Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/102/20241156

Review of Talking Head Synthesis for Driving Mechanisms and Portrait Rendering

Ziwei Liu

School of Environmental Science and Safety Engineering, Tianjin University of Technology, 391 BinShui Xidao Road, Xiqing District, 300384, China

lzw9612@gmail.com

Abstract. Talking head synthesis has emerged as a vital area of research, enabling the generation of realistic and expressive digital avatars. This paper explores the primary mechanisms driving talking head synthesis, categorized into video-driven and audio-driven methods. Video-driven techniques manipulate facial movements using key points, 3D meshes, and latent spaces, while audio-driven approaches focus on synchronizing lip movements and facial expressions with audio inputs. Recent advances in each method, highlighting key innovations and the challenges faced, such as occlusion, identity preservation, and lip synchronization are reviewed. The technology's applications span smart customer service, online education, telemedicine, and video creation. Future research directions focus on overcoming challenges like handling large-angle poses, ensuring temporal consistency, and improving multilingual performance.

Keywords: Talking Head Synthesis, Driving Mechanisms, Portrait Rendering.

1. Introduction

The rapid evolution of deep learning technology has positioned talking head synthesis at the forefront of digital communication innovation [1]. This technology holds immense potential across various domains, including enhancing accessibility for individuals with communicative impairments [2], revolutionizing educational practices through AI-driven interactive tutoring [3], and providing therapeutic support and social interaction in healthcare settings [4].

Talking head synthesis comprises two primary stages: driving mechanisms and portrait rendering. The driving mechanism stage is crucial for imbuing static images with lifelike qualities. These mechanisms utilize either video or audio inputs to orchestrate facial movements and expressions. Video-driven approaches replicate facial movements and expressions from reference footage, while audio-driven methods synchronize lip movements and facial expressions with corresponding audio tracks. The portrait rendering stage focuses on generating photorealistic facial images using advanced generative models. These include Generative Adversarial Networks (GANs) [5], Diffusion Models [6], Neural Radiance Fields (NeRF) [7], and 3D Gaussian Splatting (3DGS) [8]. These models excel in producing high-resolution static images, which serve as the foundation for subsequent animation processes.

This paper explores the primary mechanisms driving talking head synthesis, categorized into video-driven and audio-driven methods., highlighting the unique contributions and strengths of various approaches. Additionally, the paper discusses the advantages, limitations, existing challenges, and potential solutions for future research. Future research directions focusing on overcoming challenges

DOI: 10.54254/2755-2721/102/20241156

like handling large-angle poses, ensuring temporal consistency, and improving multilingual performance are elucidated.

2. Methods of driven mechanisms and portrait rendering

Driven mechanisms are essential in pushing the boundaries of talking head synthesis, facilitating the generation of realistic and expressive talking head videos. By utilizing different input modalities, these mechanisms effectively control the facial movements and expressions of synthesized characters, ensuring that the output appears natural and coherent. The approaches in this field are generally categorized into two main types: video-driven methods and audio-driven methods.

2.1. Video-driven Methods

Video-driven talking head synthesis involves manipulating defined features such as keypoints, meshes, and latent spaces, making these methods highly interpretable.

Keypoints-based warping methods generate motion flow by learning the correspondence between keypoints, thereby warping the features of the source image. Siarohin et al. developed the First Order Motion Model (FOMM), a method that animates objects in images using a set of learned keypoints along with their local affine transformations, generating a dense motion field and occlusion masks to deform the source image at the encoder's feature layer and restore it through the decoder [9]. Zhao et al. improved upon FOMM by using multi-resolution occlusion masks to achieve more effective feature fusion and incorporating an affine transformation for background prediction [10]. Hong et al. further enhanced FOMM by utilizing depth information to improve the precision of warps and reduce artifacts [11]. Hong & Xu propose a novel implicit identity representation conditioned memory compensation network for talking head video generation, in which an implicit identity representation conditioned memory module and a facial memory compensation module are designed to respectively perform the meta-memory query and feature compensation [12]. Wang et al. introduced dynamic 3D keypoints for unsupervised learning from single-source images, enabling one-shot free-view avatar synthesis [13]. Drobyshev et al. utilize dynamic 3D keypoints and propose a novel contrastive loss to achieve higher degrees of disentanglement between the latent motion and appearance representations, adding a gaze loss that increases the realism and accuracy of eye animation [14]. Drobyshev et al. improve MegaPortraits model to transfer intense expressions correctly through careful latent facial expression space development and employ the expression-enhanced loss and a minimal amount of domain-specific data [15]. Guo et al. introduce a series of significant enhancements including high-quality data curation, a mixed image and video training strategy, an upgraded network architecture, scalable motion transformation, landmark-guided implicit keypoints optimization, and cascaded loss terms [16].

