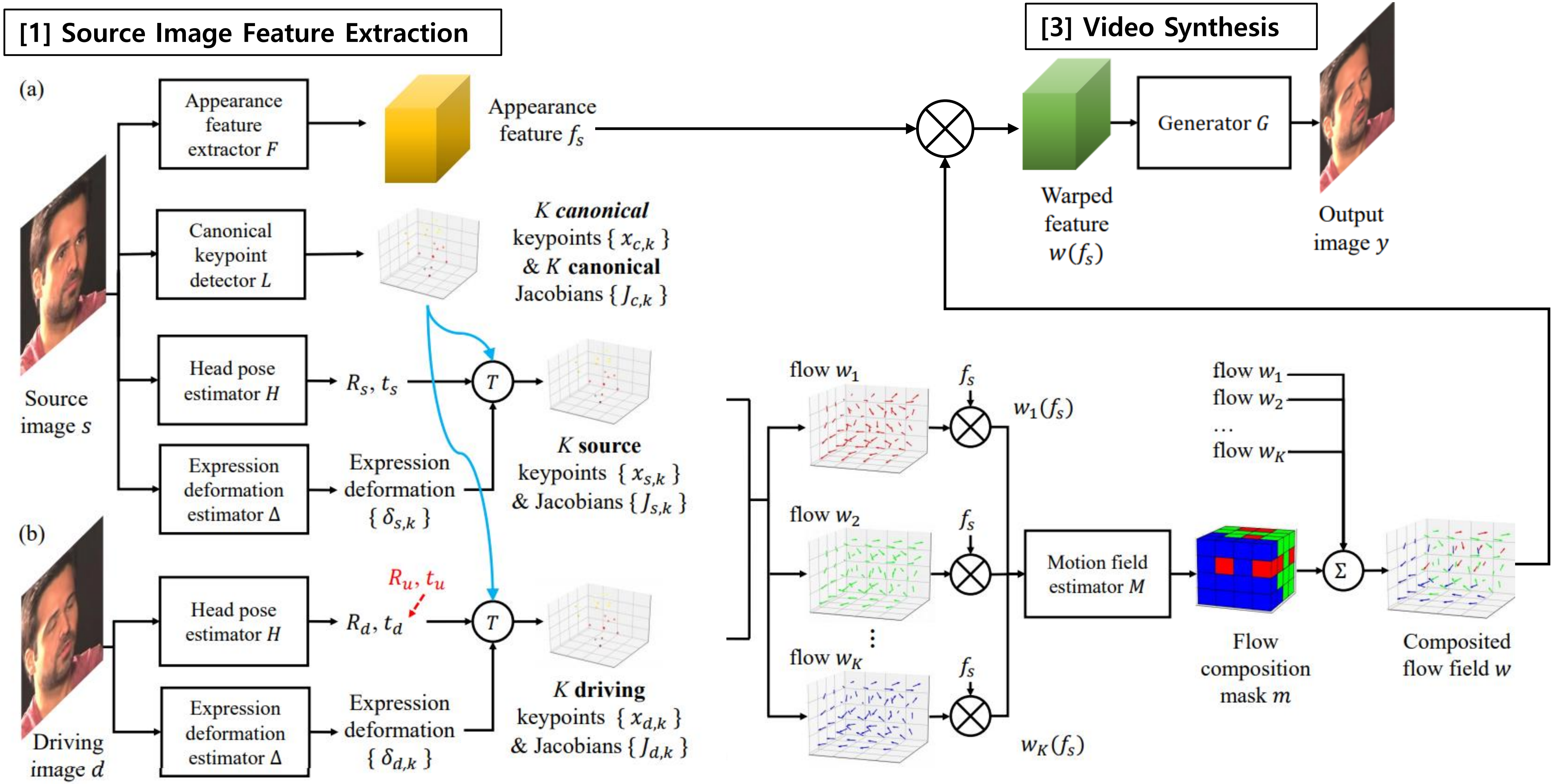
Task : Neural Talking-Head Synthesis

- **Contribution1** : A novel one-shot neural talking-head synthesis approach, which achieves better visual quality than state-of-the-art methods on the benchmark datasets.
- **Contribution2** : Local free-view control of the output video, without the need for a 3D graphics model. Our model allows changing the viewpoint of the talking-head during synthesis.
- **Contribution3** : Reduction in bandwidth for video streaming. We compare our approach to the commercial H.264 standard on a benchmark talking-head dataset and show that our approach can achieve 10x bandwidth reduction.

Overview of the Method

[1] Source Image Feature Extraction

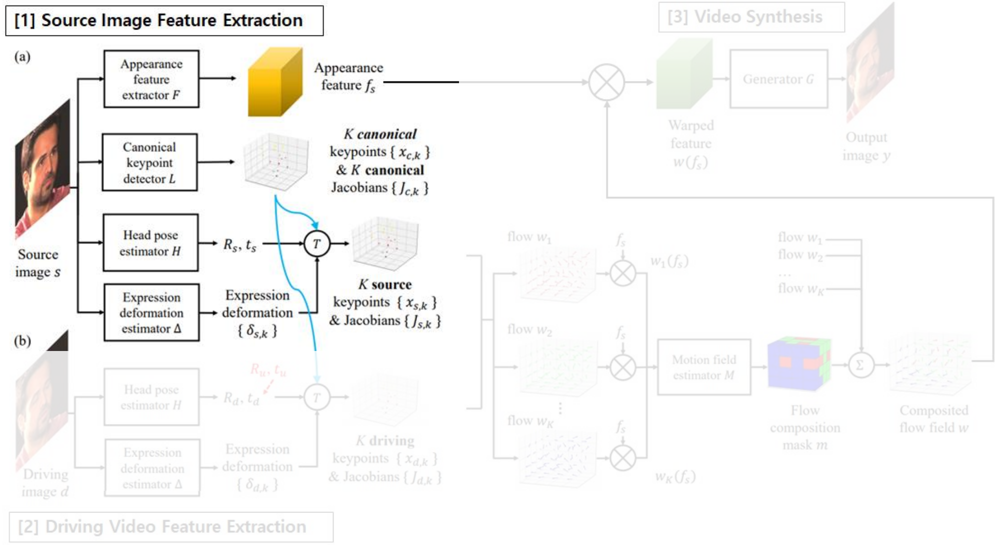
[3] Video Synthesis


(a)
Appearance feature extractor F
Appearance feature f_s
Canonical keypoint detector L
K **canonical** keypoints $\{ x_{c,k} \}$ & K **canonical** Jacobians $\{ J_{c,k} \}$
Head pose estimator H
R_s, t_s
T
Expression deformation estimator Δ
Expression deformation $\{ \delta_{s,k} \}$
K **source** keypoints $\{ x_{s,k} \}$ & Jacobians $\{ J_{s,k} \}$
Source image s
(b)
Head pose estimator H
R_u, t_u
R_d, t_d
T
Expression deformation estimator Δ
Expression deformation $\{ \delta_{d,k} \}$
K **driving** keypoints $\{ x_{d,k} \}$ & Jacobians $\{ J_{d,k} \}$
Driving image d
Warped feature $w(f_s)$
Generator G
Output image y
flow w_1
flow w_2
flow w_K
f_s
f_s
f_s
$w_1(f_s)$
$w_K(f_s)$
Motion field estimator M
Flow composition mask m
flow w_1
flow w_2
…
flow w_K
Σ
Composited flow field w

[2] Driving Video Feature Extraction

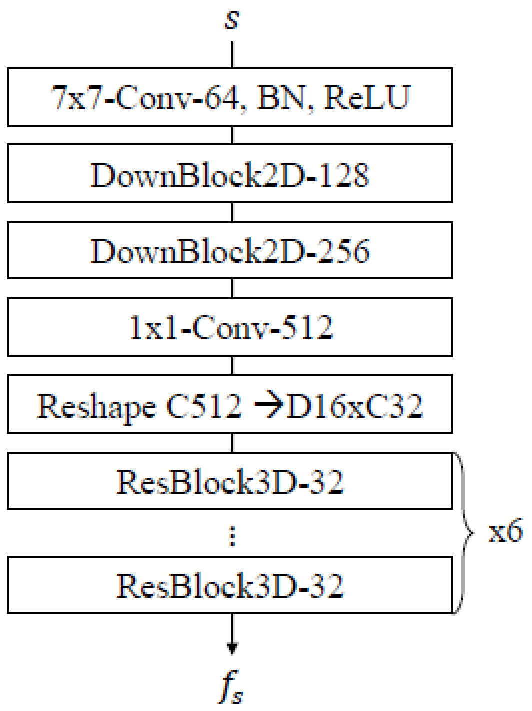
Overview of the Method

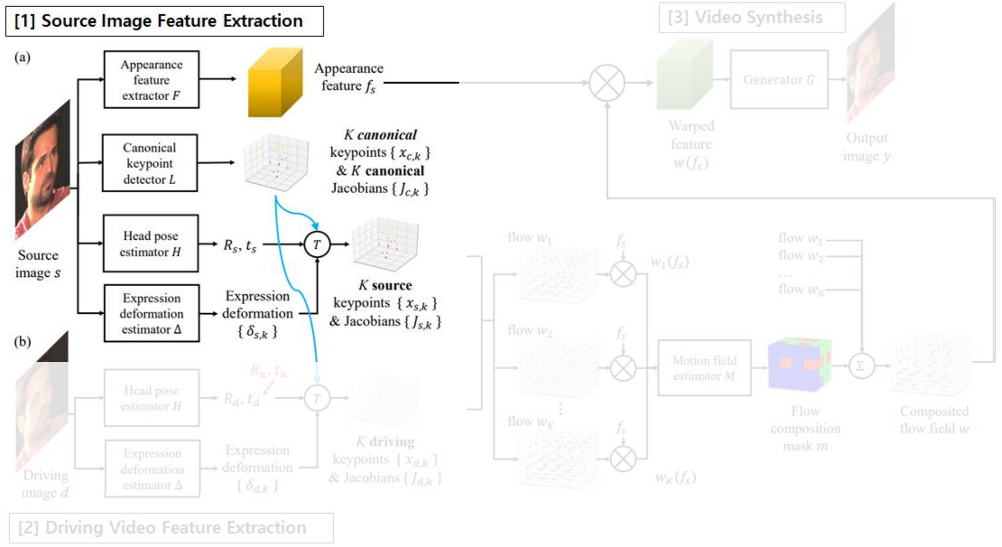
- **[1] Source Image Feature Extraction** :
  - Appearance feature extractor **F**
  - Canonical keypoint detector **L**
  - Head pose estimator **H**
  - Expression deformation estimator △
- **[2] Driving Video Feature Extraction** :
  - Head pose estimator **H**
  - Expression deformation estimator △
- **[3] Video Synthesis** :
  - 3D keypoints and Jacobians extracted from source and driving images → **Warping flow maps**
  - Generator **G**


