ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446049

DIFFDUB: PERSON-GENERIC VISUAL DUBBING USING INPAINTING RENDERER WITH DIFFUSION AUTO-ENCODER

Tao Liu[1], *Chenpeng Du*[1], *Shuai Fan*[2], *Feilong Chen*[2], [†]*Kai Yu*[1]

[1]MoE Key Lab of Artificial Intelligence, AI Institute
[1]X-LANCE Lab, Shanghai Jiao Tong University
[2]AISpeech Ltd, Suzhou China

ABSTRACT

Generating high-quality and person-generic visual dubbing remains a challenge. Recent innovation has seen the advent of a two-stage paradigm, decoupling the rendering and lip synchronization process facilitated by intermediate representation as a conduit. Still, previous methodologies rely on rough landmarks or are confined to a single speaker, thus limiting their performance. In this paper, we propose ***DiffDub***: **Diff**usion-based **dub**bing. We first craft the Diffusion auto-encoder by an inpainting renderer incorporating a mask to delineate editable zones and unaltered regions. This allows for seamless filling of the lower-face region while preserving the remaining parts. Throughout our experiments, we encountered several challenges. Primarily, the semantic encoder lacks robustness, constricting its ability to capture high-level features. Besides, the modeling ignored facial positioning, causing mouth or nose jitters across frames. To tackle these issues, we employ versatile strategies, including data augmentation and supplementary eye guidance. Moreover, we encapsulated a conformer-based reference encoder and motion generator fortified by a cross-attention mechanism. This enables our model to learn person-specific textures with varying references and reduces reliance on paired audio-visual data. Our rigorous experiments comprehensively highlight that our ground-breaking approach outpaces existing methods with considerable margins and delivers seamless, intelligible videos in person-generic and multilingual scenarios.

Index Terms— Talking Face, Diffusion, Face Animation, Dubbing

1. INTRODUCTION

Visual dubbing [1, 2, 3], an area intrinsically linked to talking head synthesis, requires the meticulous alignment of lower-face movements in a source video with corresponding driving audio, while preserving the original identity, head pose, and background depicted in the source video. The utility of this task becomes evident when there is a need to modify or substitute the audio content with alternative audio, typically for translation purposes [4], as indicated in Figure 1.

The task navigates through multiple challenges[5]: maintaining high **visual quality**, ensuring **temporal consistency**, and perfecting **lip synchronization**. Visual quality necessitates the seamless alteration of the lower facial area and a harmonious integration of the generated region with the unaltered part of the image. Temporal consistency preserves a fluid and natural motion between successive video frames, preempting jitters, or abrupt transitions. Lip synchronization mandates effective alignment between audio and visual cues. A harmonious balance between temporal consistency and lip synchronization contributes to the intelligibility of the video[6].

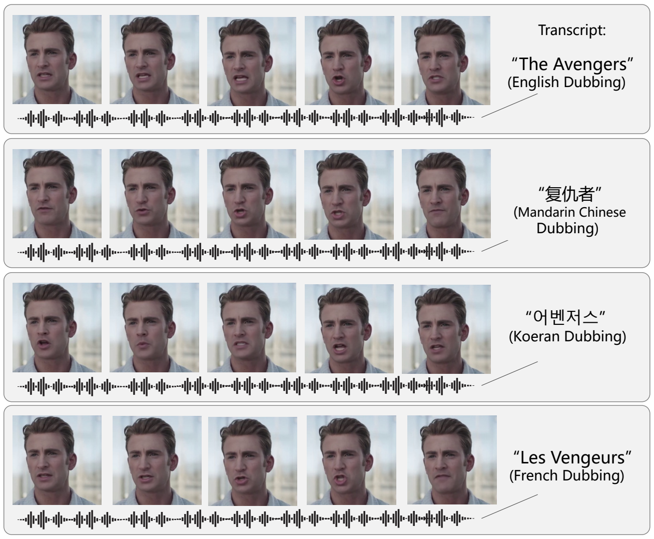

Fig. 1. **Dubbed videos with audio in various languages**. Our method can produce seamless and intelligible videos.

