2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

UniGS: Unified 3D Gaussian Splatting for Long-Haired Talking Portraits

Yanping Hu[1][0000-0003-4363-9560], Ting Liu[1], and Yuzhuo Fu[1]

Shanghai Jiao Tong University, Shanghai Minhang 200240, China
{yanping_hu,louisa_liu,yzfu}@sjtu.edu.cn

Abstract. Talking portrait synthesis is a crucial task in computer vision, enabling realistic animations for applications in virtual communication, entertainment, and digital media. Current methods primarily focus on short-hair scenarios, where they rely on rigid segmentation to separate the head from the torso, followed by head reconstruction and simple compositional strategies to combine the head back with the torso. However, these methods face significant challenges when applied to long-haired individuals due to the complex interactions between hair and body, which can lead to visual artifacts and misalignments. In this work, we introduce a novel dataset specifically designed for long-haired individuals, providing a comprehensive benchmark for evaluating head-torso separation in these complex scenarios. Building upon this dataset, we propose UniGS, a unified 3DGS-based framework that holistically models the full portrait, eliminating the need for explicit segmentation. By incorporating audio, eye, and pose features into a deformation network and utilizing a static-to-dynamic training strategy, our method achieves superior realism and coherence. Experimental results show that our approach outperforms existing state-of-the-art techniques in both visual quality and inference efficiency, and effectively handles the complex visual challenges posed by long-haired scenarios. Additional comparisons on existing short-hair datasets further confirm the robustness of our method.

Keywords: Talking Portrait Generation, Head-Torso Separation, 3D Gaussian Splatting.

1 Introduction

In recent years, neural rendering techniques, particularly Neural Radiance Fields (NeRF) [1] and 3D Gaussian Splatting (3DGS) [2], have made significant improvements in facial animation and view synthesis, drawing considerable attention to the task of talking portrait generation [3-6]. These technologies can be applied in a range of applications, including virtual assistants, film dubbing, and remote communication [7-8]. Given the human eye's sensitivity to even the smallest discrepancies, it is essential to generate lifelike and seamless talking portraits. Inaccurate head-torso separation often results in distortions such as misalignment, floating heads, or unnatural body motion, which significantly degrade the user experience in interactive applications.

Despite its importance, head-torso consistency has been largely overlooked in existing methods. Most current techniques adopt segment-dependent frameworks that divide the portrait into distinct head and torso regions with a hard boundary at the jawline to simplify modeling [5,9]. While this enables high-quality head animation, the torso is typically left static or inadequately processed; then the animated head is simply composited back onto the original body. This approach works well in simple cases like short-haired individuals, where the head and torso exhibit distinct boundaries and similar color tones, allowing for straightforward blending [3,10].

However, when applied to long-haired subjects, these methods struggle significantly. Hair often spans the head-torso boundary, causing self-occlusion, dynamic interactions with clothing, and high color contrast at the boundary. These effects disrupt traditional segmentation strategies, amplifying both geometric and photometric inconsistencies, making it difficult to maintain head-torso consistency. Existing methods fail to adequately address these issues, resulting in misalignments, artifacts, and unrealistic body motion in long-hair scenarios.

To address this challenge, we introduce a novel dataset focused on long-haired individuals, providing a comprehensive benchmark for evaluating head-torso consistency in complex scenarios. Based on this dataset, we propose UniGS, a unified 3DGS framework that addresses head-torso boundary inconsistencies through holistic modeling. By jointly encoding audio features, eye movements, and head poses within a unified Gaussian field, our method enables consistent animation across the eyes, face, and torso. Furthermore, we adopt a 'static-to-dynamic' training strategy, similar in spirit to canonical-to-deformation pipelines [11-13], which transitions from learning a stable canonical field to modeling dynamic deformations. This approach reduces dependence on accurate segmentation and removes the need to separately handle the head and torso, enabling efficient and high-fidelity portrait generation.

We validate our framework through both quantitative and qualitative experiments. Results demonstrate that our method significantly outperforms existing techniques in maintaining head-torso consistency and visual realism in long-haired scenarios, while also achieving efficient portrait rendering.