[1] Source Image Feature Extraction
(a)
Appearance feature extractor F
Appearance feature f_s
Canonical keypoint detector L
K canonical keypoints $\{x_{c,k}\}$ & K canonical Jacobians $\{J_{c,k}\}$
Source image s
Head pose estimator H
R_s, t_s
T
Expression deformation estimator Δ
Expression deformation $\{\delta_{s,k}\}$
K source keypoints $\{x_{s,k}\}$ & Jacobians $\{J_{s,k}\}$
(b)
Head pose estimator H
R_u, t_u
R_d, t_d
T
Driving image d
Expression deformation estimator Δ
Expression deformation $\{\delta_{d,k}\}$
K driving keypoints $\{x_{d,k}\}$ & Jacobians $\{J_{d,k}\}$
[2] Driving Video Feature Extraction
[3] Video Synthesis
Warped feature $w(f_s)$
Generator G
Output image y
flow w_1
flow w_2
flow w_K
f_s
$w_1(f_s)$
$w_K(f_s)$
Motion field estimator M
Flow composition mask m
flow w_1
flow w_2
…
flow w_K
Σ
Composited flow field w

Model details

- **[1] Source Image Feature Extraction** :
  - Appearance feature extractor **F**



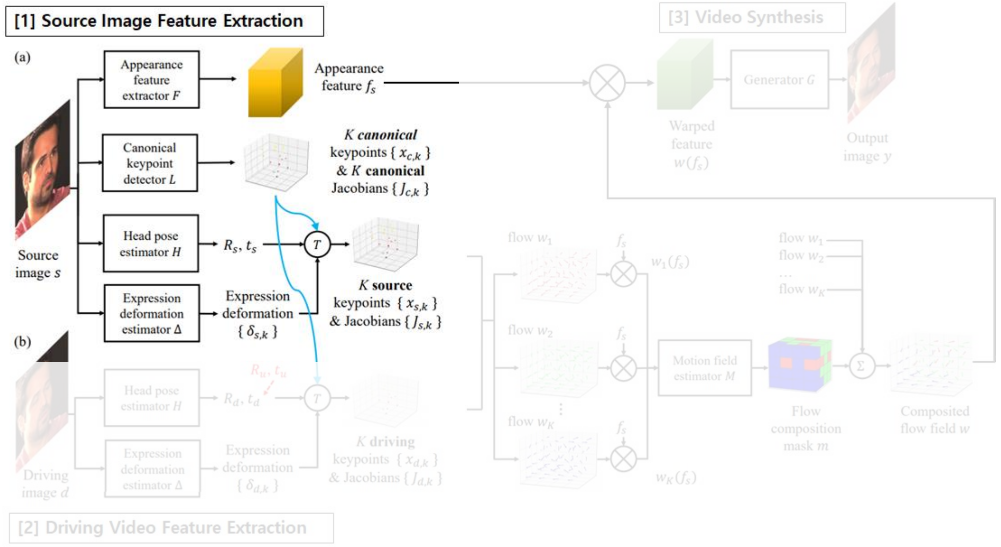


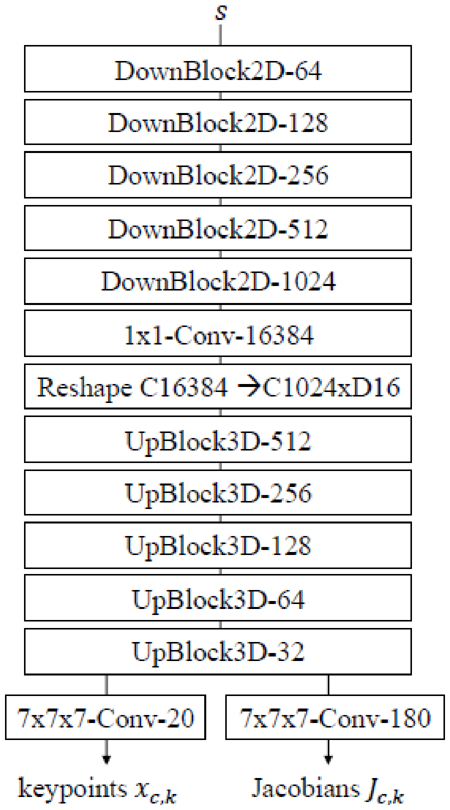
[1] Source Image Feature Extraction
[3] Video Synthesis
(a)
Appearance feature extractor F
Appearance feature f_s
Canonical keypoint detector L
K canonical keypoints $\{x_{c,k}\}$ & K canonical Jacobians $\{J_{c,k}\}$
Head pose estimator H
R_s, t_s
T
Expression deformation estimator Δ
Expression deformation $\{\delta_{s,k}\}$
K source keypoints $\{x_{s,k}\}$ & Jacobians $\{J_{s,k}\}$
Source image s
(b)
Head pose estimator H
R_d, t_d
R_u, t_u
Expression deformation estimator Δ
Expression deformation $\{\delta_{d,k}\}$
K driving keypoints $\{x_{d,k}\}$ & Jacobians $\{J_{d,k}\}$
Driving image d
[2] Driving Video Feature Extraction
flow w_1
flow w_2
flow w_K
f_s
$w_1(f_s)$
$w_K(f_s)$
Motion field estimator M
Flow composition mask m
flow w_1
flow w_2
...
flow w_K
Σ
Composited flow field w
Warped feature $w(f_s)$
Generator G
Output image y

Using a neural network *F*, the source image *s* is mapped to a 3D appearance feature volume *fs*. The network *F* consists of multiple downsampling blocks followed by a number of 3D residual blocks to compute the 3D feature volume *fs*.

Model details

- **[1] Source Image Feature Extraction** :
  - Canonical keypoint detector **L**


[1] Source Image Feature Extraction
[3] Video Synthesis
(a)
Appearance feature extractor F
Appearance feature f_s
Canonical keypoint detector L
K canonical keypoints { $x_{c,k}$ } & K canonical Jacobians { $J_{c,k}$ }
Head pose estimator H
R_s, t_s
T
Source image s
Expression deformation estimator Δ
Expression deformation { $\delta_{s,k}$ }
K source keypoints { $x_{s,k}$ } & Jacobians { $J_{s,k}$ }
(b)
Head pose estimator H
R_d, t_d
R_u, t_u
Driving image d
Expression deformation estimator Δ
Expression deformation { $\delta_{d,k}$ }
K driving keypoints { $x_{d,k}$ } & Jacobians { $J_{d,k}$ }
[2] Driving Video Feature Extraction
Warped feature $w(f_s)$
Generator G
Output image y
flow w_1
flow w_2
flow w_K
f_s
$w_1(f_s)$
$w_K(f_s)$
Motion field estimator M
Flow composition mask m
Σ
Composited flow field w



Using a canonical 3D keypoint detection network L, obtain the below, where we set K=20.

- 1. A set of K canonical 3D keypoints x_{c, k} ∈ R^3

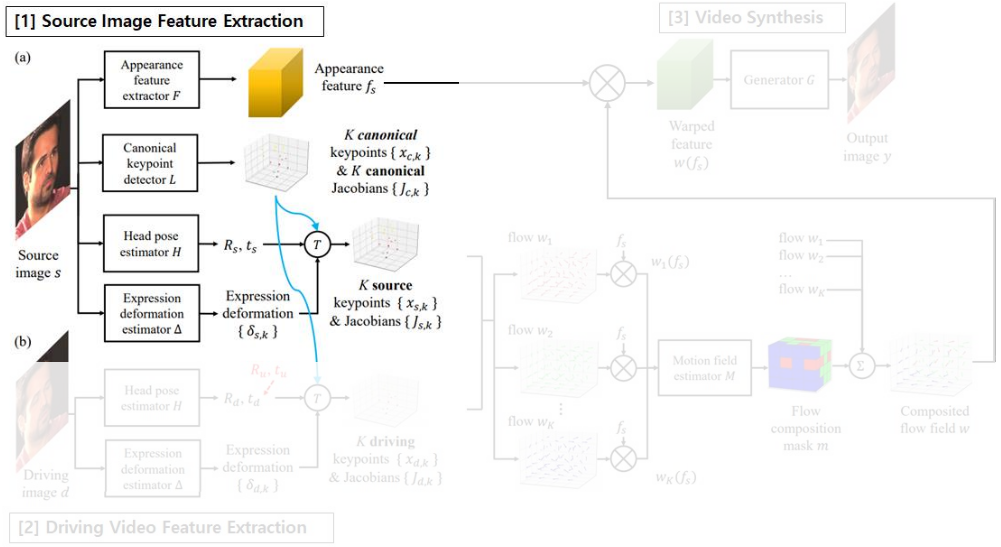
- 2. Their Jacobians J_{c, k} ∈ R^{3 x 3}