Regarding visual quality, a large proportion of methods [1, 7, 3] deploy Generative Adversarial Networks (GANs) as the rendering network. Despite this, these approaches grapple with challenges such as training instability and mode collapse [8]. The Diffusion Denoising Probabilistic Model (DDPM) [9] stands out as a probabilistic generative model that has notably displayed promising visual quality within the realm of talking head synthesis [10, 11, 12, 13]. The foundation of the diffusion model rests on the iterative addition (diffusion process) and removal (denoising process) of noise. Despite DDPM's considerable achievements in talking face generation [10, 11, 12], progress has been limited within the sphere of visual dubbing, where only the lower facial area requires modification while the rest of the image remains unaltered. This stipulation concurs with image in-painting [14, 15], where alterations are confined to the repaired region. The modified segment must blend flawlessly with the original image to emanate a natural and coherent image. However, their primary focus on static images and an absence of consideration for temporal consistency.

Concerning temporal consistency and lip synchronization, current methods [1, 16, 3] have yet to fully leverage the perks of sequence modeling or transduction methods, such as Transformer

[†]Kai Yu is the corresponding author.

979-8-3503-4485-1/24/$31.00 ©2024 IEEE

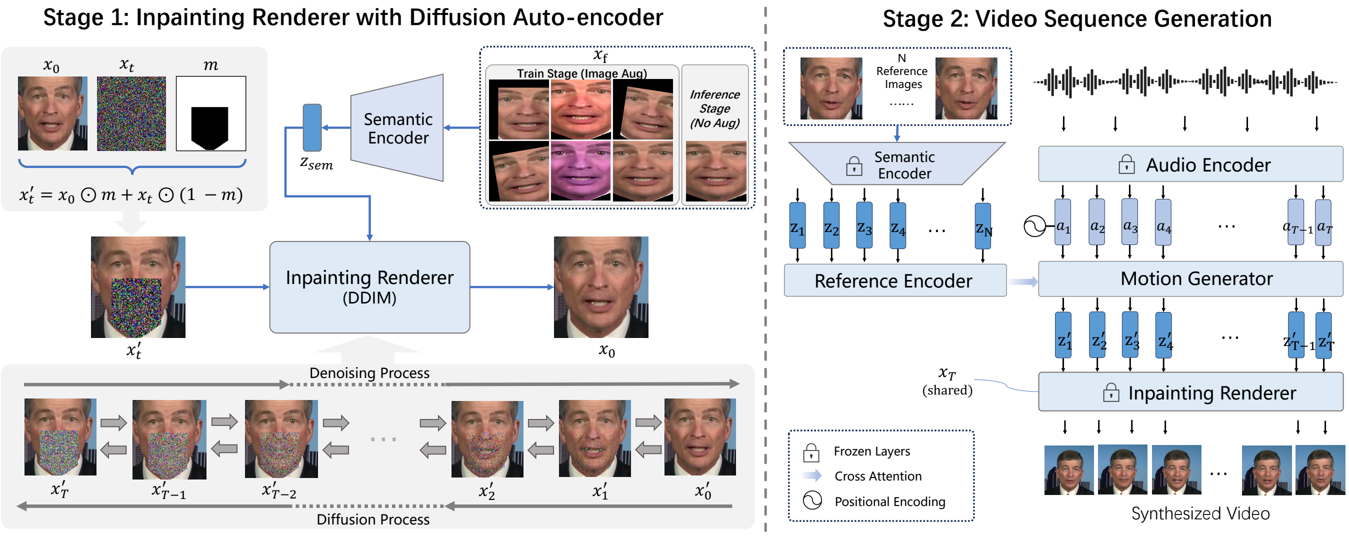

Fig. 2. **Architecture of DiffDub**. Our DiffDub approach upholds a two-stage paradigm encompassing *inpainting rendering with Diffusion Auto-encoder* and *video sequence generation*. In the first stage, we usher in a Diffusion Auto-encoder with masked conditions to generate semantic latent codes z through the semantic encoder. Subsequently, during the video generation phase, the semantic latent code z, in tandem with the audio latent code a derived from an extant model, is employed to generate the final videos.