In contrast to keypoints-based methods, mesh-based rendering methods depend on 3D head reconstruction models. 3DMM models like DeepFaceReconstruction[17], DECA [18], Emoca [19] are employed. Yao et al. proposed a method that uses 3D mesh reconstruction to guide optical flow learning for facial reenactment[20]. Guan et al. proposed a new Style-based generator using 3D mesh and re-configure the information insertion mechanisms within the noise and style space [21]. Hong et al. presented HeadNeRF, a parametric head model using NeRFs to render high-fidelity human head images, incorporating 2D neural rendering and improving detail accuracy [22]. Xu et al. utilized controllable 3D Gaussian models for high-fidelity head avatar modeling. By incorporating high-frequency dynamic details through a fully learned MLP-based deformation field, it effectively simulates a wide range of extreme expressions [23].

Latent space-based methods represent images as embedding codes. Wang et al. proposed the LIA model, achieving animation by learning orthogonal motion directions within the latent space and linearly combining them, avoiding complex processing based on structural representation [24]. Liu et al. enhances motion representation by employing metric learning, mutual information disentanglement, and Hierarchical Aggregation Layer [25]. Li et al. introduced HiDe-NeRF, which decomposes 3D dynamic scenes into canonical appearance and implicit deformation fields, maintaining facial shape and details through multi-scale volumetric features [26].

DOI: 10.54254/2755-2721/102/20241156

Video-driven methods have advanced the generation of realistic talking head animations by preserving the identity of a still image while replicating the motion from a driving video. However, they face challenges such as occlusion, maintaining identity integrity, and handling large pose variations.

2.2. Audio-driven Methods

Audio-driven methods leverage audio signals to synchronize lip movements and facial expressions with spoken content, thereby enhancing the realism and naturalness of talking head animations. This multimodal task faces several challenges due to the inherent differences between audio and visual modalities, such as lip synchronization. To address these challenges, existing approaches are classified into two main categories: explicit-based and implicit-based.

Implicit-based methods represent audio as latent space-based features. Prajwal et al. propose to use a pre-trained expert lip-sync discriminator to penalize the generator for inaccurate generation [27]. Liu et al. utilized a combination of diffusion and variance adapters to predict motion latent [25]. Xu et al. proposed diffusion transformer to model the motion distribution and generate the motion latent codes in the test time given audio and other conditions [28].

Explicit-based represent audio as 3D facial mesh and 3DMM parameters. Fan et al. leveraged self-supervised pre-trained speech representations and a transformer-based autoregressive model with a biased attention mechanism to capture long-term audio context, aligning audio-motion modalities, and accurately forecasting animated 3D facial mesh sequences for enhanced lip movement precision [29]. Huang et al. used a transformer-based encoder to map audio signals to 3DMM parameters, guiding the generation of high-quality avatars by predicting long-term audio context [30]. Zhang et al. introduce SadTalker, which synthesizes 3D motion coefficients from audio inputs and integrates them with advanced 3D perception rendering technology to accurately map audio-motion correlations, capturing detailed facial expressions and head movements [31]. Cho et al. proposed GaussianTalker, a real-time pose-controllable talking head model that integrates 3D Gaussian attributes with audio features into a shared implicit feature space, leveraging 3DGS for rapid rendering, resulting in enhanced facial fidelity, lip-sync accuracy, and superior rendering speed compared to existing models [32].

Audio-driven methods heavily rely on high-quality training data, which can cause overfitting or poor generalization when the data is insufficient or biased. Moreover, their stochastic nature often leads to inconsistent animations, particularly over longer sequences.

3. Applications

Talking head synthesis technology has a wide range of applications across various sectors, revolutionizing how we interact with digital content. This section explores some of the most promising application areas where talking head synthesis is making significant impacts.

Smart Customer Service: Talking head synthesis revolutionizes customer service by deploying virtual assistants that provide consistent, empathetic interactions. These digital agents can handle customer queries with lifelike expressions and real-time responses, improving customer satisfaction and operational efficiency.

Online Education: Talking head generation enhances online education by creating engaging virtual tutors that deliver interactive lessons with realistic facial expressions and real-time responses. This personalized approach improves student engagement, making complex concepts more accessible and enjoyable, particularly in remote and self-paced learning environments.

Telemedicine: In telemedicine, talking head technology provides virtual medical professionals that interact empathetically with patients. These lifelike avatars can guide patients through medical consultations, providing a human touch in remote diagnostics and care, enhancing patient comfort and trust.

Video Creation: Talking head technology streamlines short video content creation, allowing creators to produce engaging, lifelike characters without needing live actors. This opens up new creative possibilities for influencers and brands, reducing production time and costs while maintaining high-quality visual appeal. Sections, subsections and subsubsections

DOI: 10.54254/2755-2721/102/20241156

The use of sections to divide the text of the paper is optional and left as a decision for the author. Where the author wishes to divide the paper into sections the formatting shown in table 2 should be used.