The primary contributions of this work are as follows:

- We introduce a novel dataset focused on long-haired individuals, filling a key gap in existing benchmarks and enabling rigorous evaluation of head-torso separation in complex scenarios.
- We propose UniGS, a unified 3DGS framework that models the entire portrait holistically, offering a novel and effective solution to the head-torso separation problem without relying on explicit segmentation.
- We demonstrate that our approach outperforms state-of-the-art methods in handling long-hair scenarios, while also achieving competitive performance on existing short-hair benchmarks, ensuring efficiency, high-fidelity and generalizability.

(a) (b)

Fig. 1. Comparison of head-torso boundary conditions between existing short-haired benchmarks (a) and our long-haired dataset (b) Short-haired portraits present clear, color-consistent boundaries, enabling simple segmentation. In contrast, long-haired examples exhibit complex, occlusion-prone boundaries with high color contrast, complicating separation. See highlighted boundary zones and sampled pixel color pairs.

2 Dataset

To systematically investigate head-torso separation under visually complex conditions, we introduce a new dataset of long-haired talking portraits. Unlike prior benchmarks [14-16] which predominantly feature short-haired individuals with visually coherent and easy-to-segment boundaries, our dataset emphasizes cases with significant structural ambiguity and high appearance variation across the boundary, as shown in **Fig. 1**. These include hair occlusions, motion-induced distortions, and color discontinuities between head and torso regions, all of which pose significant challenges for both geometric modeling and photometric consistency.

We collected a total of eight videos from YouTube, selecting content published within the past five years that features long-haired individuals speaking natural sentences. Each subject appears under relatively consistent lighting and pose. The original videos are in 720p or 1080p resolution, providing high-fidelity footage suitable for both model training and evaluation.

To standardize the dataset, we apply a facial landmark detector to extract face-centered regions, using a fixed crop window per video to preserve framing consistency. All videos are resized to 480×480 resolutions to balance detail and computational efficiency. To minimize distraction and preserve portrait integrity, we manually remove segments containing occlusions such as hand gestures or large head movements.

While this dataset does not directly improve the performance of existing models, it serves as a stress test for evaluating head-torso consistency under long-hair conditions. Our analysis reveals that state-of-the-art segment-dependent frameworks exhibit significant performance degradation on this dataset, largely due to their reliance on rigid region separation and precise segmentation. These results highlight the need for a holistic modeling strategy that treats the portrait as an integrated whole. This underscores the need for holistic and unified modeling approaches. In response to this limitation, we introduce a unified 3DGS-based framework that models the entire portrait without

explicit head-torso segmentation, offering a more robust and generalizable solution to the challenges posed by long-haired scenarios. We detail this framework in the following section.

3 Methodology

We propose a unified framework for audio-driven talking portrait synthesis that leverages 3D Gaussian Splatting (3DGS) for holistic head–torso modeling, as shown in **Fig. 2**. Unlike traditional methods that treat head and torso separately, UniGS ensures spatial coherence through a single Gaussian field. The model integrates audio, eye, and pose features to achieve precise lip synchronization and realistic motion modeling.