[17] or Conformer [18]. These approaches typically operate on short audio clips, e.g., 200 ms in Wav2Lip [1]. This duration, however, might fall short for some phonemes [19], let alone their combinations. Relying on such brief audio clips introduces several challenges, including needing copious, aligned audio-visual data to achieve satisfactory results. Moreover, this approach regularly results in lip-synced but unintelligible videos [20] that often exhibit brisk lip movements or exaggerated mouth articulations [1, 16]. To surmount these challenges and harness the potential of sequence modeling, recent strategies [21, 22, 23, 13] have devised two-stage networks composed of an audio-to-representation generator and a representation-to-video rendering network. The generator, often reliant on a transformer model, maps the audio sequence onto a representative sequence while the rendering network fabricates the final video based on this representation. These methods, by decoupling the rendering process, successfully capture long-range dependencies. DAE-talker [13] advances this approach by exploiting semantic latent variables from Diffusion Autoencoders (Diff-AE) [24] as opposed to employing explicit structural representations like blendshapes [21, 22] or landmark coefficients [23]. Diff-AE can be viewed as a data compression or a representation learner by encoding an input image into a semantic latent variable and reconstructing the image from this latent space. Building upon this capability and capitalizing on Diff-AE's demonstrated competence in learning meaningful representations, DAE-talker achieves impressive outcomes. However, its applicability remains confined to a single speaker, like Obama, which dampens its generalization capacity.

In a bid to holistically address the aforementioned challenges, we introduce ***DiffDub***: a person-generic visual dubbing methodology underpinned by DDPM. This paper encapsulates several contributions itemized below.

- We have designed a potent inpainting renderer tasked with the generation of the modifiable region under the supervision of immutable components and semantic conditioning. This method can generate seamlessly blended lower facial regions that far surpass the capabilities of preceding methods.
- We have proposed diverse strategies aimed at bolstering the robustness of the semantic encoder. These tactics enable the encoder to apprehend subtle movements and supply meticulous positional data for the mouth and nose regions.
- By leveraging the Conformer model with cross-attention, we have successively adapted our methodology to cater to an assortment of references and audio sequences. This adaptation allows our model to assimilate person-specific textures while reducing reliance on paired audio-visual data.
- We have performed thorough quantitative and qualitative evaluations in both few-shot and one-shot settings on URL[1].

2. DIFFDUB FRAMEWORK

2.1. Inpainting Renderer with Diffusion Auto-encoder

This module is responsible for inpainting rendering and latent semantic representation learning. Before diving into its details, it is crucial to briefly recapitulate the noising and denoising process inherent to vanilla diffusion [9]. Given a data distribution $x_0 \sim q\left(x_0\right)$, the noising process produces a series of latent $x_1, \ldots, x_T$ by adding Gaussian noise with variance at each time t. For the denoising process, we use $\mathcal{N}(\mu_\theta(x_t, t), \sigma_t)$ to model $q\left(x_{t\text{-}1}|x_t\right)$. Thus, we can train a deep neural network p_θ to predict the mean of Gaussian noise. However, the vanilla diffusion is not controllable, and we define two guidance information for the dubbing task by segregating the full image into two segments using a lower-face mask. The first segment is the reference x_0, providing the information to be retained, including identity, pose, and background. The second segment is the facial motion x_f, supplying the information for the editable region. The mask encompasses the lower facial area and can be procured through a landmark predictor [25].

[1] https://liutaocode.github.io/DiffDub/

Our proposed approach is comprised of a semantic encoder $z_{\text{sem}} = \text{Enc}_\phi(x_{\text{f}})$ and an inpainting renderer $p(x'_{t-1} \mid x'_t, z_{\text{sem}})$. Compared to Diff-AE [24], the input of the semantic encoder is solely the facial area x_f rather than a full image x_0. We opt for this approach to ensure that the semantic encoder solely imparts facial motion information. We also incorporate image augmentation throughout the training phase to ascertain the learning of high-level features by the semantic encoder instead of trivial patterns. Moreover, the noised image x_t in Diff-AE incorporates noise added across the entire image. In contrast, in our approach, the noise is exclusively added to the masked region, thereby maintaining the unaltered section. Distinguished from x_t, the noised image here is denoted as x'_t here, as articulated in Equation 1, where the symbol $\odot$ denotes element-wise product and m designates a binary mask matrix: zero for the edited region and one for the unchanged part. This equation demonstrates that $m \odot x_0$ supplies the known pixel in the given image, and $(1-m) \odot x_t$ is a masked version of x_t for each iteration t, providing the unknown pixel. The architecture is depicted in the first stage of Figure 2.

