Outstanding Orthodontist: No More Artifactual Teeth in Talking Face

Paper ID: 5681
Zibo Su, Ziqi Zhang , Kun Wei* , Xu Yang and Cheng Deng
Xidian University, No.2 South Taibai Road, Xi'an 710071, Shaanxi, China

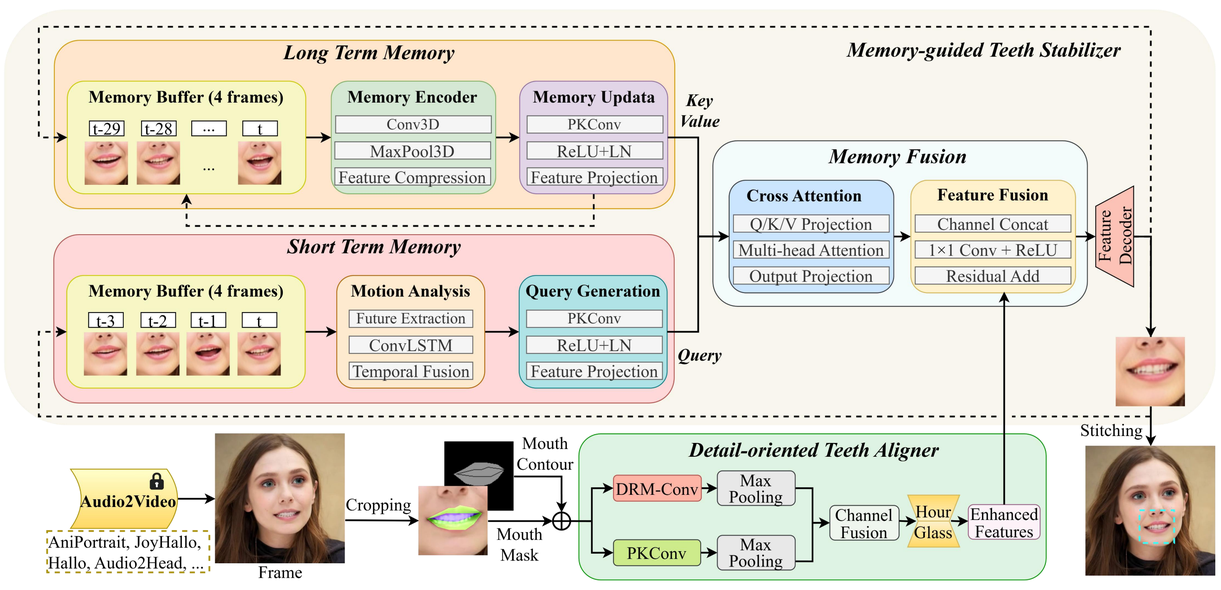
Abstract

Audio-driven talking face synthesis (TFS) enables the creation of realistic speaking videos by combining a single facial image with a speech audio clip. Unlike other facial features that naturally deform during speech, teeth represent unique rigid structures whose shape and size should remain constant throughout the video sequence. However, current methods often produce temporal inconsistencies and artifacts in the teeth region, resulting in a less realistic appearance of the generated videos. To address this, we propose **OrthoNet**, a plug-andplay framework designed to eliminate unrealistic teeth effects in audio-driven TFS. Our method introduces a **Detail-oriented Teeth Aligner** module, designed to preserve teeth details and adapt to their shape. It works with a **Memory-guided Teeth Stabilizer** that integrates a long-term memory bank for global teeth structure and a short-term memory module for local temporal dynamics. Through this framework, OrthoNet acts like an orthodontist for existing Audio2Video methods, ensuring that teeth maintain natural rigidity and temporal consistency even under varying degrees of teeth occlusion. Extensive experiments demonstrate that our method makes the teeth in generated videos appear more natural during speech, significantly enhancing the temporal consistency and structural stability of audio-driven video generation.

Motivation

A key underexplored challenge is maintaining temporal consistency and realism of teeth appearance under dynamic occlusion during speech. Since occlusion is common, the task shifts from image editing to inpainting, introducing two issues: **(1) Teeth inconsistency**—unnatural size changes across frames (Fig. a–d) caused by independent frame generation, destabilizing the rigid structure. **(2) Generation hallucination**—missing, distorted, or blurred teeth details (Fig. a, c) due to insufficient prior knowledge and complex lip–teeth interactions, producing uneven edges and unrealistic shapes. These flaws make videos appear fake, harming viewer experience.