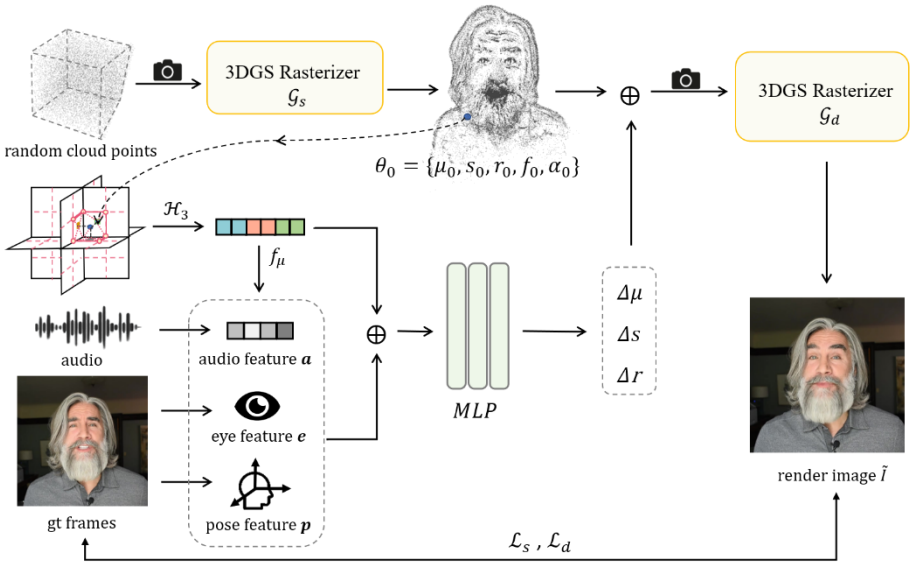

Fig. 2. Overview of UniGS. The framework models a unified portrait with 3D Gaussian Splatting, integrating dynamic head and torso motions. Random cloud points are rasterized by the 3DGS Rasterizer ($\mathcal{G}_s$) to create an initial 3D structure with initial Gaussian parameters (θ_0). Conditional inputs, audio, eye, and pose features, are concatenated with the encoded spatial locations (f_μ) derived from hash encoding ($\mathcal{H}^3$) and processed by a single MLP to predict dynamic updates ($\Delta\mu, \Delta s, \Delta r$) for the Gaussian parameters. These updates refine the 3D representation, which is then rasterized by the final 3DGS Rasterizer ($\mathcal{G}_d$) to generate the rendered image ($\tilde{I}$). Loss functions, including static training loss $\mathcal{L}_s$ and dynamic training loss $\mathcal{L}_d$, ensure the rendered output ($\tilde{I}$) aligns with the ground truth (I), achieving realistic motion and fine-grained details.

3.1 Preliminaries

3D Gaussian Splatting is a point cloud rendering technique that represents scenes using anisotropic ellipsoidal Gaussians. Each Gaussian primitive is defined by its position

$\mu \in \mathbb{R}^3$, rotation $r \in \mathbb{R}^4$, scaling $s \in \mathbb{R}^3$, opacity $\alpha \in \mathbb{R}^1$, and color feature $f \in \mathbb{R}^d$. It is expressed as:

$$g(x) = \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \tag{1}$$

where the covariance matrix Σ is derived from r and s. During rendering, a rasterizer gathers N Gaussians, combining the color c and projected opacity $\hat{\alpha}$ of each to compute the pixel color $\mathcal{C}(p)$:

$$\mathcal{C}(p) = \sum_{i \in N} c_i \, \hat{\alpha}_i \prod_{j=1}^{i-1}(1-\hat{\alpha}_j) \tag{2}$$

3DGS explicitly represents scenes with parameters $\theta = \{\mu, r, s, \alpha, f\}$, which are optimized via gradient descent under color supervision. It employs a densification strategy to grow primitives and prune redundant ones during training.

3.2 Preprocessing

To generate audio-driven talking portraits, we first preprocess the input data to extract useful features. The training data consists of a few minutes of speech video from a single person. Preprocessing begins by resampling the raw audio signal at a rate of 16 kHz. Mel-frequency cepstral coefficients (MFCC) are then extracted from each frame, which has a duration of 350 ms, using a 40-dimensional filter bank energy and a 25 ms sliding window step. We then map these MFCC features into speech embedding vectors using a pre-trained DeepSpeech encoder [17], which captures the semantic elements of the speech and provides discriminative information essential for precise lip synchronization.

Next, we use OpenFace toolkit [18] to extract facial action unit parameters, specifically those related to ocular muscles. These parameters are represented by an eye feature vector e, which encodes facial movements necessary for accurate eye and facial expression synthesis.

Finally, we employ 3D Morphable Model (3DMM) [19] to estimate the rotation matrix $R \in \mathbb{R}^{3\times3}$ and translation vector $t \in \mathbb{R}^3$, which work as the pose features p and camera viewpoints cam. The 3DMM allows us to model the three-dimensional geometry of the face and account for changes in viewpoint during the video, ensuring that the generated portrait maintains correct alignment with the speaker's position and pose.

