ABSTRACT

Previous studies have explored generating accurately lip-synced talking faces for arbitrary targets given audio conditions. However, most of them deform or generate the whole facial area, leading to non-realistic results. In this work, we delve into the formulation of altering only the mouth shapes of the target person. This requires masking a large percentage of the original image and seamlessly inpainting it with the aid of audio and reference frames. To this end, we propose the Audio-Visual Context-Aware Transformer (AV-CAT) framework, which produces accurate lip-sync with photo-realistic quality by predicting the masked mouth shapes. Our key insight is to exploit desired contextual information provided in audio and visual modalities thoroughly with delicately designed Transformers. Specifically, we propose a convolution-Transformer hybrid backbone and design an attention-based fusion strategy for filling the masked parts. It uniformly attends to the textural information on the unmasked regions and the reference frame. Then the semantic audio information is involved in enhancing the self-attention computation. Additionally, a refinement network with audio injection improves both image and lip-sync quality. Extensive experiments validate that our model can generate high-fidelity lip-synced results for arbitrary subjects.

CCS CONCEPTS
• Computing methodologies → Animation; Neural networks.

KEYWORDS
Lip-Sync, Transformer, Audio-Visual Learning.

1 INTRODUCTION

Driving human mouth movements with speech audio is in great need in the field of virtual human creation, entertainment, the film industry, and the detection of deep fakes. The ability to drive arbitrary target video can vastly improve the applicability of such models. However, previous studies focusing on arbitrary targets mostly operate on the whole face [Chen et al. 2020, 2018, 2019; Jamaludin et al. 2019; Ji et al. 2022; Song et al. 2018; Sun et al.
2 RELATED WORK

2.1 Audio-Driven Lip-Synced Talking Face

Driving a target portrait with speech audio has long been a popular research topic in the computer graphics and computer vision communities. Generally, previous approaches can be classified into person-specific and person-agnostic settings.

2.1.1 Person-Specific Lip-Sync Synthesis. Most person-specific modeling methods generate high video quality [Guo et al. 2021; Ji et al. 2021; Lahiri et al. 2021; Li et al. 2021; Liu et al. 2022; Song et al. 2020; Suwajanakorn et al. 2017; Thies et al. 2020]. Some of them focus on altering the mouth areas for photo-realistic results [Song et al. 2020; Suwajanakorn et al. 2017; Thies et al. 2020]. To synthesize high-quality talking face videos of Obama, Song et al. [2020] introduce 3DMM [Blanz and Vetter 1999] to translate the speech content into the landmarks of the mouth area. They achieve similar results for mouth rendering. On the other hand, Guo et al. [2021] and Liu et al. [2022] use NeRF [Ji et al. 2021] to model 2D landmarks and Lahiri et al. [2021] use neural rendering to synthesize the whole talking faces. All the above results enjoy high video quality, however, their methods cannot be applied to a person with few (seconds of) data, which limits their applicability.

2.1.2 Person-Agnostic Lip-Sync Synthesis. Other methods focus on the person-agnostic setting where only one image or a short clip is provided as a reference. Speech2Vid [Jamaludin et al. 2019] studies the problem for the first time in an end-to-end manner. Then Chen et al. [2019] leverage 2D facial landmarks as guidance. Zhou et al. [2020] propose to use 3D landmark displacements and synthesize videos including head poses and blinks. Zhou et al. [2021] choose our Audio-Visual Context-Aware Transformer framework generates high-fidelity lip-sync results for arbitrary subjects.
Transformer-hybrid backbone network and a Refinement Network. The Reference image $I_t$ is processed in the reference branch with Basic Transformer stages (Basic-TS). The audio information is processed in the audio branch. Both this information is fused to the CCF-Transformer stages (CCF-TS) in the main branch. The figures are selected from VoxCeleb [Nagrani et al. 2017] ©Visual Geometry Group (CC BY).

Specifically, Wav2Lip [Prajwal et al. 2020] focuses on the mouth areas and inpaints the lower half of the face using speech audio and another set of reference frames. However, they rely on CNN-based UNet with skip-connections and concatenate audio information to the bottleneck of the network, which leads to blurry results. In the pursuit of high-quality and person-agnostic lip-sync, we adopt a similar setting as Wav2Lip and identify the problem as masked lip-sync prediction. Our method can synthesize clearly more realistic results than previous person-agnostic methods.

