15th IEEE International Conference on
Automatic Face and Gesture Recognition
(FG 2020)

Head2Head: Video-based Neural Head Synthesis

Mohammad Rami Koujan*, Michail Christos Doukas*, Anastasios Roussos, Stefanos Zafeiriou

16-20 November 2020

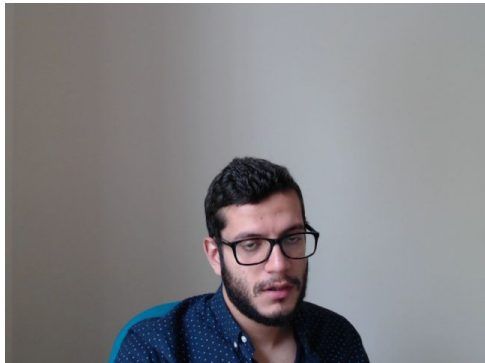
*Equal contribution



The problem of reenactment

Face reenactment: **Facial expressions** are transferred from the source to the target





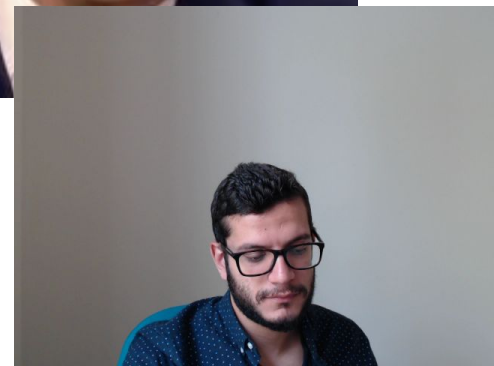
The problem of reenactment

Face reenactment: **Facial expressions** are transferred from the source to the target



Head reenactment: **Facial expressions** along with **head pose** and **eye gaze** are transferred





Head reenactment challenges

- Identity preservation problem



Head reenactment challenges

- Identity preservation problem
  > 3D face representation (NMFC) instead of facial landmarks



Head reenactment challenges

- Identity preservation problem
  > 3D face representation (NMFC) instead of facial landmarks

- Temporal coherence



Head reenactment challenges

- Identity preservation problem
  > 3D face representation (NMFC) instead of facial landmarks

- Temporal coherence
  > Sequential Generator
  > Temporal Discriminator conditioned on **facial flow**



Head reenactment challenges

- Identity preservation problem
  > 3D face representation (NMFC) instead of facial landmarks

- Temporal coherence
  > Sequential Generator
  > Temporal Discriminator conditioned on **facial flow**

- Mouth realism



Head reenactment challenges

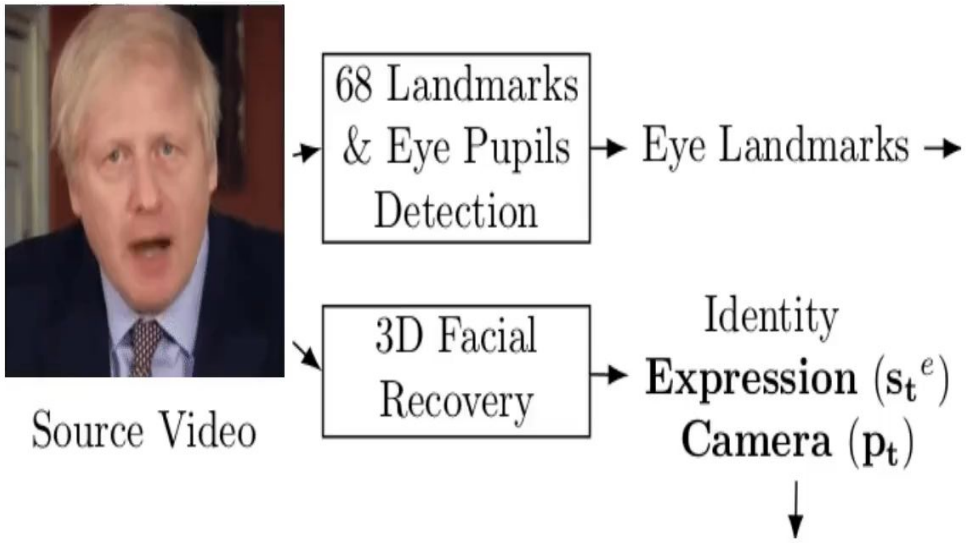
- Identity preservation problem
  > 3D face representation (NMFC) instead of facial landmarks

