Personalized Visual Dubbing through Virtual Dubber and Full Head Reenactment (Supplementary Material)

Bobae Jeon[1], Eric Paquette[2], Sudhir Mudur[1], Tiberiu Popa[1]

[1]Concordia University, Montreal, Canada
[2]École de technologie supérieure, Montreal, Canada

This document contains additional diagrams, information, and results, which were not included in the main paper. Figure 1 shows the detailed steps of our preprocessing stage, as discussed in Section 2.1 of the paper. Figure 2 illustrates details about compositing, related to the full-head reenactment (Section 2.2). Figure 3 visualizes our identity swapping network architecture. Figure 4 presents qualitative results of our pipeline compared to other methods, showing consecutive frames to evaluate expressiveness and temporal continuity.

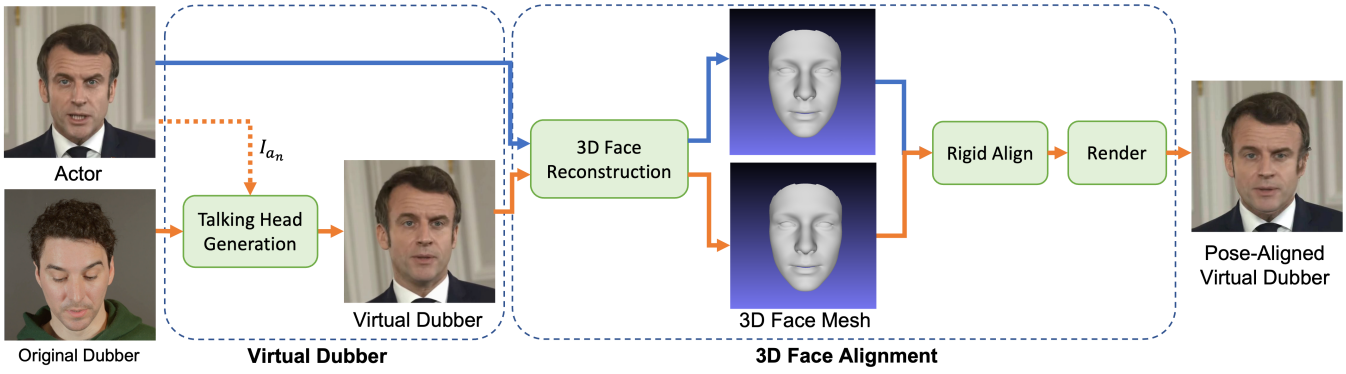

Figure 1: ***Details of our preprocessing stage.*** *Firstly, we select a front-facing frame of an actor with a neutral expression and use it along with the original dubber video for the talking head generation. This process generates the virtual dubber, a synthesized version of the actor that mimics the dubber expressions. However, the virtual dubber follows the head pose of the original dubber, whereas it should match the actor's. Therefore, we perform 3D face reconstruction, generating a face mesh with vertices and 3D facial landmarks. Using these, we rigid align the virtual dubber's face mesh to match the head pose of the actor's. Finally, we render the aligned virtual dubber face onto the actor frames.*