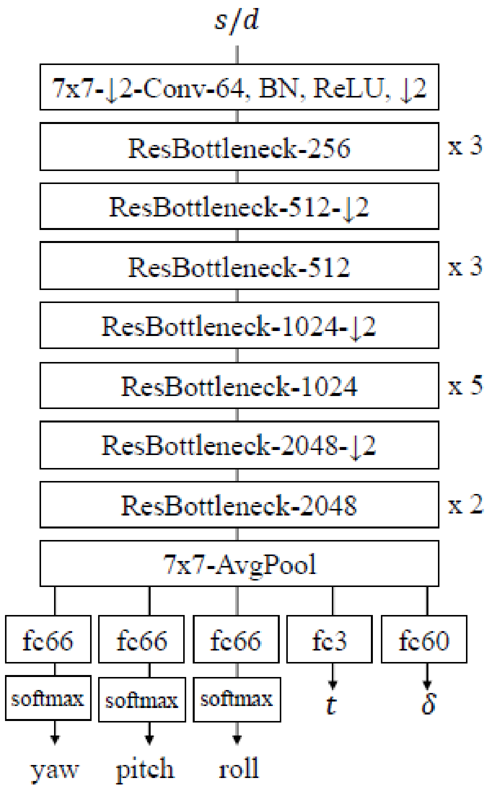
The Jacobians represent how a local patch around the keypoint can be transformed into a patch in another image via an affine transformation. The authors have used a U-Net style encoder-decoder to extract canonical keypoints.

We note that the extracted keypoints and Jacobians are meant to be independent of the face's pose and expression.

Model details

- **[1] Source Image Feature Extraction** :
  - Head pose estimator **H** & Expression deformation estimator △


[1] Source Image Feature Extraction
(a)
Appearance feature extractor F
Appearance feature f_s
Canonical keypoint detector L
K canonical keypoints $\{x_{c,k}\}$ & K canonical Jacobians $\{J_{c,k}\}$
Source image s
Head pose estimator H
R_s, t_s
T
Expression deformation estimator Δ
Expression deformation $\{\delta_{s,k}\}$
K source keypoints $\{x_{s,k}\}$ & Jacobians $\{J_{s,k}\}$
(b)
Head pose estimator H
R_u, t_u
R_d, t_d
Expression deformation estimator Δ
Expression deformation $\{\delta_{d,k}\}$
K driving keypoints $\{x_{d,k}\}$ & Jacobians $\{J_{d,k}\}$
Driving image d
[2] Driving Video Feature Extraction
[3] Video Synthesis
Warped feature $w(f_s)$
Generator G
Output image y
flow w_1
flow w_2
flow w_K
$w_1(f_s)$
$w_K(f_s)$
Motion field estimator M
Flow composition mask m
Composited flow field w



H, obtain the below,

- 1. A rotation matrix R_{s} ∈ R^{3x3}

- 2. A translation vector t_{s} ∈ R^{3}

* R_{s} in practice I s composed of three atrices – yaw, pitch, roll

The architecture is used to extract motion-related information from the image.

We adopt the same architecture as in Ruiz et al*. (The full angle range is divided into 66 bins for rotation angles, and the network predicts which bin the target angle is in.)

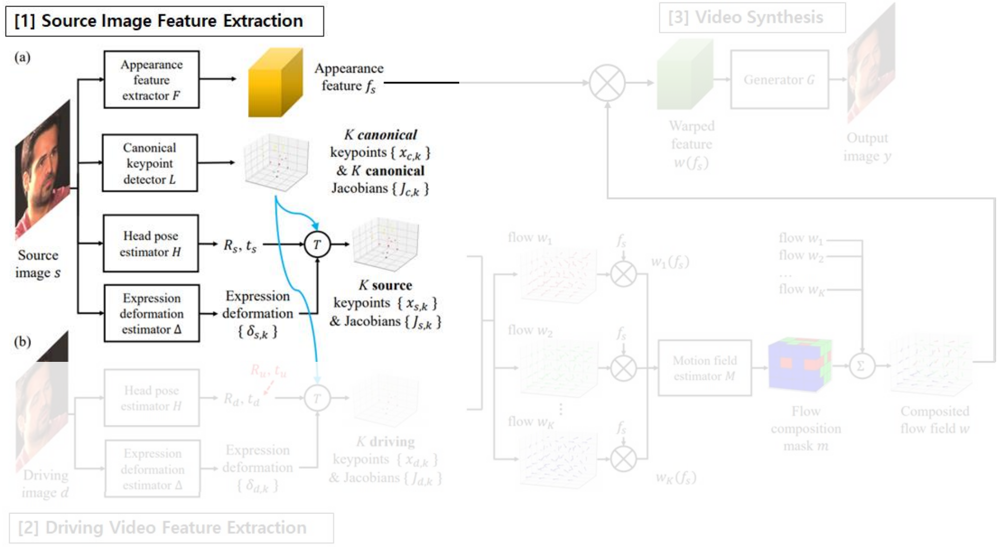
https://openaccess.thecvf.com/content_cvpr_2018_workshops/w41/html/Ruiz_Fine-Grained_Head_Pose_CVPR_2018_paper.html

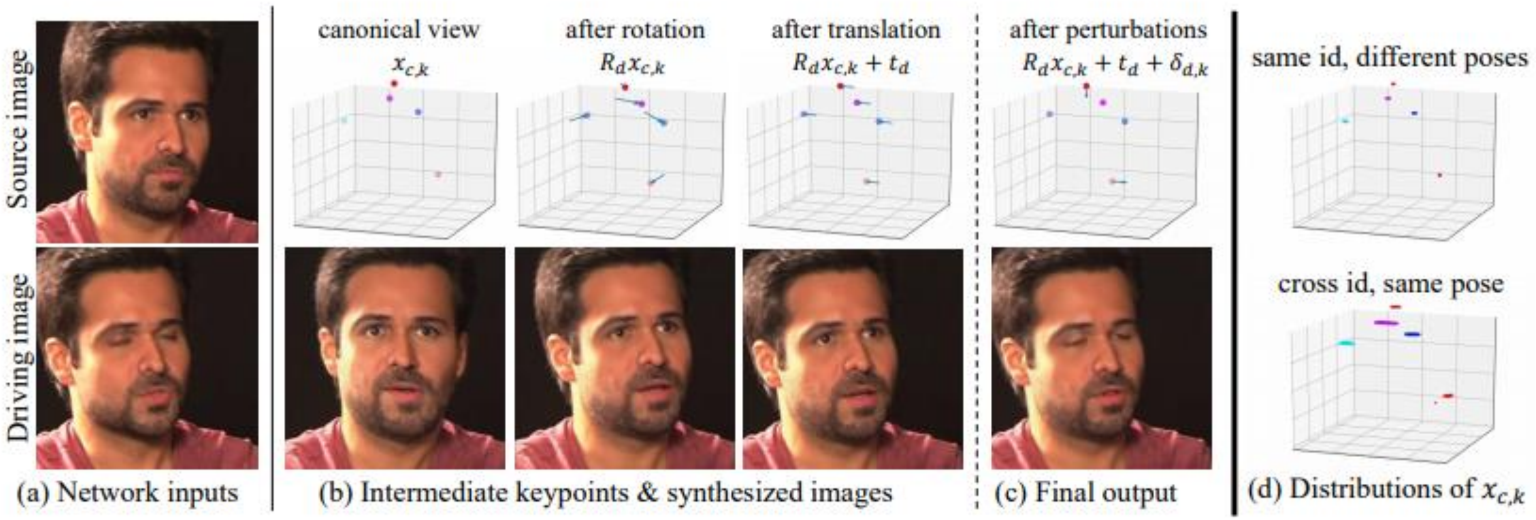
Model details

- **[1] Source Image Feature Extraction** :
  - Using the information from all 3 architectures, the authors have **proposed a transformation** *T* **to obtain the final 3D keypoints** *x s,k* **and their Jacobians** *Js,k* **for the source image**. *Tx* is applied to the keypoints and *Tj* to the Jacobians such that:

$$x_{s,k} = T_x(x_{c,k}, R_s, t_s, \delta_{s,k}) \equiv R_s x_{c,k} + t_s + \delta_{s,k} \qquad (1)$$

$$J_{s,k} = T_J(J_{c,k}, R_s) \equiv R_s J_{s,k}. \qquad (2)$$


[1] Source Image Feature Extraction
[3] Video Synthesis
(a)
Appearance feature extractor F
Appearance feature f_s
Canonical keypoint detector L
K canonical keypoints { $x_{c,k}$ } & K canonical Jacobians { $J_{c,k}$ }
Head pose estimator H
R_s, t_s
T
Source image s
Expression deformation estimator Δ
Expression deformation { $\delta_{s,k}$ }
K source keypoints { $x_{s,k}$ } & Jacobians { $J_{s,k}$ }
(b)
Head pose estimator H
R_u, t_u
R_d, t_d
Expression deformation estimator Δ
Expression deformation { $\delta_{d,k}$ }
K driving keypoints { $x_{d,k}$ } & Jacobians { $J_{d,k}$ }
Driving image d
[2] Driving Video Feature Extraction
Warped feature $w(f_s)$
Generator G
Output image y
flow w_1
flow w_2
flow w_K
f_s
$w_1(f_s)$
$w_K(f_s)$
Motion field estimator M
Flow composition mask m
Σ
Composited flow field w