3.3 Static to Dynamic Gaussian Deformation Training

We initialize a randomly distributed point cloud within a cube, which is fed into a 3DGS rasterizer to learn the coarse static appearance and structure of the portrait. This process generates static Gaussian primitives $\theta_0 = \{\mu_0, s_0, r_0, f_0, \alpha_0\}$, and the rendered static image is:

$$I_s = \mathcal{G}_s(\theta_0, cam) \tag{3}$$

While Gaussian primitives effectively represent the global 3D head, they lack fine-grained regional position encoding. To address this, we adopt three orthogonal 2D multi-resolution hash grids [20] to encode projected coordinates $\mu = (x, y, z)$:

$$\mathcal{H}^{AB}: (a, b) \rightarrow f_{ab}^{AB} \tag{4}$$

where the output $f_{ab}^{AB} \in \mathbb{R}^{LF}$ is the plane-level geometry feature for the projected coordinate (a, b), with the number of levels L and feature dimensions per entry F. Then the outputs of these three encoders will be concatenated as:

$$f_{\mu} = \mathcal{H}^{XY}(x, y) \oplus \mathcal{H}^{YZ}(y, z) \oplus \mathcal{H}^{XZ}(x, z) \tag{5}$$

where $\oplus$ denotes concatenation, and f_{μ} captures plane-level geometry features.

In the deformation prediction network, we employ a 3-layer Multilayer Perceptron (MLP) to predict the deformation parameters for each Gaussian primitive. The input to the MLP consists of concatenated feature vectors, including encoded positional information f_{μ}, audio features a, eye features e, and pose features p. Specifically, for each Gaussian primitive indexed by i, the deformation prediction is computed as follows:

$$\{\Delta\mu_i, \Delta s_i, \Delta r_i\} = MLP(f_{\mu_i} \oplus a_i \oplus e_i \oplus p_i) \tag{6}$$

By adding these deformed parameters to the static Gaussian primitives, we obtain dynamic primitives $\theta_d = \{\mu_0 + \Delta\mu, s_0 + \Delta s, r_0 + \Delta r, f_0, \alpha_0\}$. The final rendered image $\tilde{I}$ is then generated using the 3DGS rasterizer:

$$\tilde{I} = \mathcal{G}_d(\theta_d, cam) \tag{7}$$

3.4 Loss Function

During initialization, we adopt the training strategy of 3DGS [2], using pixel-wise L_1 loss and the D-SSIM term to encourage structural similarity between the rendered image I_s (generated by parameters θ_0) and the ground truth image I. The static stage loss is formulated as:

$$\mathcal{L}_s = \mathcal{L}_1 + \lambda_1 \mathcal{L}_{D-SSIM} \tag{8}$$

Following the static network initialization, we proceed to train the dynamic deformation network. This stage aims to refine the model's capacity to capture dynamic regions of interest, such as the hair, torso, and mouth, enhancing motion consistency and local detail fidelity.

To guide the learning of dynamic deformations, we introduce a region-specific L_1 loss:

$$L_j = \lambda_j \cdot \|I_j - \tilde{I_j}\|_1 \quad where \quad I_j = I \cdot M_j, \quad j \in \{mouth, hair, torso\} \tag{9}$$

Here, I_j denotes the masked ground truth image for region j, with M_j representing the corresponding binary mask. This term enforces pixel-wise accuracy within dynamically deformable regions, ensuring the preservation of fine-grained details and motion alignment.

In addition, we introduce a motion regularization term to promote temporal smoothness and penalize abrupt changes in motion-related parameters, including translation

(μ), rotation (r), opacity (α), and scale (s). The motion loss term $\mathcal{L}_{motion}$ is defined as the sum of the absolute differences of these parameters:

$$\mathcal{L}_{motion} = \lambda_{motion} \cdot (\|\Delta\mu\|_1 + \|\Delta r\|_1 + \|\Delta\alpha\|_1 + \|\Delta s\|_1) \tag{10}$$

Furthermore, inspired by prior works [10,15,21], we also incorporate a perceptual loss based on LPIPS, computed over randomly sampled image patches. This loss encourages the preservation of perceptual similarity and local texture details.