$$x'_t = m \odot x_0 + (1-m) \odot x_t \tag{1}$$

Furthermore, it is crucial to observe that the facial image x_f encompasses an additional eye area, extending beyond the masked region. The inclusion of this extra eye is anchored on the reasoning that it assists in localizing the position of the nose and mouth, thereby enhancing the stability of the nose and mouth across frames.

In the DDPM's training stage, we employ the simplified loss objective delineated in [9] and incorporate a specific mask that ensures only the loss in the editable facial motion area is computed, as depicted in Equation 2, where ϵ represents the actual noise. For inference, in light of the extensive iteration steps associated with diffusion, we opt for the Denoising Diffusion Implicit Model (DDIM) [26]—an alternative non-Markovian noising process—as the solver to accelerate the sampling process.

$$L_{\text{simple}} = E_{t,x_0,\epsilon}\left[\left\|(1-m) \odot \left(\epsilon - \epsilon_\theta\left(x'_t, t, z_{\text{sem}}\right)\right)\right\|^2\right] \tag{2}$$

2.2. Video Sequence Generation

This phase aims to produce person-specific synthesized videos by processing N reference facial images and singular driving audio. It enlists the help of a reference encoder and a motion generator based on the Conformer [18]. Unlike methods [1, 3] that restricted inputs to fixed frames, our methodology permits input lengths to vary. Incorporating the Conformer also enables us to capture global and local facial motion interactions, a considerable leap from previous methods [1, 16, 3] that only facilitated interactions of short durations.

Besides, in this stage, we rely on latent codes instead of predefined structural representations [21, 22, 23]. The latent codes fall into two categories: the semantic latent code z and the audio latent code a. The semantic encoder, which remains frozen, is tasked with extracting the semantic latent code from facial images. Concurrently, the audio model, using a self-supervised approach, extracts the latent code from the driving audio. The reference encoder gleans person-specific facial texture information from N visual latent codes $z_{\text{1:N}}$, and the motion generator performs a one-to-one mapping, transforming T audio latent codes $a_{\text{1:T}}$ into the corresponding T visual latent codes $z'_{\text{1:T}}$. Ultimately, the visual latent codes are fed into the inpainting renderer to synthesize the images.

To efficiently incorporate personalized textures, we introduce a cross-attention mechanism [17], a pivot from the direct concatenation of the reference images [1, 3, 10]. This mechanism employs the output of the reference decoder as the query, while the audio latent codes operate as the key and value in a multi-head attention operation, thus generating person-aware facial motion latent codes.

To enhance the robustness of the audio encoder, we retrieve the audio latent code through a weighted sum [27] of all layers of the self-supervised models. This approach deviates from the commonly used Mel-based feature representation, thus granting added language flexibility. Furthermore, mirroring the approach adopted in DAE-Talker [13], we use a shared x_T for DDIM as the starting point for all images, thereby ensuring that the DDIM generates deterministic and consistent results.

3. EXPERIMENTS

3.1. Experimental Setups

Dataset. We utilize the HDTF dataset [28] for our experiments. The HDTF dataset is comprised of high-resolution, real-world talking head videos collected from YouTube. The dataset has a total duration of around 16 hours, partitioning 245 clips for training and 68 clips for testing, aligning with the framework in [3]. Notably, the actual tally of videos utilized in our experimental setup marginally trails the official release, owing to the unavailability of 6 online videos. As part of the preprocessing endeavor, all videos in the HDTF dataset are reformatted to a fixed resolution of 256 × 256 pixels and standardized to a frame rate of 25 frames per second (FPS).