canonical view $x_{c,k}$

after rotation $R_d x_{c,k}$

after translation $R_d x_{c,k} + t_d$

after perturbations $R_d x_{c,k} + t_d + \delta_{d,k}$

same id, different poses

cross id, same pose

(a) Network inputs

(b) Intermediate keypoints & synthesized images

(c) Final output

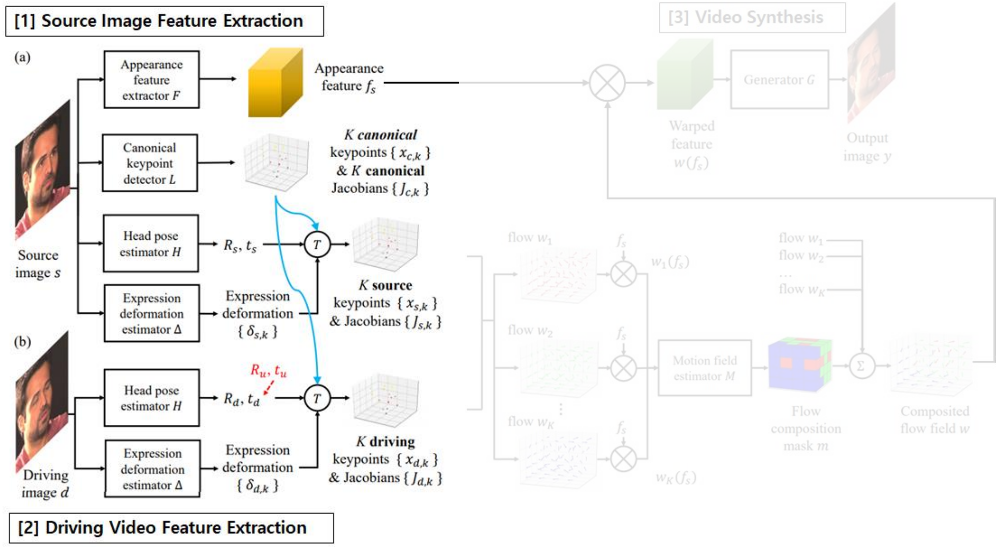
(d) Distributions of $x_{c,k}$

Model details

- **[2] Driving Video Feature Extraction** :
  - From Head pose estimator **H &** Expression deformation estimator △**,** we get

$$x_{d,k} = T_x(x_{c,k}, R_d, t_d, \delta_{d,k}) = R_d x_{c,k} + t_d + \delta_{d,k} \quad (3)$$

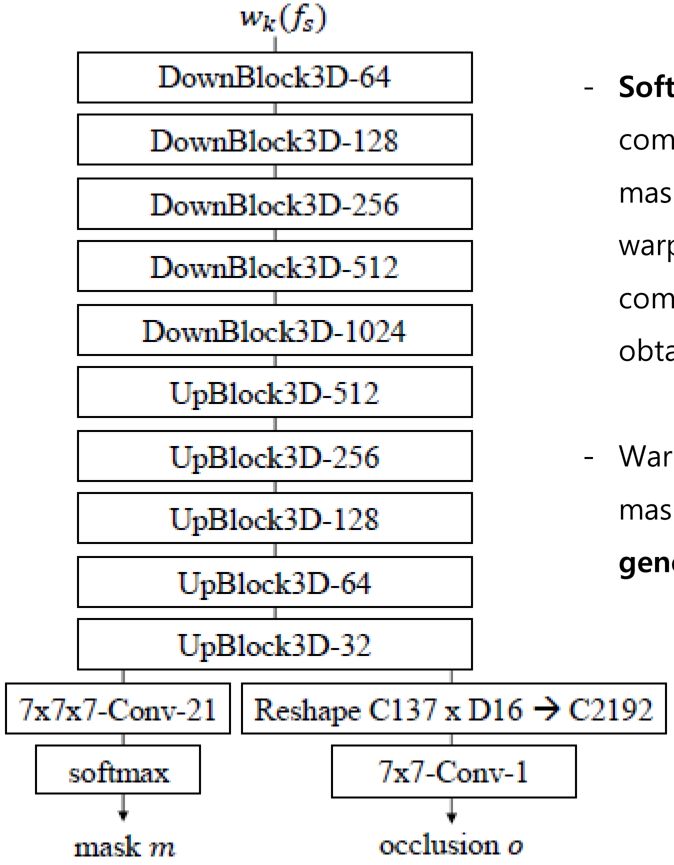
$$J_{d,k} = T_J(J_{c,k}, R_d) = R_d J_{c,k}. \quad (4)$$


[1] Source Image Feature Extraction
(a)
Appearance feature extractor F
Appearance feature f_s
Canonical keypoint detector L
K canonical keypoints $\{x_{c,k}\}$ & K canonical Jacobians $\{J_{c,k}\}$
Source image s
Head pose estimator H
R_s, t_s
T
Expression deformation estimator Δ
Expression deformation $\{\delta_{s,k}\}$
K source keypoints $\{x_{s,k}\}$ & Jacobians $\{J_{s,k}\}$
(b)
Head pose estimator H
R_u, t_u
R_d, t_d
T
Driving image d
Expression deformation estimator Δ
Expression deformation $\{\delta_{d,k}\}$
K driving keypoints $\{x_{d,k}\}$ & Jacobians $\{J_{d,k}\}$
[2] Driving Video Feature Extraction
[3] Video Synthesis
Warped feature $w(f_s)$
Generator G
Output image y
flow w_1
flow w_2
flow w_K
$w_1(f_s)$
$w_K(f_s)$
Motion field estimator M
Flow composition mask m
Composited flow field w

- The driving video is used to extract motion-related information. To this end, head pose estimation network *H* and expression deformation estimator network △ is used.
- Our approach allows manual changes to the 3D head pose during synthesis. Let Ru and tu be user-specified rotation and translation, respectively. The final head pose in the output image is given by Rd → RuRd and td → tu + td.

Model details

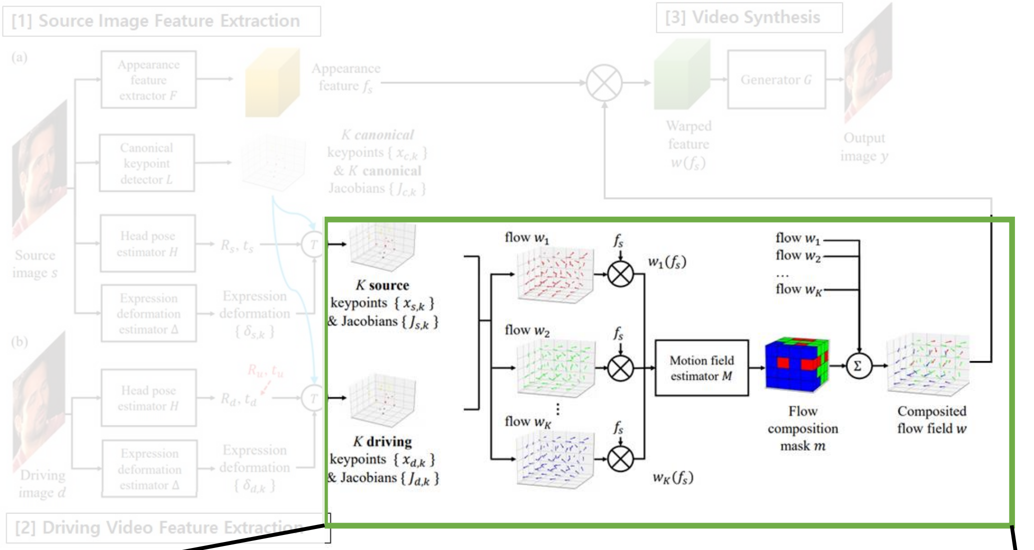
- **[3] Video Synthesis** :
  - Estimate K warping flow maps, ***wk*** is used to warp the source feature *fs* where $k \in \{1,2,..,K\}$.



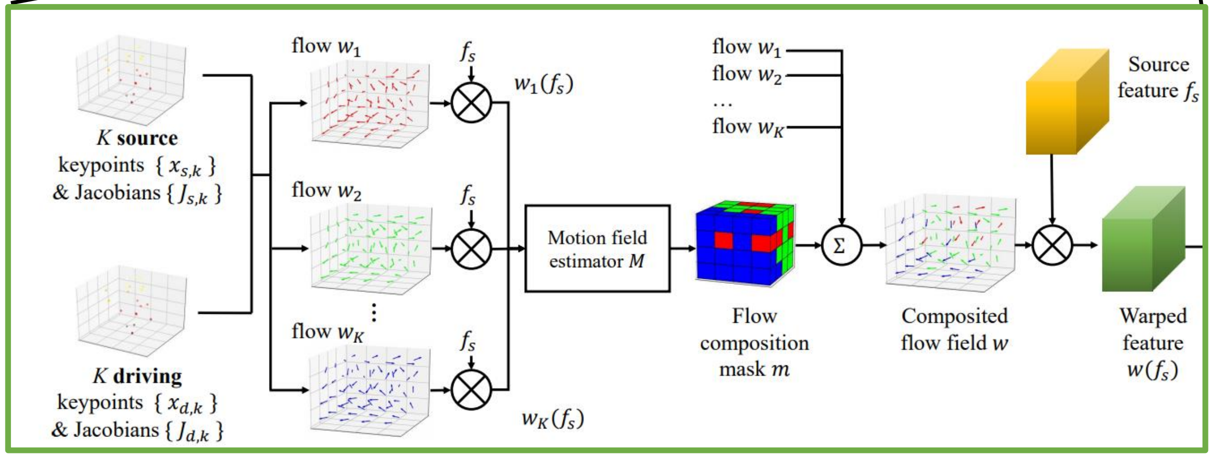
- **Softmax activation** is used to obtain the flow composition mask m, which consists of K 3D masks. These are again combined with K warping flow maps w_k to obtain the final composite flow field w. This is finally used to obtain the warped source feature w(f_s).

- Warping leads to **occlusions**. A **2D occlusion** mask o is predicted to be inputted to the **generator G**.