The overall loss function in dynamic training is formulated as:

$$\mathcal{L}_d = \mathcal{L}_s + \sum_j \mathcal{L}_j + \mathcal{L}_{motion} + \lambda_2 \mathcal{L}_{LPIPS} \tag{11}$$

4 Experiments

4.1 Experimental Settings

Implementation Details. Our method is implemented in PyTorch. Adam [22] and AdamW [23] optimizers are used in training. In the loss functions, λ_1, λ_{mouth}, λ_{hair}, λ_{torso}, λ_{motion} and λ_2 are set to 0.2, 0.1, 0.05, 0.1, 1e-5 and 0.2, respectively. We train our model on a single RTX3090 GPU with 180k iterations, which costs about 1 hour.

Evaluation Protocol. In addition to the *longhair* dataset introduced in Section 2, we also collected publicly available speech videos ("May" and "Lieu") used in previous works [10,21] to form *shorthair* dataset. We evaluate our method on both datasets by dividing each video into training and testing sets with a 10:1 ratio. We use standard metrics for quantitative evaluation: PSNR, LPIPS [24], and SSIM [25] assess reconstruction quality; LMD [26] evaluates facial keypoints alignment; and Sync-C, derived from SyncNet [27,28], measures audio-visual synchronization. Additionally, we report training time and inference speed (frames per second, FPS) to evaluate efficiency. Moreover, Wav2Lip uses ground-truth images, invalidating its PSNR, LPIPS, and SSIM scores, and its Sync-C is also excluded from discussion since it is used as a training loss.

4.2 Quantitative Results

We compared the proposed method with several existing approaches, including NeRF-based methods ER-NeRF [21], as well as the 3DGS-based method TalkingGaussian [10]. These methods typically model the head and torso separately or focus exclusively on the head, which inherently results in head-torso separation issues. Additionally, we conducted comparative analyses with 2D-based methods, such as Wav2Lip [4] and IP-LAP [29]. Due to the single branch of the Gaussian field, our method achieves an inference speed of 98 FPS, slightly surpassing the 90 FPS of TalkingGaussian [10]. As shown in **Table 1**, our method outperforms others in all evaluated metrics in *longhair* dataset.

Our method also achieves competitive results on the *shorthair* dataset, despite being trained primarily for challenging long-hair scenarios. Notably, our unified 3DGS

framework delivers strong performance in LMD and Sync-C, indicating its robustness and generalization ability across varying portrait types.

Table 1. Comparison of various methods on Rendering Quality, Motion Quality, and Efficiency on *longhair* and *shorthair* dataset. This dual evaluation demonstrates the generalizability of our method. The best and second-best methods are in **bold** and underline, respectively.

Dataset	Method	PSNR↑	LPIPS↓	SSIM↑	LMD↓	Sync-C↑	Time↓	FPS↑
longhair	Wav2Lip	-	-	-	3.705	8.747	-	21.6
	IP-LAP	17.43	0.2798	0.594	3.838	3.899	-	3.18
	ER-NeRF	17.68	0.2579	0.623	3.324	2.581	2h	32
	TalkingGaussian	23.48	0.1339	0.804	2.311	4.470	1.5h	90
	UniGS(Ours)	**25.37**	**0.1049**	**0.808**	**2.241**	**5.047**	**1h**	**98**
shorthair	Wav2Lip	-	-	-	2.991	11.112	-	21.6
	IP-LAP	**34.98**	0.0422	0.907	3.236	**7.471**	-	3.18
	ER-NeRF	30.97	0.0354	0.906	2.911	6.451	2h	32
	TalkingGaussian	31.81	**0.0335**	**0.921**	**2.695**	6.616	1.5h	90
	UniGS(Ours)	31.35	0.0349	0.910	2.864	7.262	**1h**	**98**

4.3 Qualitative Results

Fig. 3 presents a qualitative comparison of talking portrait synthesis results across various methods. As it is shown, our proposed method demonstrates superior performance in generating seamless and high-quality talking portraits in challenging long-haired scenarios that involve complex interactions between the head and torso.