Model Details. Similar to Diff-AE [24], the diffusion model draws on U-Net [29] and the dimension of the semantic latent code $\mathbf{z}_{\text{sem}}$ is 512. We employ several techniques to augment the data, including horizontal flipping, color jitter, Gaussian blur, shifting, scaling, and rotation. The time step T for DDIM is set to 20. The off-the-shelf audio encoder is a pre-trained Hubert-large model [30]. The reference frame number N is set to 75 (3 seconds) for few-shot experiments. The reference encoder and motion generator use a 2-layer and 8-layer conformer, both employing two attention heads and relative positional encoding [31]. The model undergoes training for an initial four epochs in the first stage, followed by an additional 100 epochs in the second stage.

Evaluation Metric. As for **Visual Quality (VQ)**, we employ Peak Signal-to-Noise Ratio (PSNR), Structured similarity (SSIM) [32] and Learned Perceptual Image Patch Similarity (LPIPS) [33] as metrics to quantify the similarity between the generated and ground truth images. Since masked areas vary across methods, we resize all images to an identical resolution for a fair comparison and only evaluate the lower face area. As for **Synchronization (SYNC)**, We utilize lip-sync-error distance (LSE-D), lip-sync-error confidence (LSE-C) [1, 34], and landmarks distance (LMD) [2]. The LSE metrics measure the degree of lip-sync alignment between the generated lips and the audio, while the LMD metric assesses the reconstructed shape of the lower face region relative to the ground truth.

Baseline systems. We compare our method with several state-of-the-art person-generic methods [1, 16, 23, 13]. **Wav2Lip** [1] employs an auto-encoder trained through adversarial methods. **PC-AVS** [16] introduces a pose-controllable audio-visual talking face generation method. **IP-LAP** [23] and **DAE-Talker** [13] represent two-stage methodologies. IP-LAP employs landmark, whereas DAE-Talker utilizes latent code. For equal comparison, adjustments are made to accommodate the nature of each method. Since PC-AVS is influenced by pose, we utilize ground-truth pose information for its evaluation. Furthermore, due to the speaker-specific design of DAE-Talker, we engage in a retraining process on HDTF.

3.2. Method Comparison

We compare methods across three typical scenarios: *reconstruction*, *dubbing*, and *one-shot*. For the reconstruction and dubbing, we synthesized a talking head with audio and they differ in that the reconstruction employs ground-truth audio, while the dubbing utilizes audio from an alternate dialogue or other speakers. For one-shot, a single portrait is exploited to synthesize the outcome.

Table 1. Quantitative Results on HDTF Reconstruction

Method	VQ			SYNC		
	PSNR↑	SSIM↑	LPIPS↓	LSE-D↓	LSE-C↑	LMD↓
Wav2Lip [1]	26.12	0.83	0.066	7.48	7.84	1.10
PC-AVS [16]	22.15	0.68	0.058	**6.66**	**8.69**	2.03
IP-LAP [23]	26.05	0.84	0.058	8.78	5.59	**0.79**
DAE-Talker [13]	17.81	0.47	0.129	8.90	6.15	2.78
GT	*N/A*	*1*	*0*	*6.73*	*9*	*0*
Ours	**28.18**	**0.87**	**0.035**	8.16	7.10	0.95

Quantitative Comparison. The results are recorded in Table 1. Our methodology exhibits commendable results in terms of the visual quality metric, corroborating the efficacy of the rendering module in the preliminary stage. However, when examined from a synchronization perspective, Wav2Lip and PC-AVS outpace our method across lip-sync-error (LSE) parameters, and IP-LAP surpasses our performance regarding the landmarks distance (LMD). This discrepancy stems from these methods' direct optimization focus on either LSE or LMD, a focus our method does not explicitly pursue. Nonetheless, our method delivers competitive results despite the absence of an explicit optimization for these distinct metrics.