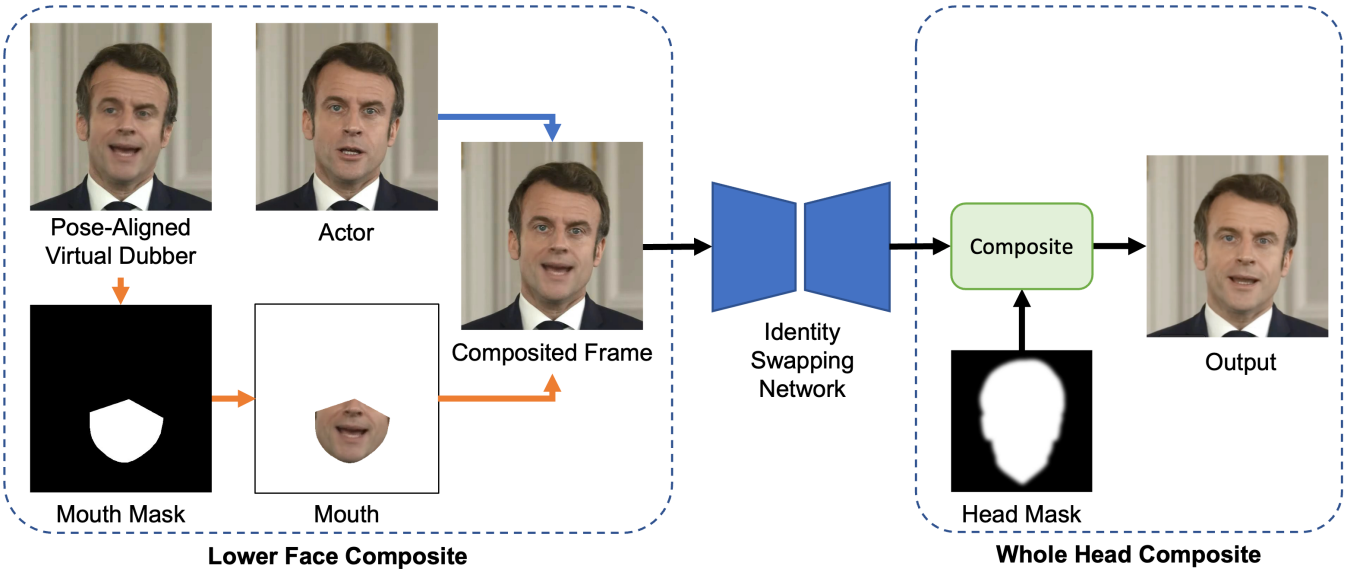

Figure 2: ***Details of the full-head reenactment.*** *Despite the good quality virtual dubber's mouth shape, the whole face of the virtual dubber often contains detrimental artifacts, such as inconsistent eyes and eyebrows. As such, in a first compositing step, we paste only the mouth region of the virtual dubber on top of the actor frame before sending it to the identity swapping network. After the network, we have a second compositing step which, this time, pastes the whole head, greatly reducing artifacts such as disconnected wrinkles and the double chin problem outlined in the paper.*

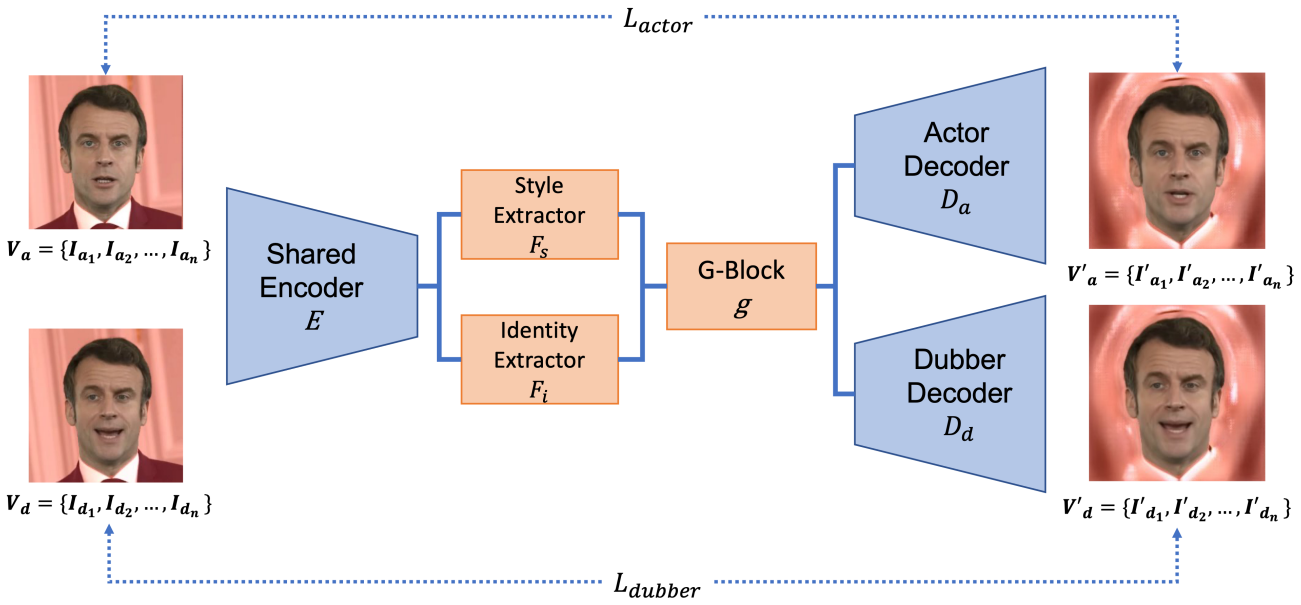

Figure 3: ***Identity swapping network architecture.*** *We train this network to reconstruct the hair, face, ears, and neck, using a head mask, which is applied to the shown images. The masked out background and clothes are depicted here with a red overlay.*