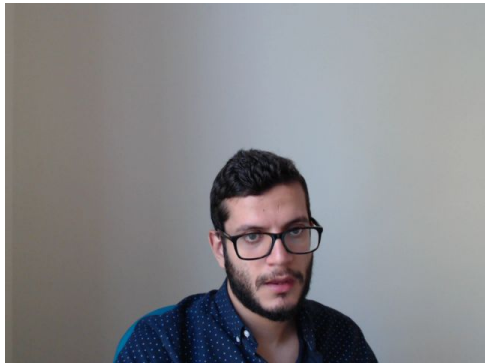
- Temporal coherence
  > Sequential Generator
  > Temporal Discriminator conditioned on **facial flow**

- Mouth realism
  > Mouth Discriminator



Head2Head pipeline

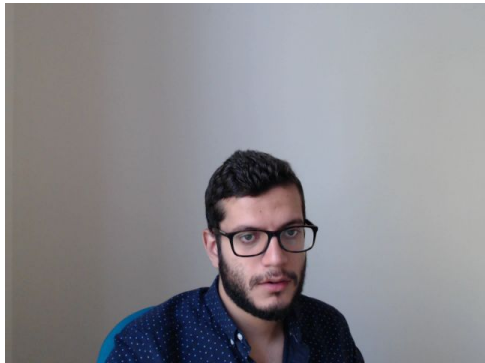




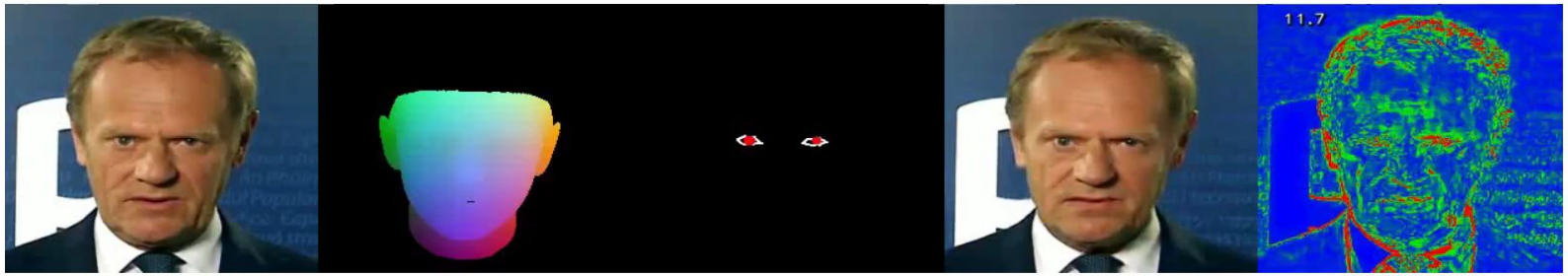
Head2Head reenactment results

Turnbull to Biden





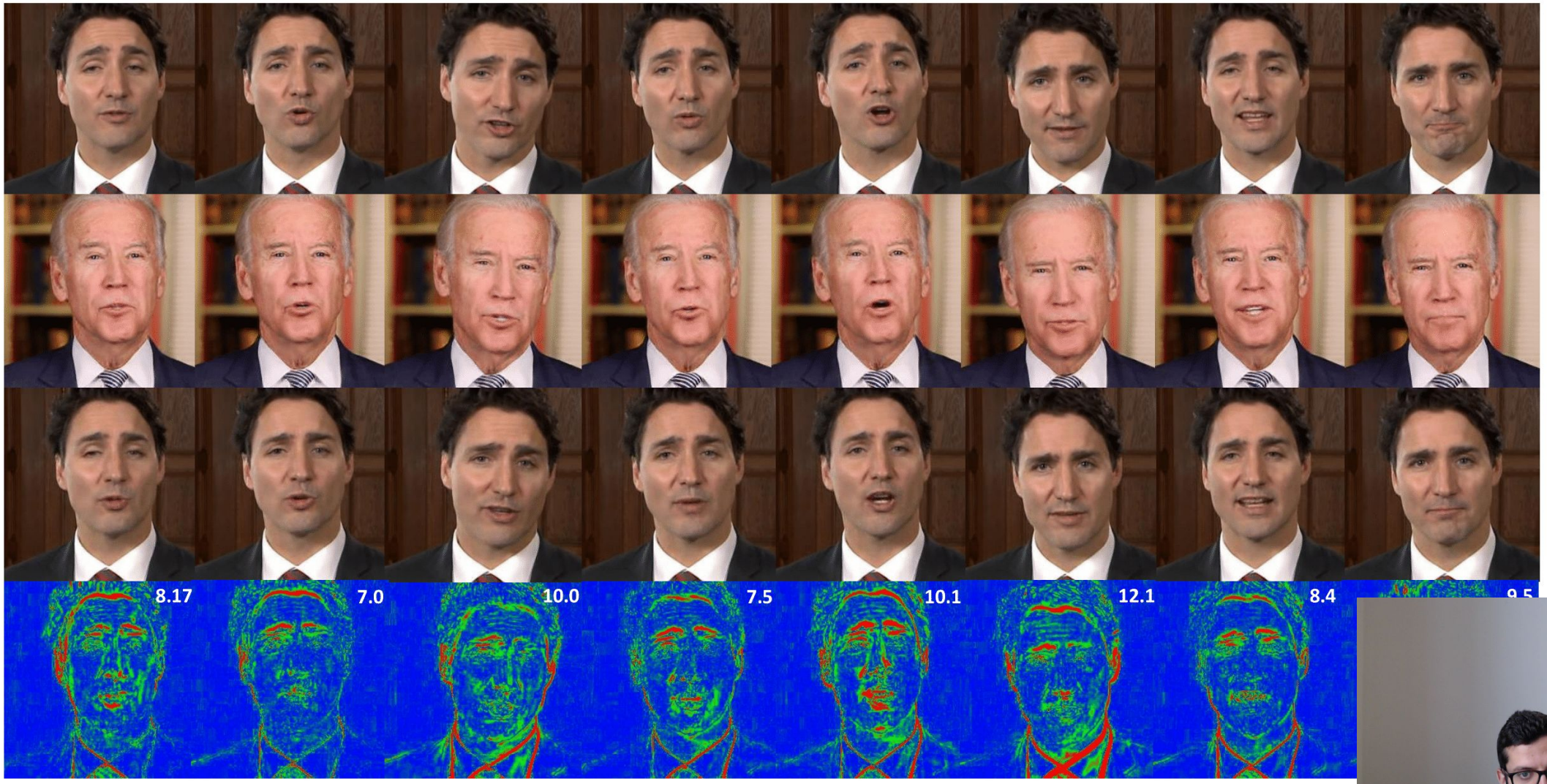
Quantitative evaluation – Self reenactment



- Self reenactment gives access to ground truth data
- Average pixel distance for evaluation of reconstruction perfo



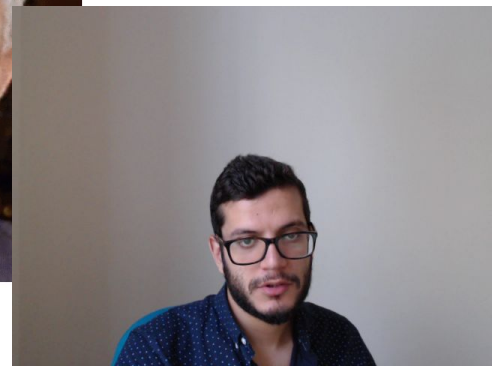
Quantitative evaluation - Cycle reenactment



The problem of landmarks in reenactment

- Qualitative comparison with **vid2vid** system, conditioned on facial sketches draw from **68 landmarks**.