## 2.2 Image Synthesis with Transformers

The Transformer architecture [Vaswani et al. 2017] has received growing interest from various tasks in computer vision [Bao et al. 2021; Chang et al. 2022; Dosovitskiy et al. 2021; Esser et al. 2021a; He et al. 2021; Li et al. 2022; Liu et al. 2021]. Recently, Transformer has also extended its reach to low-level vision tasks, such as image synthesis [Esser et al. 2021a; Van Den Oord et al. 2017] and image restoration [Li et al. 2022; Yang et al. 2020, 2021; Yu et al. 2021; Zheng et al. 2021].

For conditional image restoration tasks (e.g., super-resolution and inpainting), Yang et al. [2020] propose a texture transformer network for image super-resolution by leveraging different high/low-resolution references. Yang et al. [2021] present a transformer-based framework for continuous image super-resolution with implicit function. Li et al. [2022] introduce a mask-aware transformer (MAT) for large-hole inpainting. Similar to the Transformer-based approaches mentioned above, our AV-CAT also aims to fill in the masked region and synthesize high-fidelity images with reasonable textural appearance and correct semantic mouth movements.

## 3 METHODOLOGY

In this section, we introduce our Audio-Visual Context-Aware Transformer (AV-CAT) as illustrated in Fig. 2. The whole framework contains a convolution-Transformer-hybrid backbone network that accounts for information fusion at coarse level and a convolutional Refinement Network that fixes details. We will first introduce our masked lip-sync prediction training setting in Section 3.1. Then we detailledly depict the design of the backbone network and information fusion strategies. Finally, we introduce the Refinement Network and the training objectives.

### 3.1 Training Setting

Our masked lip-sync prediction setting is similar to Wav2Lip [Prajwal et al. 2020]. Given a training video $V = \{I_1, \ldots, I_T\}$ with corresponding audio clip $a = \{a_1, \ldots, a_T\}$, our model processes one target frame $I_t \in V$ at a time. We mask out most areas on the lower half of $I_t$ to $I^m_t = M \ast I_t$ with a fixed mask $M$. The training goal is to recover $I_t$ with the corresponding audio $a_t$. To recover the mouth-shape irrelevant information such as textures on the face as well as the masked hair and backgrounds, a reference frame $I_r \in V$ is involved. We expect the reference frame to provide textual information only and avoid exacting the mouth shapes directly from it. Thus it is sampled from the same video but at a different timestamp. The reference frame and the target can be selected as the same during testing.
3.2 Audio-Visual Context-Aware Transformer

The convolution-Transformer-hybrid backbone network consists of three branches: the main branch, reference branch and the audio branch. They target processing the target frame, reference frame, and audio information, respectively. The main branch consists of a convolutional encoder head, 5 Cross-Modal Contextual Fusion Transformer (CCF-Transformer) stages, and a convolutional decoder. Specifically, the CCF-Transformer is built upon the Basic Transformer blocks modified from Swin Transformers [Liu et al. 2021].

We design our reference branch without involving extra parameters. The audio branch encodes an audio feature with encoder $E_a$ and decoder $F_a$. Both $F_a$ and intermediate features of the Basic-Transformer blocks are sent into the CCF-Transformer blocks for information fusion.

3.2.1 Convolutional Encoder and Decoder. We choose to preserve 4 layers of convolution operations for the low-level representation downsampling in encoder $E_r$ and another 4 for upsampling in decoder $Net_d$. It has been verified that the inductive bias at early layers is effective for Transformer learning [Li et al. 2022; Raghu et al. 2021; Xiao et al. 2021]. Specifically, the convolutional encoder $E_r$ processes an image to a feature map $F$ with 1/4 the size of $I_r$. It encodes the target and the reference frame to feature maps $F_r$ and $F_r$, respectively.

3.2.2 Basic Transformer Structure. The reference feature map $F_r \in \mathbb{R}^{h \times w \times C_0}$ is then sent into our Basic Transformer of 5 stages, each containing several blocks. It is built based on Swin Transformer [Liu et al. 2021] with modifications on the hierarchical representations inspired by [Li et al. 2022], including the removal of the layer normalization. While Swin Transformer downsamples the feature patches at each stage by fusing neighboring $2 \times 2$ patches, we consider a downsampling-upsample strategy given our generative task property. At the beginning of the first three stages, we downsample the feature map with a stride-2 convolution with $2 \times 2$ kernel size. Differently, we upsample the feature maps at the last two stages with transposed convolutions.