The white dash boxes highlight the head-torso separation artifact, which is common in segment-dependent modeling methods such as ER-NeRF [21] and TalkingGaussian [10]. These approaches often adopt oversimplified representations of the torso region, which fail to handle challenging long-haired scenarios involving complex motion or occlusions. In contrast, our unified modeling framework effectively integrates the dynamics of both the head and torso, significantly reducing head-torso separation artifacts. This holistic approach overcomes the limitations of independent modeling strategies, resulting in improved alignment and reconstruction quality. Furthermore, Wav2Lip [4] produces overly smoothed lip motions, and IP-LAP [29] lacks accurate lip synchronization and detailed representations such as teeth, while our method generates more precise and realistic facial dynamics.

Overall, these results underscore the effectiveness of our unified framework, which allows for more accurate and natural interactions between the head and torso, particularly in dynamic and occluded scenarios.

Fig. 3. Qualitative comparison of long-haired talking portrait synthesis by different methods. The white dash boxes highlight the unnatural separation of the head and torso, while the blue boxes indicate incorrect mouth motion or blurred reconstructions. Please zoom in for better visualization.

4.4 Ablation Study

To evaluate the effectiveness of various modules and parameter settings in our method, we conducted ablation experiments. First, we tested different parameter configurations of the multi-resolution hash grid shown as **Table 2**. The experimental results indicate that when $L = 12$ and $F = 2$, the model achieves the optimal balance between reconstruction quality and computational efficiency. Thus, this configuration was adopted for all experiments. Second, we investigated the impact of Gaussian attribute deformation shown as **Table 3**. Although altering the position and shape of the Gaussian attributes slightly reduced the reconstruction quality, it significantly improved dynamic effects, demonstrating its effectiveness in generating natural dynamic performance. Overall, these ablation studies highlight the importance of carefully selecting parameter configurations and the value of Gaussian attribute deformation in enhancing dynamic realism, ensuring a robust and efficient model for high-quality reconstruction and natural dynamic performance.

Table 2. the optimal level (L) and dimension (F) of the multi-resolution hash grid

Methods	PSNR ↑	LPIPS ↓	LMD ↓	Sync-C ↑
L=12 F=1	25.07	0.1223	2.252	5.007
L=24 F=1	**25.66**	0.1177	2.247	4.948
L=12 F=2	25.37	**0.1049**	**2.241**	**5.047**
L=6 F=4	25.30	0.1061	2.268	5.096

Table 3. Ablation study on deformed attributes of Gaussian primitives

Methods	PSNR ↑	LPIPS ↓	LMD ↓	Sync-C ↑
$\Delta f, \Delta\alpha$	24.61	0.1457	2.677	4.745
$\Delta\mu, \Delta s, \Delta r$	25.37	**0.1049**	**2.241**	**5.047**
$\Delta f, \Delta\alpha, \Delta\mu, \Delta s, \Delta r$	**25.40**	0.1120	2.379	5.004

5 Ethical Consideration

We hope our method can promote the healthy development of digital industries. However, it is crucial to acknowledge the potential for misuse, which could lead to harmful consequences. In recognition of this, we are committed to supporting the development of tools for detecting deepfakes. We strongly recommend the responsible application of this technology and similar techniques to mitigate any negative impact.

6 Conclusion

We present UniGS, a unified 3D Gaussian Splatting framework that effectively addresses the persistent issue of head-torso consistency in talking portrait synthesis, particularly under the challenging long-haired scene. We introduce a novel *longhair* dataset, which serves as a comprehensive benchmark for evaluating head-torso consistency under complex conditions. By modeling the entire portrait as a single Gaussian field, our framework eliminates the need for explicit head-torso segmentation, enabling natural and coherent motion across both regions. Extensive evaluations show that our UniGS not only surpasses existing 2D-based, NeRF-based, and 3DGS-based baselines in realism and consistency but also achieves high inference efficiency in long-haired scenarios. This contribution paves the way for more consistent and expressive talking portraits in intricate visual settings. Furthermore, our method demonstrates strong generalization capabilities, as shown by its competitive performance on a widely used *shorthair* dataset, proving its broad applicability and robustness.