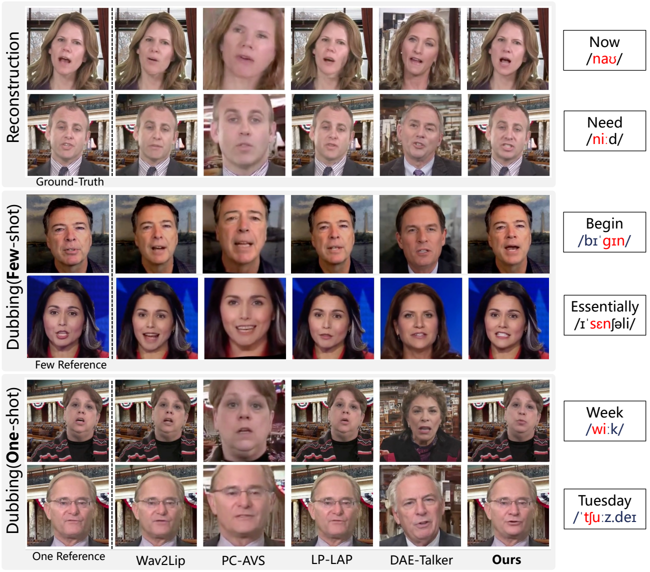

Fig. 3. Qualitative results on Reconstruction & Dubbing. The corresponding pronounced syllables are highlighted in red.

Qualitative Comparison. The results are presented in Figure 3. It is noteworthy that the generated lower face and teeth regions manifest improved clarity, mirroring the positive visual quality metric unearthed in the quantitative comparison. Additionally, our generated outputs exhibit enhanced consistency with the longer pronunciation units such as syllables. This outcome reaffirms our method's capacity to produce more comprehensible results, synchronizing with the conclusion that previous strategies frequently culminate in lip-synced but unintelligible videos [20].

Table 2. Ablation Study on HDTF Reconstruction

Method	VQ			SYNC		
	PSNR↑	SSIM↑	LPIPS↓	LSE-D↓	LSE-C↑	LMD↓
Ours w/o eye	28.84	0.88	0.033	8.85	5.76	1.53
Ours w/o aug.	28.45	0.88	0.034	9.60	4.75	0.94
Ours w/o w.s.	27.48	0.86	0.040	8.40	6.50	1.90
Ours w $\frac{1}{10}$ data	28.12	0.87	0.033	8.34	7.01	0.97
Ours	28.18	0.87	0.035	8.16	7.10	0.95

Ablation Study. We conducted several ablation experiments to evaluate the effectiveness of the proposed model. In the first ablation, we remove the eye area from the input for the semantic encoder. This exclusion led to a particular degradation in LMD, highlighting the importance of eyes in guiding the localization of the nose and mouth. In the second ablation, we deactivated the augmentation in the first stage. The results revealed a reduction in LSE, suggesting that augmentation is crucial in learning high-level semantic information from facial images. In the third ablation, we adopted the Mel-based feature instead of weighted sum (w.s.), and we observed a decline across all metrics. In the fourth ablation, we curtailed the amount of paired audio-visual data utilized in the second stage to a tenth of the original data, equivalent to approximately 1.5 hours. Remarkably, we observe no significant degradation, implying that our method exhibits enhanced robustness with constrained training data.

Table 3. User Study with Mean Option Score (MOS)

Methods	MOS-MF	MOS-RI	MOS-N	MOS-CL
Wav2Lip [1]	3.46	4.07	3.26	3.42
PC-AVS [16]	3.74	3.46	3.08	3.50
IP-LAP [23]	3.53	3.70	3.95	2.42
DAE-Talker [13]	3.19	3.33	2.61	2.23
Ours	**4.62**	**4.59**	**4.60**	**4.17**

Subjective Evaluation. We conducted a user study with 12 participants rating our method across Mouth Fidelity (MF), Reading Intelligibility (RI), Naturalness (N), and Cross-Lingual performance (CL). Five videos were selected randomly from HDTF dataset. For CL, the original videos were in English, and we translated them into Mandarin Chinese, Korean, and French. The results, outlined in Table 3, demonstrate the language generality of our method.

4. CONCLUSION

This paper presents ***DiffDub***, an innovative approach for person-generic dubbing without necessitating fine-tuning. We have devised a potent inpainting renderer with meticulously tailored strategies that can seamlessly integrate the editable region. We employ a conformer-based method with a cross-attention mechanism to learn person-specific textures and details, thus accommodating varied references and reducing dependence on the training corpus. Quantitative and qualitative experiments underline that our proposed method yields seamless and intelligible outcomes. Subjective evaluations further affirm that DiffDub outstrips the baseline by a margin.