4. Challenges and Future Directions

4.1. Pretrained large models

Current frameworks for talking head synthesis often rely on pretrained large models, such as diffusion models, which are designed with highly specific capabilities tailored to particular tasks. While these models offer impressive performance in certain areas, their rigid architecture can hinder innovation and adaptability for more diverse or specialized use cases. Over-reliance on a single pretrained model restricts flexibility and may result in models that struggle to generalize to new tasks or novel domains. A promising avenue for future research is to explore training individual components or modules on large-scale, diverse datasets. This approach allows for more modular and adaptable architectures that can be combined or fine-tuned to suit specific tasks without being overly dependent on one core model. Furthermore, hybrid architectures that integrate pretrained models with task-specific layers or modules can offer the benefits of both pretrained expertise and targeted adaptability, potentially leading to more efficient and versatile systems for a wide range of applications.

4.2. Handling Large-Angle Poses

Large-angle poses present a significant challenge in talking head synthesis due to the limited availability of training data for extreme angles, often leading to suboptimal performance. These angles introduce substantial changes in facial geometry and can obscure parts of the face, compounding the issue. To mitigate this, expanding datasets to include a wider range of large-angle poses, as well as incorporating multi-view training strategies, can help models handle such poses more effectively.

4.3. Temporal consistency

One of the most important challenges in talking head synthesis is maintaining temporal consistency, particularly when producing video outputs that need to look smooth and natural. Inadequate handling of time-series data can result in visual discontinuities such as jitter, sudden jumps between frames, or unnatural changes in facial expressions and movements. These artifacts disrupt the fluidity of the synthesized video, reducing its overall realism. The key to addressing this issue lies in both the quality of the datasets used and the methods employed for processing sequential data. High-quality datasets with strong frame-to-frame continuity are essential to train models that can accurately capture the natural flow of movements over time. Additionally, developing improved algorithms that better account for temporal dependencies in time-series data is crucial. Future research should prioritize the integration of advanced temporal processing techniques, such as recurrent neural networks (RNNs) or temporal attention mechanisms, to enhance the model's ability to produce consistent, lifelike sequences across extended periods of time.

4.4. Multilingual Challenges

Most existing audio-driven talking head synthesis models are primarily designed and optimized for English, creating challenges when extending the technology to other languages. This limitation stems from a lack of annotated, high-quality datasets for a wide range of languages, especially those with different phonetic structures, lip movements, or cultural nuances. When models are trained exclusively on English data, their ability to accurately generate lip movements, expressions, and speech dynamics in other languages is significantly diminished, leading to less accurate and natural outputs. Addressing this issue requires the collection and curation of diverse multilingual datasets, which capture the unique linguistic and phonetic features of different languages. Furthermore, employing advanced techniques like self-supervised learning, where models can learn useful representations from unlabelled data, and transfer learning, where knowledge gained from English can be adapted to other languages, can help improve the performance and flexibility of talking head synthesis across various linguistic contexts.

DOI: 10.54254/2755-2721/102/20241156

These approaches can enhance the scalability and global applicability of such systems, making them more inclusive and adaptable to a wider audience.

5. Conclusion

This review provided an in-depth analysis of talking head synthesis, examining both video-driven and audio-driven mechanisms and their impact on generating realistic head animations. While significant progress has been made in improving facial movement accuracy and synchronization, challenges remain, particularly in managing large-angle poses, maintaining temporal consistency, and expanding multilingual support. Applications of this technology are already transforming fields such as customer service, education, and telemedicine. Future research should focus on creating more adaptable models, expanding datasets, and developing hybrid architectures to overcome current limitations and enhance the versatility and realism of talking head synthesis.

References

[1] Wang, T. C., Mallya, A., & Liu, M. Y. (2021). One-shot free-view neural talking-head synthesis for video conferencing. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10039-10049).

[2] Johnson, E., Hervás, R., Gutiérrez López de la Franca, C., Mondéjar, T., Ochoa, S. F., & Favela, J. (2018). Assessing empathy and managing emotions through interactions with an affective avatar. Health informatics journal, 24(2), 182-193.

[3] Bozkurt, A., Junhong, X., Lambert, S., Pazurek, A., Crompton, H., Koseoglu, S., ... & Romero-Hall, E. (2023). Speculative futures on ChatGPT and generative artificial intelligence (AI): A collective reflection from the educational landscape. Asian Journal of Distance Education, 18(1), 53-130.

[4] Leff, J., Williams, G., Huckvale, M., Arbuthnot, M., & Leff, A. P. (2014). Avatar therapy for persecutory auditory hallucinations: What is it and how does it work?. Psychosis, 6(2), 166-176.

[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139-144.