- A linear combination of **m** and **wk**'s then produces the composited flow field **w**



[1] Source Image Feature Extraction
[3] Video Synthesis
(a)
Appearance feature extractor F
Appearance feature f_s
Generator G
Warped feature $w(f_s)$
Output image y
Canonical keypoint detector L
K canonical keypoints $\{x_{c,k}\}$ & K canonical Jacobians $\{J_{c,k}\}$
Source image s
Head pose estimator H
R_s, t_s
Expression deformation estimator Δ
Expression deformation $\{\delta_{s,k}\}$
(b)
Driving image d
R_d, t_d
Expression deformation $\{\delta_{d,k}\}$
[2] Driving Video Feature Extraction
K source keypoints $\{x_{s,k}\}$ & Jacobians $\{J_{s,k}\}$
K driving keypoints $\{x_{d,k}\}$ & Jacobians $\{J_{d,k}\}$
flow w_1
flow w_2
flow w_K
f_s
$w_1(f_s)$
$w_K(f_s)$
Motion field estimator M
Flow composition mask m
Composited flow field w

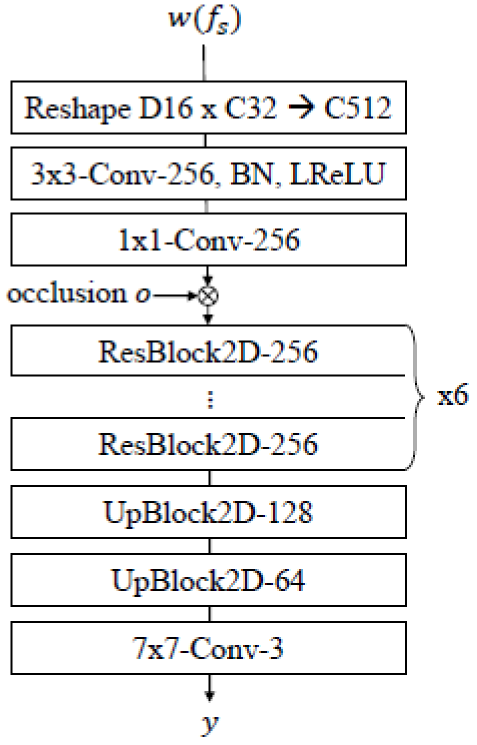


K source keypoints $\{x_{s,k}\}$ & Jacobians $\{J_{s,k}\}$
K driving keypoints $\{x_{d,k}\}$ & Jacobians $\{J_{d,k}\}$
flow w_1
flow w_2
flow w_K
f_s
$w_1(f_s)$
$w_K(f_s)$
Motion field estimator M
Flow composition mask m
flow w_1
flow w_2
…
flow w_K
Σ
Composited flow field w
Source feature f_s
Warped feature $w(f_s)$

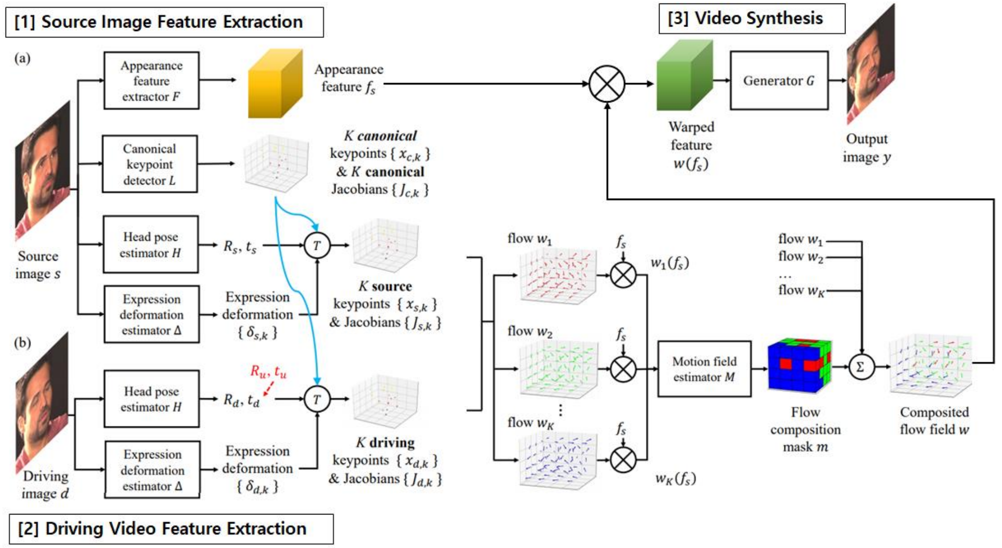
We first compute the warping flow wk induced by the k-th keypoint using **the first order approximation*** , which is reliable only around the neighborhood of the keypoint. * First Order Motion Model for Image Animation by Siarohin et al

Model details

- **[3] Video Synthesis** :
  - A generator network *G* that takes the warped 3D source feature map *w*(*fs*) and first projects them back to the 2D feature.



- This is then multiplied with the occlusion mask *o* followed by a series of 2D residual blocks and upsampling layers to obtain the final image.



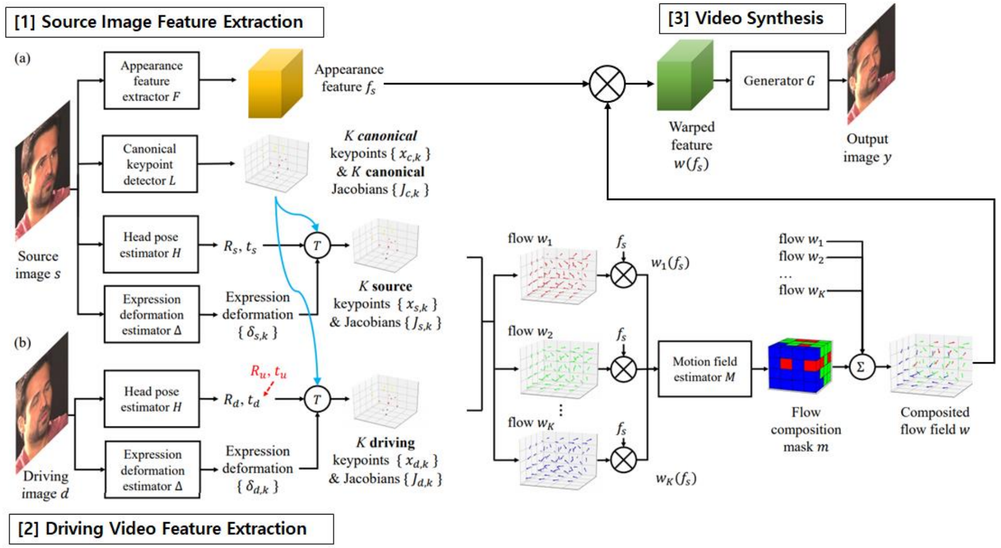
[1] Source Image Feature Extraction
[3] Video Synthesis
(a)
Appearance feature extractor F
Appearance feature f_s
Warped feature $w(f_s)$
Generator G
Output image y
Canonical keypoint detector L
K canonical keypoints $\{x_{c,k}\}$ & K canonical Jacobians $\{J_{c,k}\}$
Source image s
Head pose estimator H
R_s, t_s
T
K source keypoints $\{x_{s,k}\}$ & Jacobians $\{J_{s,k}\}$
Expression deformation estimator Δ
Expression deformation $\{\delta_{s,k}\}$
(b)
Driving image d
Head pose estimator H
R_u, t_u
R_d, t_d
T
K driving keypoints $\{x_{d,k}\}$ & Jacobians $\{J_{d,k}\}$
Expression deformation estimator Δ
Expression deformation $\{\delta_{d,k}\}$
flow w_1
flow w_2
flow w_K
f_s
$w_1(f_s)$
$w_K(f_s)$
Motion field estimator M
Flow composition mask m
flow w_1
flow w_2
...
flow w_K
Σ
Composited flow field w
[2] Driving Video Feature Extraction

Training details

- **Losses**: $$\mathcal{L} = \lambda_P \mathcal{L}_P(d, y) + \lambda_G \mathcal{L}_G(d, y) + \lambda_E \mathcal{L}_E(\{x_{d,k}\}, \{J_{d,k}\}) + \lambda_L \mathcal{L}_L(\{x_{d,k}\}) + \lambda_H \mathcal{L}_H(R_d, \bar{R}_d) + \lambda_\Delta \mathcal{L}_\Delta(\{\delta_{d,k}\})$$

  - **Perpetual Loss (L*P*):** a pre-trained VGG, L1 distance the ground truth **image** 5 features and the reconstructed image 5 features with 3 multiple resolution.
  - **GAN Loss** (L*G*)**:** patch GAN with hinge loss. Two-scale discriminator for 512x.
  - **Equivalence Loss** (L*E*)**:** This loss ensures the **consistency** of the estimated **keypoints**. Let *xd* be the detected keypoints for the input image *d*. When a known transformation ***T*** is applied to the image *T*(*d*), the detected keypoints should be transformed in the same way. *L*1 distance is minimized such that $\|x_d - \mathbf{T}^{-1}(x_{\mathbf{T}(d)})\|_1$ tends to zero. Here T^{-1} is the inverse of the known transform. The same logic is applicable to the jacobians of the keypoints. (T : 2D transformations, so project 3D keypoints by dropping the z values)
  - **Key Prior Loss** (*LL*)**:** This loss encourages the estimated image-specific **keypoints** *xd,k* to **spread out across the face region**, instead of crowding around a small neighborhood. Distance is computed between the keypoint pairs and penalized if the distance is below some threshold.

$$\mathcal{L}_L = \sum_{i=1}^{K} \sum_{j=1}^{K} \max(0, D_t - \|x_{d,i} - x_{d,j}\|_2^2) + \|Z(x_d) - z_t\|$$