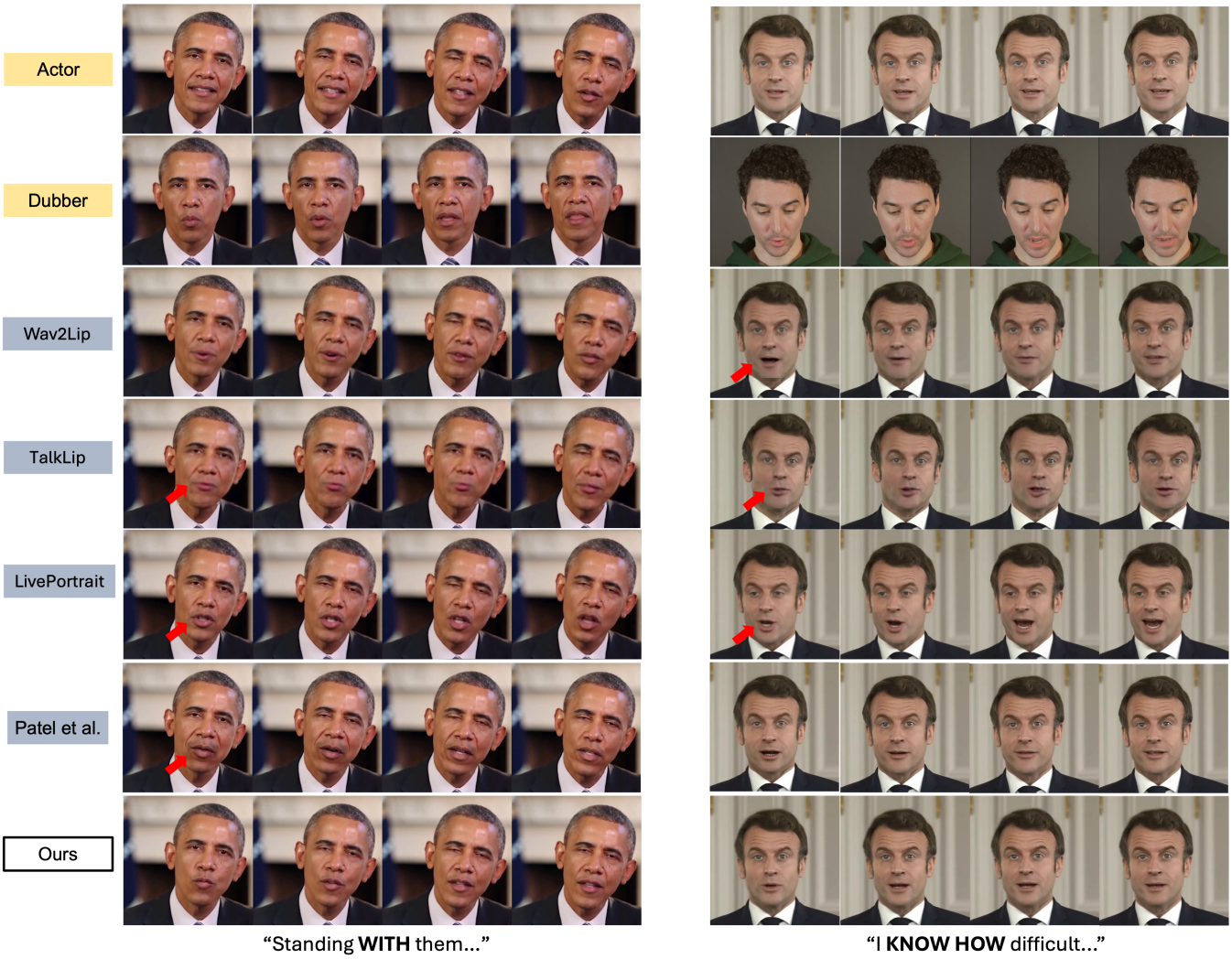

Figure 4: ***Qualitative evaluation.*** *(left) Our method recreates mouth puckering, which other methods fail to capture adequately. (right) The dubber's head is tilted downward, yet our method generates the expressions with minimal loss of detail.*