[6] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015, June). Deep unsupervised learning using nonequilibrium thermodynamics. In International conference on machine learning (pp. 2256-2265). PMLR.

[7] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1), 99-106.

[8] Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Trans. Graph., 42(4), 139-1.

[9] Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019). First order motion model for image animation. Advances in neural information processing systems, 32.

[10] Zhao, J., & Zhang, H. (2022). Thin-plate spline motion model for image animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3657-3666).

[11] Hong, F. T., Zhang, L., Shen, L., & Xu, D. (2022). Depth-aware generative adversarial network for talking head video generation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3397-3406).

[12] Hong, F. T., & Xu, D. (2023). Implicit identity representation conditioned memory compensation network for talking head video generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 23062-23072).

[13] Wang, T. C., Mallya, A., & Liu, M. Y. (2021). One-shot free-view neural talking-head synthesis for video conferencing. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10039-10049).

DOI: 10.54254/2755-2721/102/20241156

[14] Drobyshev, N., Chelishev, J., Khakhulin, T., Ivakhnenko, A., Lempitsky, V., & Zakharov, E. (2022, October). Megaportraits: One-shot megapixel neural head avatars. In Proceedings of the 30th ACM International Conference on Multimedia (pp. 2663-2671).

[15] Drobyshev, N., Casademunt, A. B., Vougioukas, K., Landgraf, Z., Petridis, S., & Pantic, M. (2024). EMOPortraits: Emotion-enhanced Multimodal One-shot Head Avatars. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8498-8507).

[16] Guo, J., Zhang, D., Liu, X., Zhong, Z., Zhang, Y., Wan, P., & Zhang, D. (2024). LivePortrait: Efficient Portrait Animation with Stitching and Retargeting Control. arXiv preprint arXiv:2407.03168.

[17] Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., & Tong, X. (2019). Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 0-0).

[18] Feng, Y., Feng, H., Black, M. J., & Bolkart, T. (2021). Learning an animatable detailed 3D face model from in-the-wild images. ACM Transactions on Graphics (ToG), 40(4), 1-13.

[19] Daněček, R., Black, M. J., & Bolkart, T. (2022). Emoca: Emotion driven monocular face capture and animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20311-20322).

[20] Yao, G., Yuan, Y., Shao, T., & Zhou, K. (2020, October). Mesh guided one-shot face reenactment using graph convolutional networks. In Proceedings of the 28th ACM international conference on multimedia (pp. 1773-1781).

[21] Guan, J., Xu, Z., Zhou, H., Wang, K., He, S., Zhang, Z., ... & Liu, Z. (2024). ReSyncer: Rewiring Style-based Generator for Unified Audio-Visually Synced Facial Performer. arXiv preprint arXiv:2408.03284.

[22] Hong, Y., Peng, B., Xiao, H., Liu, L., & Zhang, J. (2022). Headnerf: A real-time nerf-based parametric head model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20374-20384).

[23] Xu, Y., Chen, B., Li, Z., Zhang, H., Wang, L., Zheng, Z., & Liu, Y. (2024). Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1931-1941).

[24] Wang, Y., Yang, D., Bremond, F., & Dantcheva, A. (2022). Latent image animator: Learning to animate images via latent space navigation. arXiv preprint arXiv:2203.09043.

[25] Liu, T., Chen, F., Fan, S., Du, C., Chen, Q., Chen, X., & Yu, K. (2024). AniTalker: Animate Vivid and Diverse Talking Faces through Identity-Decoupled Facial Motion Encoding. arXiv preprint arXiv:2405.03121.

[26] Li, W., Zhang, L., Wang, D., Zhao, B., Wang, Z., Chen, M., ... & Li, X. (2023). One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17969-17978).

[27] Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2020, October). 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 (pp. 484-492).

[28] Xu, S., Chen, G., Guo, Y. X., Yang, J., Li, C., Zang, Z., ... & Guo, B. (2024). Vasa-1: Lifelike audio-driven talking faces generated in real time. arXiv preprint arXiv:2404.10667.

[29] Fan, Y., Lin, Z., Saito, J., Wang, W., & Komura, T. (2022). Faceformer: Speech-driven 3d facial animation with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18770-18780).

[30] Huang, R., Zhong, W., & Li, G. (2022, October). Audio-driven talking head generation with transformer and 3d morphable model. In Proceedings of the 30th ACM International Conference on Multimedia (pp. 7035-7039).

[31] Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., ... & Wang, F. (2023). Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face

DOI: 10.54254/2755-2721/102/20241156

animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8652-8661).

[32] Cho, K., Lee, J., Yoon, H., Hong, Y., Ko, J., Ahn, S., & Kim, S. (2024). GaussianTalker: Real-Time High-Fidelity Talking Head Synthesis with Audio-Driven 3D Gaussian Splatting. arXiv preprint arXiv:2404.16012.