[1] Source Image Feature Extraction
(a)
Appearance feature extractor F
Appearance feature f_s
Canonical keypoint detector L
K canonical keypoints $\{x_{c,k}\}$ & K canonical Jacobians $\{J_{c,k}\}$
Source image s
Head pose estimator H
R_s, t_s
T
Expression deformation estimator Δ
Expression deformation $\{\delta_{s,k}\}$
K source keypoints $\{x_{s,k}\}$ & Jacobians $\{J_{s,k}\}$
(b)
Driving image d
Head pose estimator H
R_u, t_u
R_d, t_d
T
Expression deformation estimator Δ
Expression deformation $\{\delta_{d,k}\}$
K driving keypoints $\{x_{d,k}\}$ & Jacobians $\{J_{d,k}\}$
[2] Driving Video Feature Extraction
[3] Video Synthesis
Warped feature $w(f_s)$
Generator G
Output image y
flow w_1
flow w_2
flow w_K
f_s
$w_1(f_s)$
$w_K(f_s)$
Motion field estimator M
Flow composition mask m
flow w_1
flow w_2
…
flow w_K
Σ
Composited flow field w

- **Head Pose Loss (***LH***)**: *L*1 distance is computed between the estimated **head pose** *Rd* and the one predicted by a pre-trained estimator *R̄ d*. This approximation is as good as the pre-trained model head pose estimator. $\mathcal{L}_H = \|R_d - \bar{R}_d\|_1$

- **Deformation Prior Loss (***L*△**):** *δd,k* is the **deviation** from the canonical keypoints. Their magnitude should not be too large, L△=||*δd,k*||1. *

Experiments

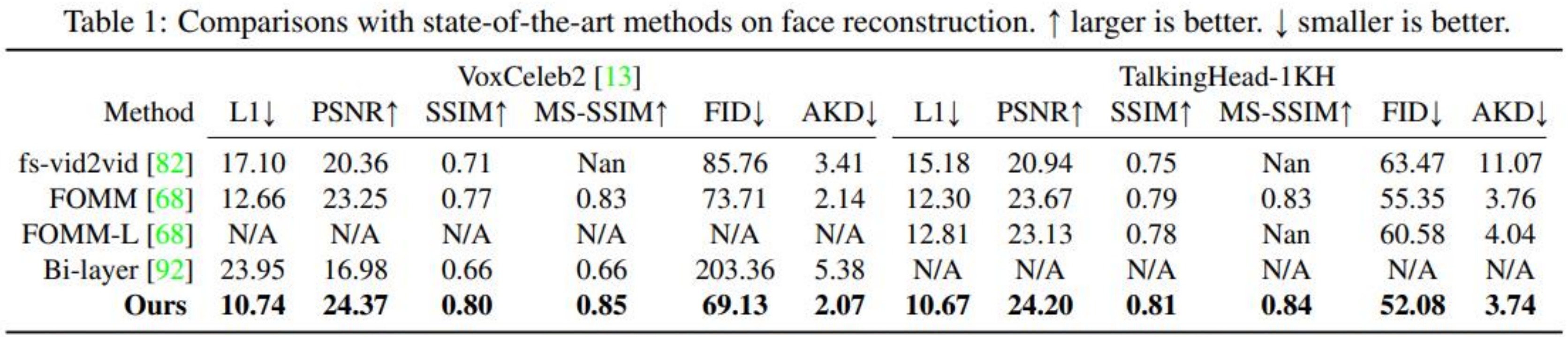


Table 1: Comparisons with state-of-the-art methods on face reconstruction. ↑ larger is better. ↓ smaller is better.

	VoxCeleb2 [13]						TalkingHead-1KH					
Method	L1↓	PSNR↑	SSIM↑	MS-SSIM↑	FID↓	AKD↓	L1↓	PSNR↑	SSIM↑	MS-SSIM↑	FID↓	AKD↓
fs-vid2vid [82]	17.10	20.36	0.71	Nan	85.76	3.41	15.18	20.94	0.75	Nan	63.47	11.07
FOMM [68]	12.66	23.25	0.77	0.83	73.71	2.14	12.30	23.67	0.79	0.83	55.35	3.76
FOMM-L [68]	N/A	N/A	N/A	N/A	N/A	N/A	12.81	23.13	0.78	Nan	60.58	4.04
Bi-layer [92]	23.95	16.98	0.66	0.66	203.36	5.38	N/A	N/A	N/A	N/A	N/A	N/A
Ours	**10.74**	**24.37**	**0.80**	**0.85**	**69.13**	**2.07**	**10.67**	**24.20**	**0.81**	**0.84**	**52.08**	**3.74**

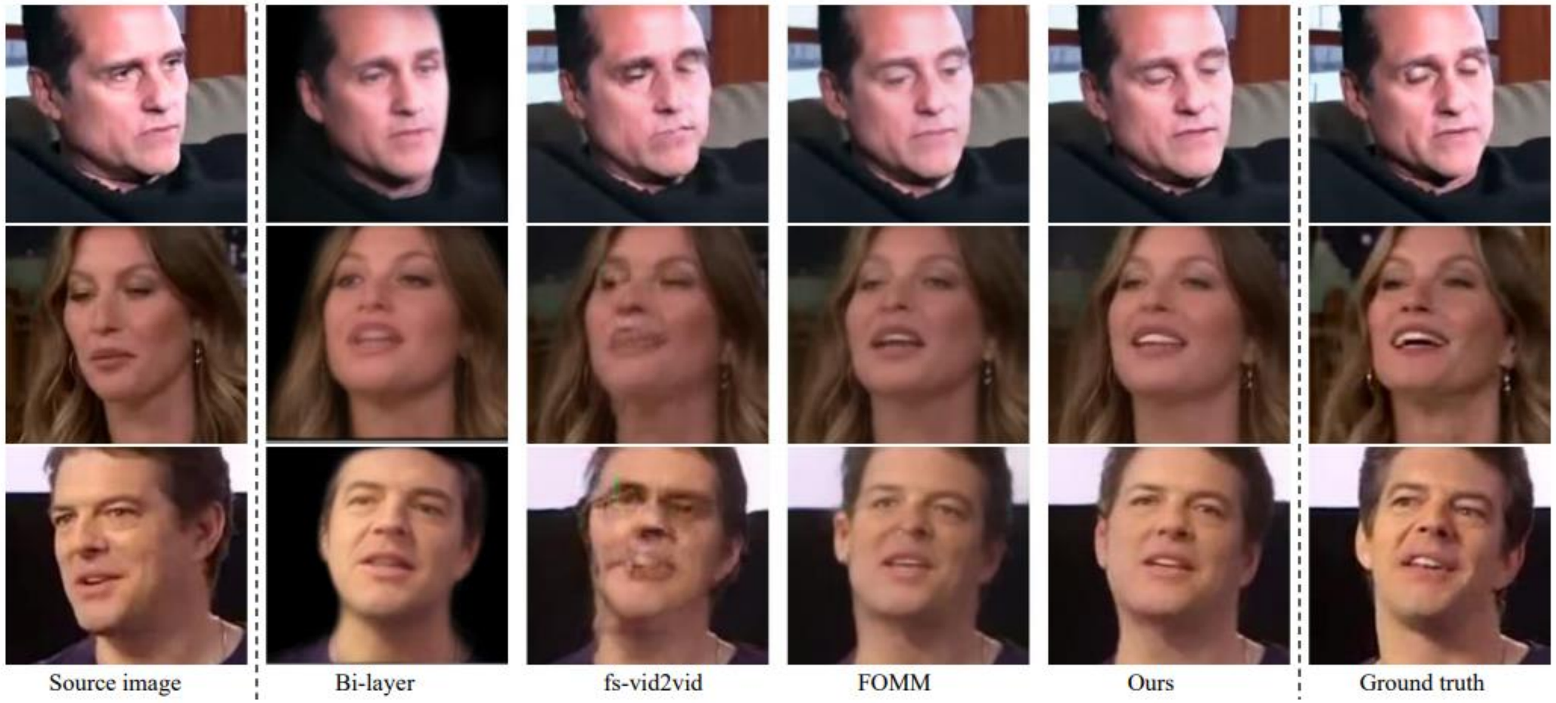


Figure 6: Qualitative comparisons on the Voxceleb2 dataset [13]. Our method better captures the driving motions.