5. ACKNOWLEDGEMENTS

China NSFC Project (No. 92370206), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102) and the State Key Laboratory of Media Convergence Production Technology and Systems Project (No. SKLMCPTS2020003).

6. REFERENCES

[1] K. Prajwal *et al.*, “A lip sync expert is all you need for speech to lip generation in the wild,” in *Proceedings of the 28th ACM international conference on multimedia (ACM MM)*, 2020.

[2] T. Xie *et al.*, “Towards realistic visual dubbing with heterogeneous sources,” in *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*, 2021.

[3] Z. Zhang *et al.*, “Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video,” *Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[4] P. KR *et al.*, “Towards automatic face-to-face translation,” in *Proceedings of the 27th ACM international conference on multimedia (ACM MM)*, 2019.

[5] C. Sheng *et al.*, “Deep learning for visual speech analysis: A survey,” *arXiv preprint arXiv:2205.10839*, 2022.

[6] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” *Nature*, 1976.

[7] Y. Zhou *et al.*, “Makelttalk: speaker-aware talking-head animation,” *ACM Transactions On Graphics (TOG)*, 2020.

[8] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” *Advances in neural information processing systems*, 2021.

[9] J. Ho *et al.*, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, 2020.

[10] S. Shen *et al.*, “Difftalk: Crafting diffusion models for generalized audio-driven portraits animation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[11] M. Stypułkowski *et al.*, “Diffused Heads: Diffusion Models Beat GANs on Talking-Face Generation,” in *https://arxiv.org/abs/2301.03396*, 2023.

[12] D. Bigioi *et al.*, “Speech driven video editing via an audio-conditioned diffusion model,” *arXiv preprint arXiv:2301.04474*, 2023.

[13] C. Du *et al.*, “Dae-talker: High fidelity speech-driven talking face generation with diffusion autoencoder,” *Proceedings of the 31th ACM International Conference on Multimedia (ACM MM)*, 2023.

[14] A. Lugmayr *et al.*, “Repaint: Inpainting using denoising diffusion probabilistic models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[15] O. Avrahami *et al.*, “Blended diffusion for text-driven editing of natural images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[16] H. Zhou *et al.*, “Pose-controllable talking face generation by implicitly modularized audio-visual representation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021.

[17] A. Vaswani *et al.*, “Attention is all you need,” *Advances in neural information processing systems*, 2017.

[18] A. Gulati *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” *Conference of the International Speech Communication Association (InterSpeech)*, 2020.

[19] Igras *et al.*, “Length of phonemes in a context of their positions in polish sentences,” in *2013 International Conference on Signal Processing and Multimedia Applications (SIGMAP)*. IEEE, 2013.

[20] J. Wang *et al.*, “Seeing what you said: Talking face generation guided by a lip reading expert,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[21] Y. Fan *et al.*, “Faceformer: Speech-driven 3d facial animation with transformers,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[22] Q. Chen *et al.*, “Improving few-shot learning for talking face system with tts data augmentation,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[23] W. Zhong *et al.*, “Identity-preserving talking face generation with landmark and appearance priors,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[24] K. Preechakul *et al.*, “Diffusion autoencoders: Toward a meaningful and decodable representation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[25] A. Bulat and G. Tzimiropoulos, “How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks),” in *International Conference on Computer Vision (ICCV)*, 2017.

[26] J. Song *et al.*, “Denoising diffusion implicit models,” in *International Conference on Learning Representations (ILCR)*, 2020.

[27] S.-w. Yang *et al.*, “Superb: Speech processing universal performance benchmark,” *Conference of the International Speech Communication Association (InterSpeech)*, 2021.

[28] Z. Zhang *et al.*, “Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[29] Ronneberger *et al.*, “Convolutional networks for biomedical image segmentation.” Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2022.

[30] W.-N. Hsu *et al.*, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 2021.

[31] Z. Dai *et al.*, “Transformer-xl: Attentive language models beyond a fixed-length context,” *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[32] Z. Wang *et al.*, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, 2004.

[33] R. Zhang *et al.*, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[34] J. S. Chung *et al.*, “Out of time: automated lip sync in the wild,” in *Asian Conference on Computer Vision (ACCV) Workshops*, 2017.