The self-attentions is computed within local windows of size $w_w \times w_w$ as performed in [Liu et al. 2021], which saves computational cost. We adopt the standard multi-head self-attention (MSA) in the Basic Transformer blocks:

$$\text{Attention}(Q, K, V) = \text{Softmax}(QK^T / \sqrt{d_k})V,$$  

where $Q, K, V$ are all tokens within the same window. After each attention operation, the window is shifted by $(\lfloor w_w / 2 \rfloor, \lfloor w_w / 2 \rfloor)$ pixels.

We visualize the architecture of one Basic Transformer block in Fig. 3 (a). Note that all the learnable parameters in the Basic Transformer are jointly learned with CCF-Transformer. This is to ensure the same semantics among both Transformer blocks and memory saving. The block is modified according to [Li et al. 2022].

3.2.3 Cross-Modal Contextual Fusion Transformer Blocks. The target’s feature map $F_t \in \mathbb{R}^{h \times w \times C_0}$ is sent into CCF-Transformer which shares similar setups as the Basic Transformer, e.g., the resolution of each stage, the sliding window operation, and the computation of self-attentions. Differently, the CCF-Transformer receives extra contextual information from the reference branch and the audio branch: 1) The cross-frame reference information provided from the Basic Transformer blocks. It is fused to CCF-Transformer blocks. 2) The cross-modal audio information is mapped to the last 3 stages through a fully connected layer.

Detailedly, at each block of the CCF-Transformer, an additional multi-head mutual attention (MMA) is developed in parallel with MSA between the target feature map $F_t(k,l)$ and reference feature map $F_r(k,l)$, where $l$ and $k$ denote the $l$th block of the $k$th stage. In the MMA, the tokens from $F_t(k,l)$ are treated as queries and $F_r(k,l)$ as keys and values. This is to retrieve desired appearance information for inpainting missing areas. Then, the MMSA and MSA results are concatenated together and mapped to the original feature dimension by a fully connected layer.

Then for the last three stages of CCF-Transformer, the audio feature $f_a$ is mapped to $f_a(k,l)$ and sent into the Transformer blocks to operate on the fused feature mentioned above. The overall computation of one CCF-Transformer block can be written as:

$$z_i^f(k,l) = \text{FC}_i(\text{Cat}\{\text{MSA}(F_t(k,l)), \text{MMA}(F_r(k,l), F_t(k,l))\}),$$   

$$z_i^f(k,l) = \text{FC}_2(f_a(k,l) + z_i^f(k,l)) \quad (\text{if } k > 2),$$   

$$z_i^f(k,l) = z_i^f(k,l) + F_t(k,l),$$   

$$F_t(k,l+1) = \text{MLP}(z_i^f \{F_t(k,l)\}) + z_i^f.$$   

which renders the feature map $F_t(k,l+1)$ of the next block. In the equations, Cat[] denotes feature concatenation and FC_i are different fully connected layers that are not shared with Basic Transformer blocks. Notably, the intuition for only injecting three layers of audio information is that audio usually interacts with high-level visual features [Prajwal et al. 2020; Zhou et al. 2019]. Experiments show that injecting audio into all stages makes little difference.

The final output of the CCF-Transformer is sent into the convolutional decoder $Net_d$ to predict an intermediate result $I_p$.

3.3 Refinement Network

Note that though the convolution-Transformer-hybrid backbone (CTH backbone) manages to fuse effective information, it fails at handling fine-grained details. On the other hand, we identify that
UNet-like [Ronneberger et al. 2015] structures are suitable for improving roughly good results. Thus, we propose the Refinement Network $G_R$. It takes the $I_t^i$ predicted from the CTH backbone and the original masked target $I_t^M = M * I_t$ for generating more harmonic results with higher quality. Specifically, the input of the Refinement Network is $I_t^i = M * I_t + (1 - M) * I_t^P$ and it predicts $I_t^p = G_R(I_t^i)$.