Experiments

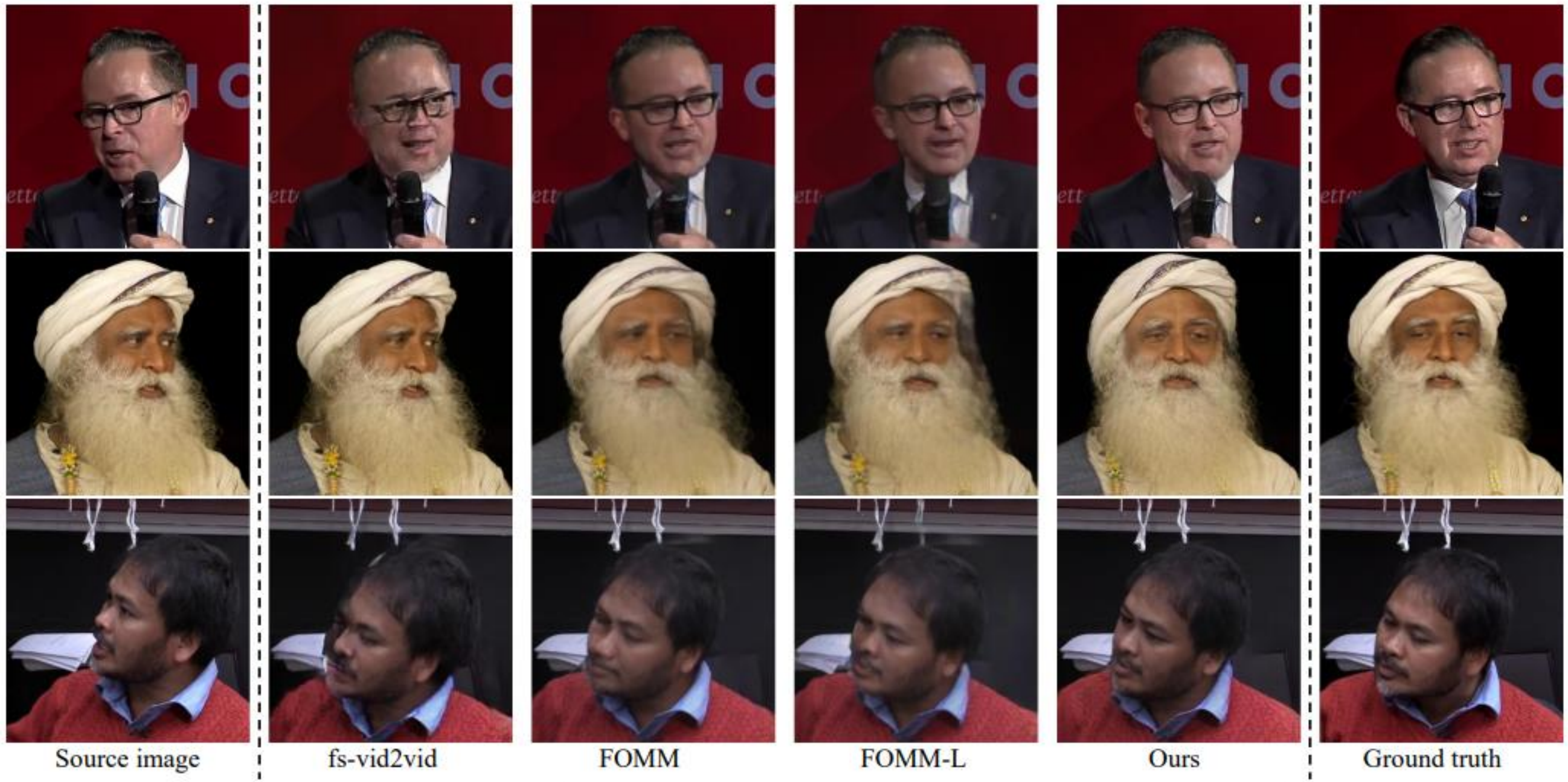


Figure 7: Qualitative comparisons on the TalkingHead-1KH dataset. Our method produces more faithful and sharper results.

Experiments

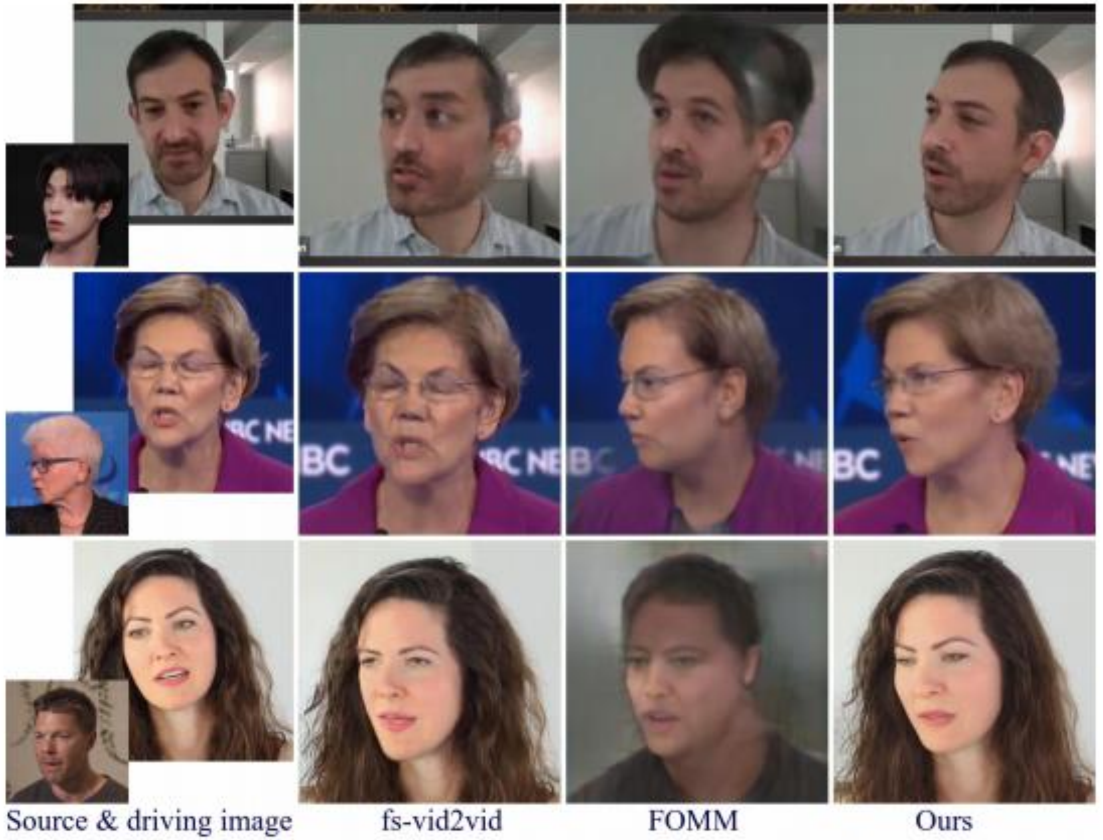


Figure 8: Qualitative comparisons for cross-subject motion transfer. Ours can capture the motion and preserve the identity better.

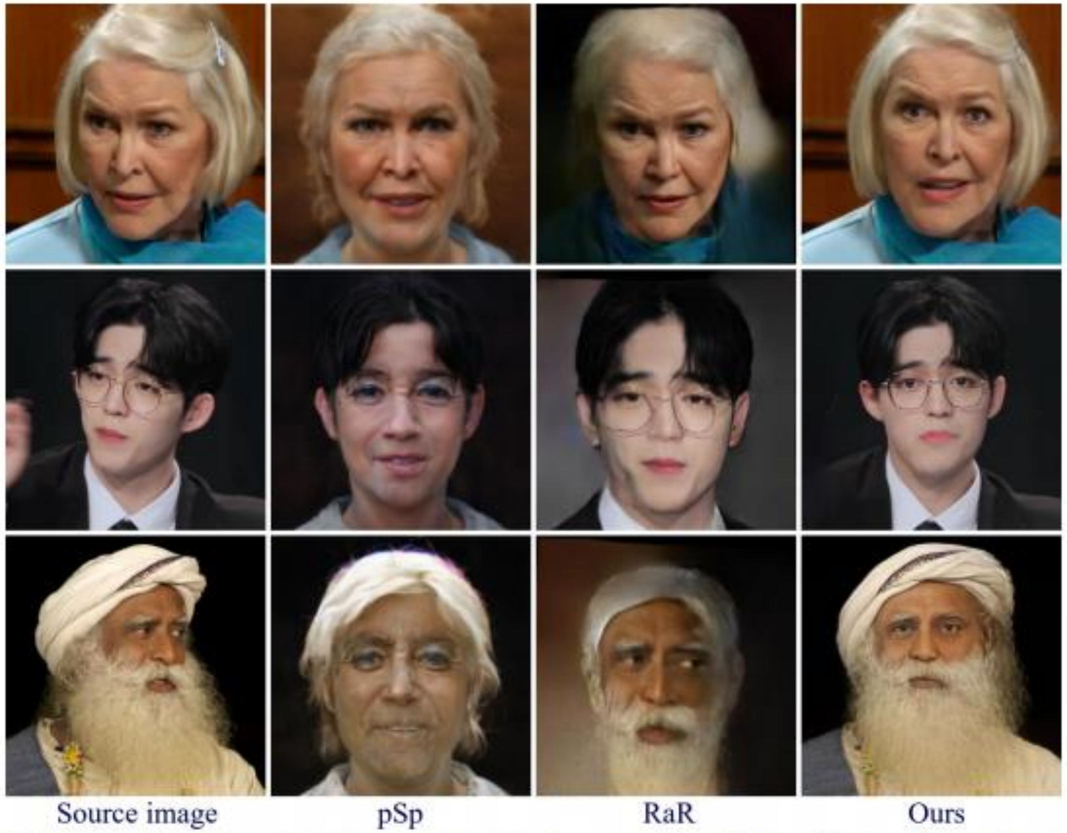


Figure 9: Qualitative comparisons for face frontalization. Our method more realistically frontalizes the faces compared to others.