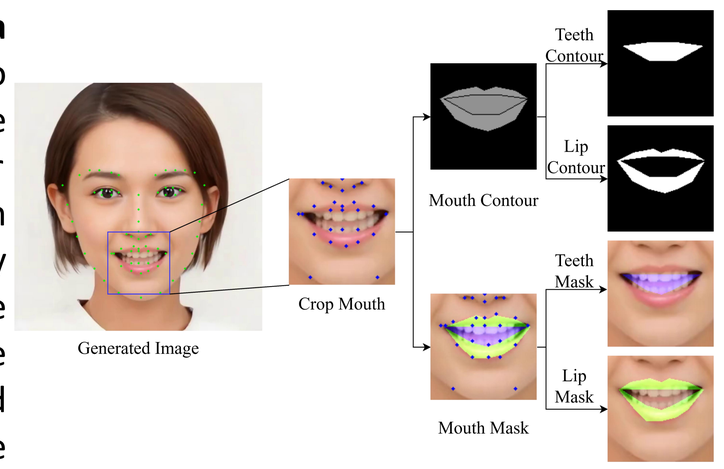
Method



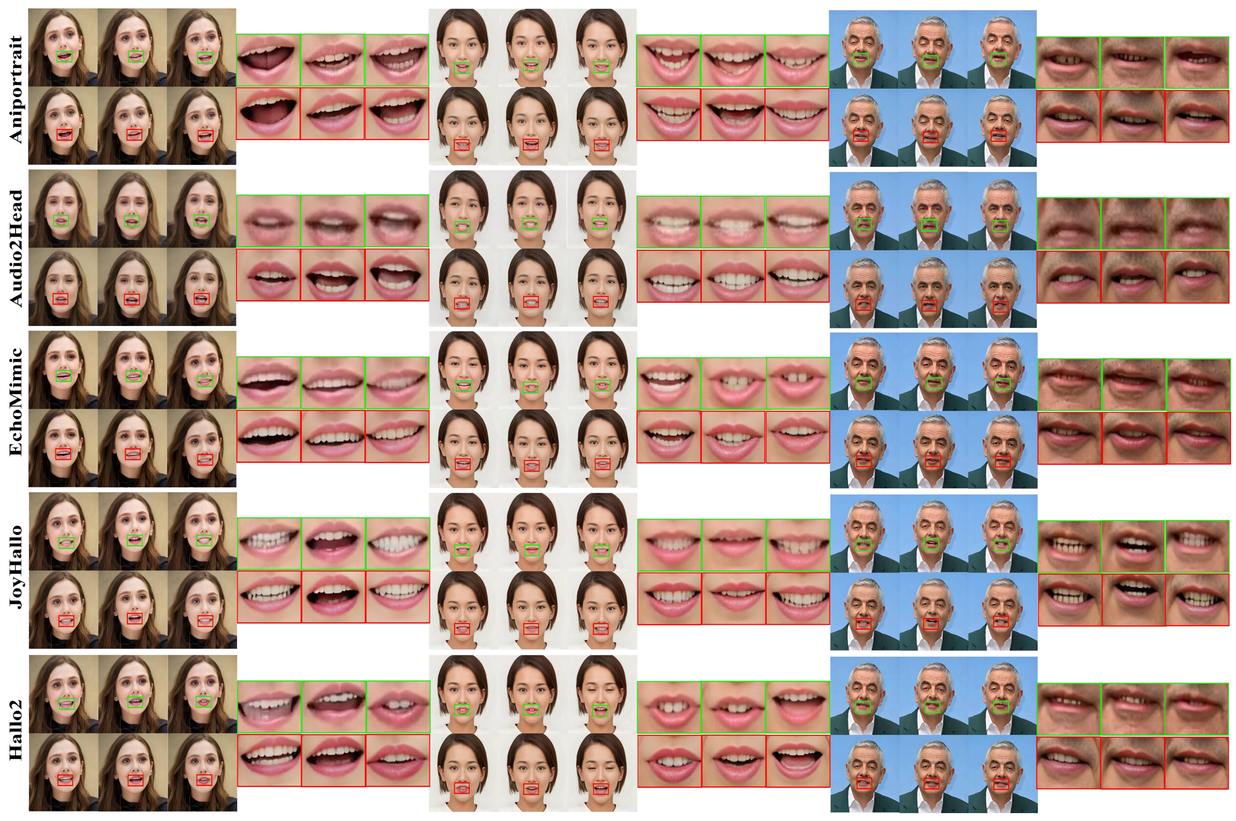
As shown in the image, we present **OrthoNet**, a plug-and-play framework for teeth restoration in talking face videos that operates as a virtual orthodontist. It has two main parts: **Detail-oriented Teeth Aligner** for extracting precise teeth features and **Memory-guided Teeth Stabilizer** for maintaining stability over time. Detail-oriented Teeth Aligner uses dualbranch: DRM-Conv preserves fine-grained teeth details, and PKConv captures the overall shape of the teeth, ensuring both small parts and the whole structure are accurately modeled. Memory-guided Teeth Stabilizer has a dual-memory stream inspired by post-orthodontic treatment. LTM maintains the temporal consistency of the teeth, and STM enables teeth adaptation to speech dynamics, similar to how orthodontists use different retainers to retain strength while adapting naturally during speech.