References

1. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM. 65, 99–106 (2021).
2. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM transactions on graphics(TOG) **42**(4) (2023).
3. Doukas, M.C., Zafeiriou, S., Sharmanska, V.: Headgan: One-shot neural head synthesis and editing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14398–14407 (2021).

4. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: 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 (2020).
5. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8652–8661 (2023).
6. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: Makelttalk: speaker-aware talking-head animation. ACM Transactions On Graphics (TOG). 39, 1–15 (2020).
7. D-ID: D-ID: AI generated video from a single image, https://www.d-id.com/, accessed: 2025-03-30.
8. Synthesia: Synthesia AI, https://www.synthesia.io/, accessed: 2025-03-30.
9. Wang, T.-C., Liu, M.-Y., Tao, A., Liu, G., Kautz, J., Catanzaro, B.: Few-shot video-to-video synthesis. Advances in Neural Information Processing Systems. vol. 32. (2019).
10. Li, J., Zhang, J., Bai, X., Zheng, J., Ning, X., Zhou, J., Gu, L.: Talkinggaussian: Structure-persistent 3d talking head synthesis via gaussian splatting. In: European conference on computer vision. pp. 127–145. Springer (2024).
11. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5865–5874 (2021).
12. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-NeRF: Neural Radiance Fields for Dynamic Scenes. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10313–10322. (2021).
13. Weng, C.-Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: Humannerf: Free-viewpoint rendering of moving people from monocular video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition. pp. 16210–16220 (2022).
14. Tang, J., Wang, K., Zhou, H., Chen, X., He, D., Hu, T., Liu, J., Zeng, G., Wang, J.: Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. arXiv preprint arXiv:2211.12368. (2022).
15. Guo, Y., Chen, K., Liang, S., Liu, Y.-J., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5784–5794 (2021).
16. Peng, Z., Hu, W., Shi, Y., Zhu, X., Zhang, X., Zhao, H., He, J., Liu, H., Fan, Z.: Synctalk: The devil is in the synchronization for talking head synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 666–676 (2024).
17. Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., others: Deep speech 2: End-to-end speech recognition in english and mandarin. In: International conference on machine learning. pp. 173–182. PMLR (2016).
18. Baltrušaitis, T., Robinson, P., Morency, L.-P.: Openface: an open source facial behavior analysis toolkit. In: 2016 IEEE winter conference on applications of computer vision (WACV). pp. 1–10. IEEE (2016).
19. Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A., Dunaway, D.: A 3d morphable model learnt from 10,000 faces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5543–5552 (2016).
20. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM transactions on graphics (TOG). 41, 1–15 (2022).

21. Li, J., Zhang, J., Bai, X., Zhou, J., Gu, L.: Efficient Region-Aware Neural Radiance Fields for High-Fidelity Talking Portrait Synthesis. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7534–7544. IEEE, Paris, France (2023).
22. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. (2014).
23. Loshchilov, I.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018).
24. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018).
25. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing. 13, 600–612 (2004).
26. Chen, L., Li, Z., Maddox, R.K., Duan, Z., Xu, C.: Lip movements generation at a glance. In: Proceedings of the European conference on computer vision (ECCV). pp. 520–535 (2018).
27. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Computer vision–ACCV 2016: 13th asian conference on computer vision, taipei, taiwan, november 20-24, 2016, revised selected papers, part II 13. pp. 87–103. Springer (2017).
28. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Computer vision–ACCV 2016 workshops: ACCV 2016 international workshops, taipei, taiwan, november 20-24, 2016, revised selected papers, part II 13. pp. 251–263. Springer (2017).
29. Zhong, W., Fang, C., Cai, Y., Wei, P., Zhao, G., Lin, L., Li, G.: Identity-Preserving Talking Face Generation with Landmark and Appearance Priors. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9729–9738. IEEE, (2023).