Ablation study

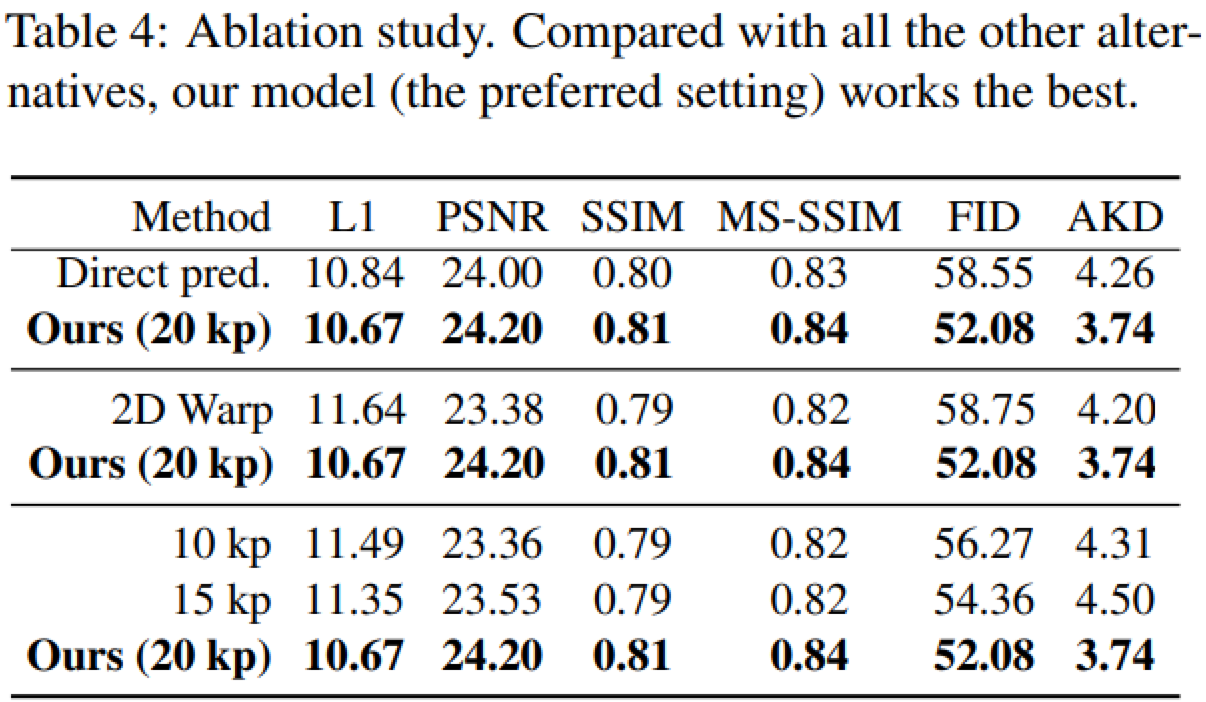
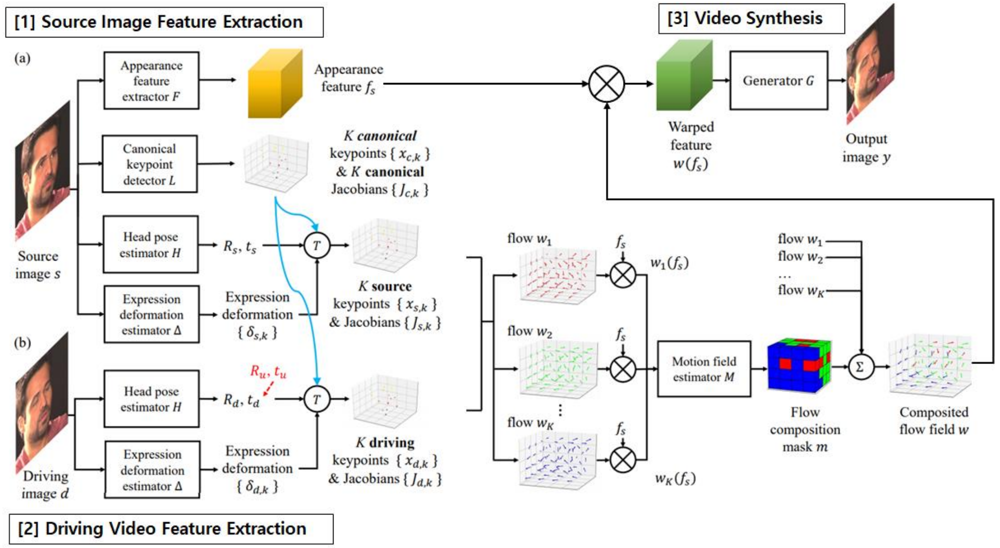


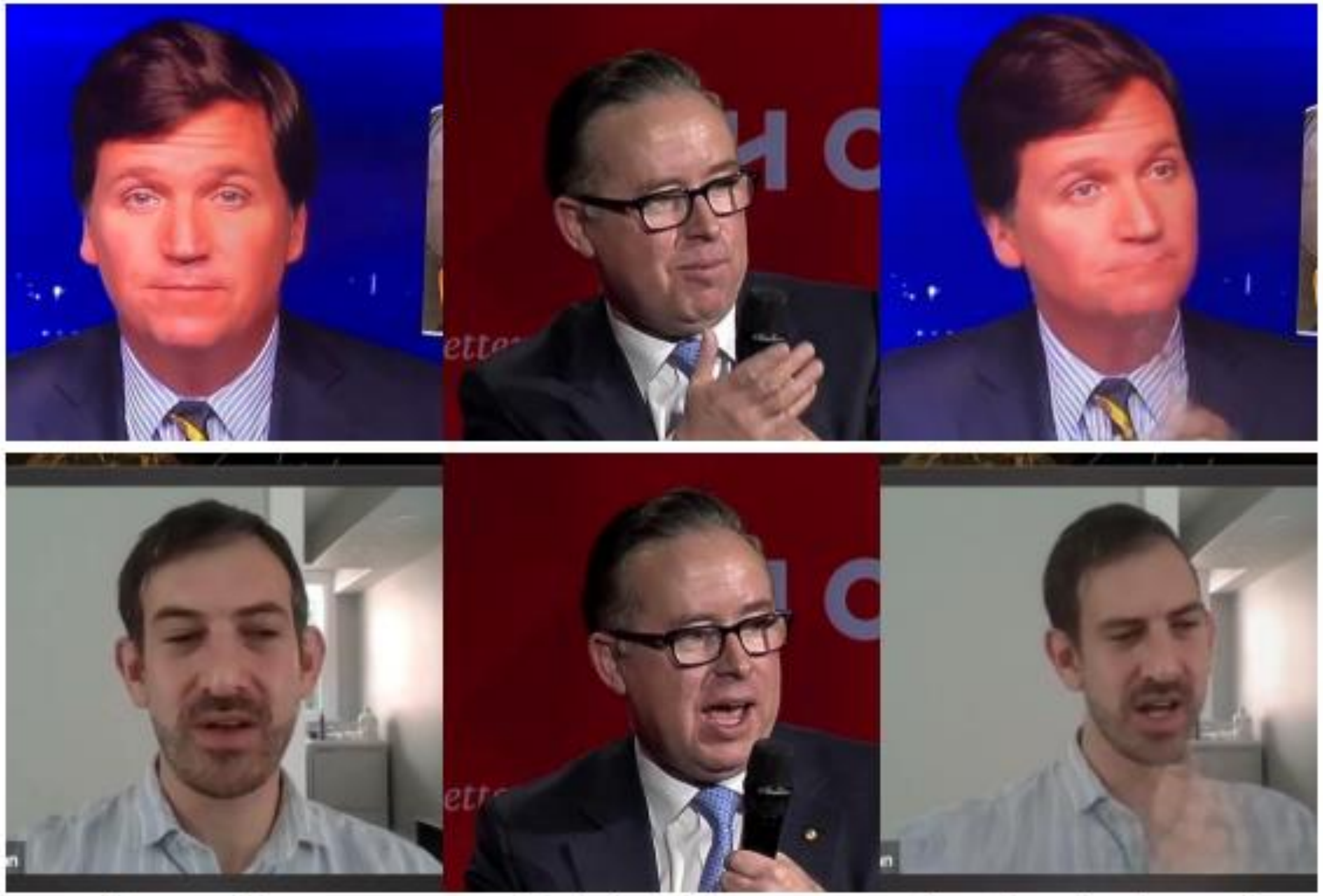
Table 4: Ablation study. Compared with all the other alternatives, our model (the preferred setting) works the best.

Method	L1	PSNR	SSIM	MS-SSIM	FID	AKD
Direct pred.	10.84	24.00	0.80	0.83	58.55	4.26
Ours (20 kp)	**10.67**	**24.20**	**0.81**	**0.84**	**52.08**	**3.74**
2D Warp	11.64	23.38	0.79	0.82	58.75	4.20
Ours (20 kp)	**10.67**	**24.20**	**0.81**	**0.84**	**52.08**	**3.74**
10 kp	11.49	23.36	0.79	0.82	56.27	4.31
15 kp	11.35	23.53	0.79	0.82	54.36	4.50
Ours (20 kp)	**10.67**	**24.20**	**0.81**	**0.84**	**52.08**	**3.74**



[1] Source Image Feature Extraction
(a)
Appearance feature extractor F
Appearance feature f_s
Canonical keypoint detector L
K ***canonical*** keypoints $\{x_{c,k}\}$ & K **canonical** Jacobians $\{J_{c,k}\}$
Head pose estimator H
R_s, t_s
T
Source image s
Expression deformation estimator Δ
Expression deformation $\{\delta_{s,k}\}$
K **source** keypoints $\{x_{s,k}\}$ & Jacobians $\{J_{s,k}\}$
(b)
Head pose estimator H
R_u, t_u
R_d, t_d
T
Expression deformation estimator Δ
Expression deformation $\{\delta_{d,k}\}$
K **driving** keypoints $\{x_{d,k}\}$ & Jacobians $\{J_{d,k}\}$
Driving image d
[2] Driving Video Feature Extraction
[3] Video Synthesis
Warped feature $w(f_s)$
Generator G
Output image y
flow w_1
f_s
$w_1(f_s)$
flow w_2
f_s
flow w_K
f_s
$w_K(f_s)$
Motion field estimator M
Flow composition mask m
flow w_1
flow w_2
…
flow w_K
Σ
Composited flow field w

Failure cases



Source image Driving frame Synthesized result

Figure 14: Example failure cases. Our method still struggles when there are occluders such as hands in the image.

* While our model is in general robust to different situations, it cannot handle large occlusions well. For example, when the face is occluded by the person's hands or other objects, the synthesis quality will degrade

End

- 논문 : https://arxiv.org/abs/2011.15126
- 참고자료 : https://wandb.ai/ayush-thakur/face-vid2vid/reports/One-Shot-Free-View-Neural-Talking-Head-Synthesis-for-Video-Conferencing--Vmlldzo1MzU4ODc

- Related works
  - Fs-vid2vid (NeurIPS 2019): https://nvlabs.github.io/few-shot-vid2vid/
  - FOMM (NeurIPS 2019): https://arxiv.org/abs/2003.00196