Specifically, motivated by previous research on talking head generation that audio features can be effectively expressed with modulated convolution [Zhou et al. 2021], we adopt a similar protocol for enhancing cross-modal learning. We map audio feature $f_a$ to a style vector $s$ for each convolution operation in the UNet. For each value $W_{mnq}$ in any convolution kernel weight $W$, where $m$ channel-wise position of the input, $n$ is its position on the output channel and $q$ stands for spatial location, we modulate it according to $s$’s given its channel-wise position $m$:

$$W_{mnq}' = \frac{s_m \cdot W_{mnq}}{\sqrt{\sum_{m,q} (s_m \cdot W_{mnq})^2 + \epsilon}}$$

where $\epsilon$ is set as a small value to avoid numerical errors.

### 3.4 Learning Objectives

The Learning objectives are mainly the VGG and GAN losses that are leveraged in various generation tasks [Isola et al. 2017; Wang et al. 2018b; Zhou et al. 2021]. Normally the VGG loss is used for perceptual similarity. Here we extend the VGG loss by computing from the first layer of the convolution to the $N_{vgg}$’s layer. This accounts for low-level information reconstruction. The loss functions are written as:

$$L_{GAN} = \min \max_D (E_I \left[ \log D(I_t) \right] - E_p \left[ \log (1 - D(I_t^p)) \right]),$$

$$L_{VGG} = \sum_{n=1}^{N_{vgg}} \|VGG_n(I_t) - VGG_n(I_t^p)\|_1,$$

where $D$ denotes a discriminator.

Besides, we also re-train the SyncNet [Chung and Zisserman 2016b] discriminator for constraining temporal consistency and audio-lip synchronization as performed in Wav2Lip [Prajwal et al. 2020]. It takes 5 consecutive predicted frames $I_{t+5}^p$ as input. The loss is carried out between $I_{t+5}^p$ and their corresponding audio representations $a_{t+5}$ and $a_{t+5}^s$.

$$L_{sync} = \text{SyncNet}(I_{t+5}^p, a_{t+5}).$$

The overall learning objective can be summarized as:

$$L_{\text{all}} = L_{GAN} + \lambda_{vgg} L_{VGG} + \lambda_{sync} L_{sync}.$$
Figure 4: Qualitative Results. The top row shows the corresponding videos of the input audio. PC-AVS [Zhou et al. 2021] produces accurate lips, it is constrained to the cropped area. MakeitTalk [Zhou et al. 2020] fails to generate accurate mouth shape and lacks head dynamics. Wav2Lip [Prajwal et al. 2020] syncs well with audios, however, the mouth areas are blurry with their method. Our AV-CAT generates realistic results with synced mouth movements. The figures are selected from VoxCeleb ©Visual Geometry Group (CC BY) and LRW ©BBC.

Table 1: Quantitative results on LRW and VoxCeleb. For LMD the lower the better, and the higher the better for other metrics.

| Method               | LRW [Chung and Zisserman 2016a] | VoxCeleb [Nagrani et al. 2017] |
|----------------------|---------------------------------|---------------------------------|
|                      | SSIM ↑  | PSNR ↑  | LMD ↓  | Sync\textsubscript{conf} ↑  | SSIM ↑  | PSNR ↑  | LMD ↓  | Sync\textsubscript{conf} ↑  |
| Wav2Lip [Prajwal et al. 2020] | 0.937   | 34.25   | 2.54   | 6.7   | 0.885   | 32.34   | 8.44   | 5.2   |
| MakeitTalk [Zhou et al. 2020]   | 0.690   | 31.09   | 5.03   | 3.1   | 0.814   | 29.51   | 29.13  | 2.3   |
| PC-AVS [Zhou et al. 2021]     | 0.895   | 33.87   | 3.04   | 6.2   | 0.865   | 32.67   | 8.91   | 5.4   |
| Ground Truth             | 1.000   | 100.00  | 0.00   | 6.5   | 1.000   | 100.00  | 0.00   | 5.4   |
| AV-CAT (Ours)            | 0.938   | 36.34   | 2.83   | 6.0   | 0.889   | 33.41   | 8.64   | 5.1   |

AV-CAT also achieves comparable performance with the state-of-the-art lip-sync performance on both datasets. Our LMD score is slightly worse than Wav2Lip, possibly caused by the observation that we tend to generate more exaggerated lip movements. This stands the same for the SyncNet scores. It is possible that our self-trained SyncNet discriminator does not perform as well as the one used in Wav2Lip.

Particularly, we argue there is generally a trade-off between perceptual quality and synchronization and our model balances it better than traditional CNN-based models. We also re-train a Wav2Lip model with the same setting and dataset as ours (denoted as Wav2Lip-L) and find it achieves worse performance than the original model. Both the limited receptive fields and the skip-connections at higher resolutions restrict the network’s ability.