Experiment

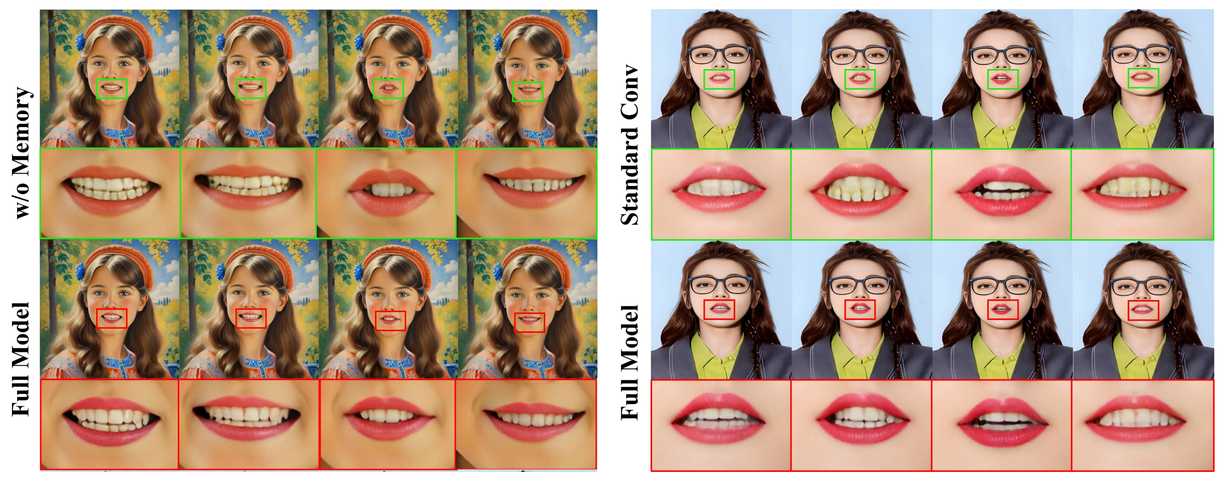
Details regarding data processing. For each video frame that is generated, face landmarks are detected in order to precisely locate the mouth region. This is achieved by identifying crucial points like the left and right corners of the mouth, the tip of the nose, and the jaw. Subsequently, the mouth region is cut out.



Moreover, masks for the teeth and lips, along with their contours, are created based on these identified key points. Eventually, the cropped regions and the generated masks are resized to a resolution of 96×96 pixels, which helps to guarantee consistent resolution for all the input data.



Qualitative comparison. In the images, **green boxes** highlight results generated by the original baseline methods, while **red boxes** show results after incorporating our proposed approach. Our OrthoNet framework consistently improves temporal stability and temporal consistency of teeth across different baseline methods, particularly under varying degrees of teeth occlusion.



Visualization of partial ablation study results. As shown in the image, without memory modules, the model exhibits irregular teeth variations and temporal instability, with sudden changes in teeth shape and size between consecutive frames and degradation of fine-grained details during mouth movements; additionally, the use of standard convolution leads to detail loss and unrealistic artifacts, such as blurred teeth edges and inconsistent gaps, failing to capture the natural structural patterns of real teeth during speech.

Conclusion

We propose **OrthoNet**, a plug-and-play framework for enhancing teeth realism in audio-driven TFS. It features a **Teeth Aligner** to preserve details during mouth movements and a **Teeth Stabilizer** that uses a memory system to ensure structural consistency. Easily integrated into existing systems, OrthoNet mitigates teeth hallucination and inconsistency, achieving notable gains in temporal stability and visual realism.