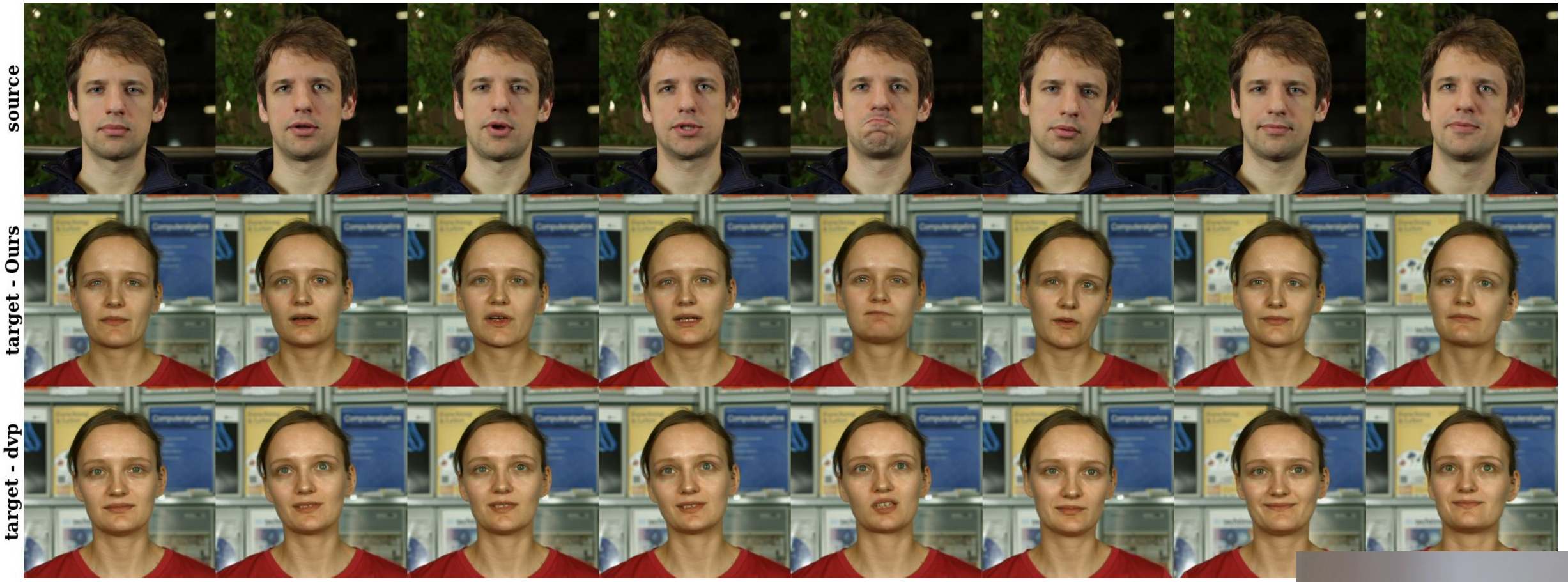
The problem of landmarks in reenactment

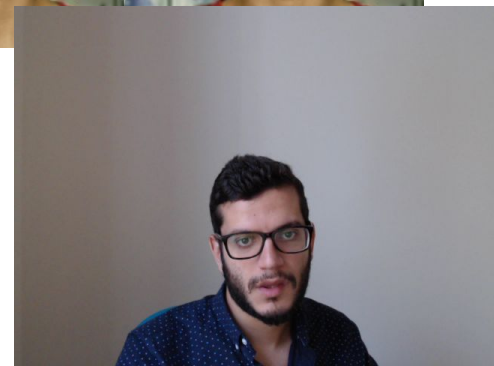
- Qualitative comparison with **vid2vid** system, conditioned on facial sketches draw from **68 landmarks**.

- We tackle the target identity preservation problem by conditioning synthesis on NMFC images.



Comparison with Deep Video Portraits





Ablation Study

TABLE I: Ablation study results.

(a) Average pixel distance obtained under a self-reenactment setup on the videos in figure 2 when combining our method with either FlowNet2 [47] or our facial flow.

Video	Ours (↓) FlowNet2	Ours (↓) Face-Flow
Justin	6.9	6
Joe	9.3	7.7
Merkel	9	7.5
Turnbull	9.4	6.5

(b) Average pixel distance around the mouth area, obtained under a self-reenactment scenario, with and without our dedicated mouth discriminator D_M.

Video	Ours (↓) w/o D_m	Ours (↓) w/ D_m
May	11.2	10.1
Justin	11	10.4
Turnbull	11.3	10.7
Obama	22.3	19.4



Ablation Study

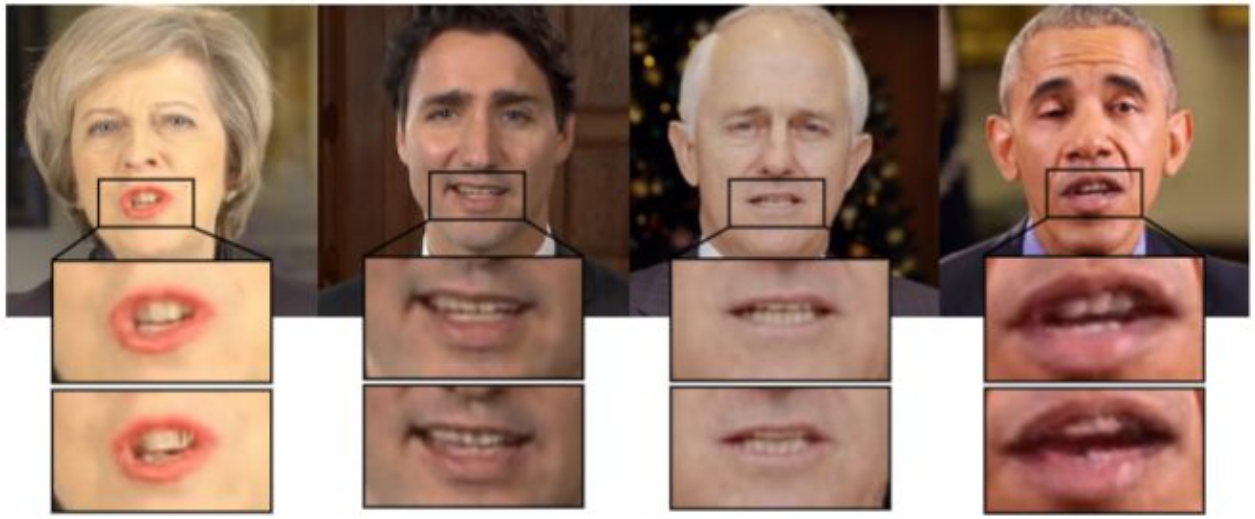
TABLE I: Ablation study results.