4.2 Qualitative Evaluation

We also conduct a qualitative comparison, including the results in Fig. 4 and a subjective evaluation. For video samples, please refer to our supplementary materials\(^1\). The first line of Fig 4 shows the frames of the original videos that correspond to input audio. As shown in Fig 4, MakeitTalk fails to predict accurate mouth shapes due to the prediction error from the 3D landmarks. Moreover, its generated head movements are relatively subtle compared with original frames. On the other hand, PC-AVS synthesizes basically correct mouth shapes, but their results are limited in the cropped facial areas, making them non-realistic. Wav2Lip synthesizes satisfactory lip-sync performance. However, the inpainted mouth region is always blurry, given that they can only handle images of low

\(^{1}\)Demo videos are available at https://hangz-nju-cuhk.github.io/projects/AV-CAT.
We conduct ablation studies further to demonstrate the contributions of different components in our AV-CAT. Specifically, we construct three variants by removing the specified components: 1) First, to evaluate the effectiveness of the transformer architecture, we remove the CTH backbone, and directly feed the masked frames to the larger version of Wav2Lip in our model (Wav2Lip-L + RN). 2) Then in order to evaluate the contribution of the Refinement Network, we remove the Refinement Network, and evaluate the intermediate output of the transformer backbone (denoted as “w/o RN”). 3) Finally, to evaluate the contribution that SyncNet makes to the lip-synchronization quality, we train another model without using the $L_{\text{sync}}$ (denoted as “w/o Sync”).

The numerical results are shown in Table 3 and the visual results are shown in Fig. 5, respectively. Compared to our Full Model, “RN only” achieves the worst results on LMD and SyncNet scores due to the redundant mouth shape information from the reference frame (the results are close to the target template as shown in Fig. 5). “W/o RN” obtains degraded performance in all metrics, which represents the inconsistent texture of the mouth region (see second row of Fig. 5). The effectiveness of our CTH backbone network can further be shown as the results combining Wav2Lip-L with RN perform worse than our full model. “w/o Sync” fails to generate synchronized lip movements. Both the numerical results and visual results demonstrate the effectiveness of our components in AV-CAT.

5 CONCLUSION

In this work, we present the Audio-Visual Context-Aware Transformer (AV-CAT), which produces high-quality lip-synced talking faces. We highlight several important properties. 1) We formulate the problem in a masked lip-sync prediction manner, and propose to learn cross-frame and cross-modal context information with Transformers. 2) Our design shows that audios can modulate UNet structures for quality improvement. 3) Our results are validated to be more realistic than the previous state of the arts.

Limitations and Future Work. 1) The model is not sensitive to certain consonants. Possibly, the data are not evenly distributed. The state of the reference image could also slightly affect the results. 2) Moreover, our model cannot handle delicate details such as mimicking the speaking style of a specific person or casting shadows from the jaw. 3) More advanced audio representations [Baevski et al. 2020; Chen et al. 2022b] can be involved in future studies.

Ethical Considerations. Our method might be leveraged for malicious uses such as creating deepfakes. We will restrict the license of our model to research use only and share it with the deepfake

| Method                        | SSM ^ | PSNR ↑ | LMD ↓ | Sync^conf ↑ |
|-------------------------------|-------|--------|-------|-------------|
| RN only                       | 0.812 | 30.61  | 10.68 | 3.9         |
| w/o RN                        | 0.748 | 29.31  | 10.16 | 4.9         |
| Wav2Lip-L + RN                | 0.874 | 32.29  | 8.72  | 4.8         |
| w/o Sync                      | 0.868 | 32.93  | 9.34  | 4.7         |
| Full Model                    | 0.889 | 33.41  | 8.64  | 5.1         |

Table 3: Ablation study with quantitative comparisons on VoxCeleb. The results are shown when we vary the accessibility of the Transformer network, refinement network, and the optimization objective.
Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. 2021. Tfill: Image completion via a transformer-based architecture. arXiv preprint arXiv:2104.00845 (2021).
Hang Zhou, Yu Liu, Ziwei Liu, Peng Luo, and Xiaogang Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. 9299–9306.
Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4176–4186.
Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. 2020. Make iTalk: speaker-aware talking-head animation. ACM Transactions on Graphics (TOG) 39, 6 (2020), 1–15.