(a) Average pixel distance obtained under a self-reenactment setup on the videos in figure 2 when combining our method with either FlowNet2 [47] or our facial flow.

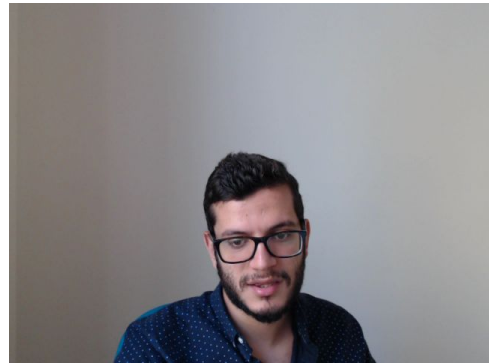
Video	Ours (↓) FlowNet2	Ours (↓) Face-Flow
Justin	6.9	6
Joe	9.3	7.7
Merkel	9	7.5
Turnbull	9.4	6.5

(b) Average pixel distance around the mouth area, obtained under a self-reenactment scenario, with and without our dedicated mouth discriminator D_M.

Video	Ours (↓) w/o D_m	Ours (↓) w/ D_m
May	11.2	10.1
Justin	11	10.4
Turnbull	11.3	10.7
Obama	22.3	19.4



The effect of Mouth Discriminator



User Study

- 95 participants
- *“On a scale of 1-5, how real does the video look?”*

TABLE II: Ratings on the self-reenactment task. Columns 1-5 show the number of participants that gave this rating, while column ”real” shows the percentage of people that rated the video with a 4 or 5.

Video	Our synthesised videos						Real videos					
	1	2	3	4	5	”real” (↑)	1	2	3	4	5	”real”
May	5	8	12	34	36	**74%**	3	6	6	26	54	84%
Turnbull	11	15	20	29	20	**52%**	2	6	14	40	33	77%
Putin*	4	10	18	36	27	**66%**	2	3	13	37	40	81%
Justin	5	8	18	36	28	**67%**	5	7	11	28	44	76%
dvp*	3	11	14	26	41	**71%**	9	12	21	29	24	56%
average	6	10	16	32	30	**66%**	4	7	13	32	39	**75%**

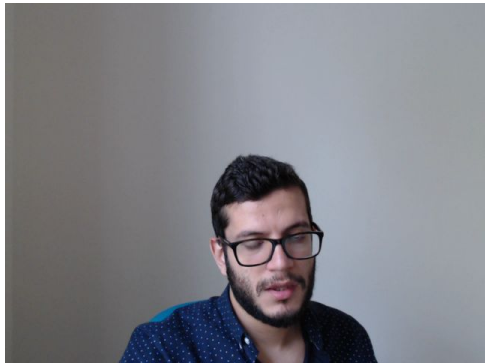
TABLE III: Ratings on the face-reenactment and full-reenactment tasks.

Video	Our synthesised videos						Real videos					
	1	2	3	4	5	”real” (↑)	1	2	3	4	5	”real”
Merkel (face-reenact.)	16	21	20	22	16	**40%**	1	6	17	31	40	**75%**
Justin (full-reenact.)	8	12	25	39	11	**53%**	5	7	11	28	44	76%
dvp* (full-reenact.)	37	19	10	22	7	**31%**	9	12	21	29	24	56%
average (full-reenact.)	22	16	18	30	9	**42%**	7	10	16	28	34	**66%**



Conclusions

- A **robust** and **accurate full-head re-enactment** approach is proposed
- Our framework **seamlessly** transfer **facial expressions**, **head pose** and **eye movements** from a source to a target video
- A **photo-realistic** synthesis is performed with the aid of facial dynamics and mouth discriminators and a novel 3D-based facial conditioning



Thank you!

Code is available here:

https://github.com/michaildoukas/head2head

