iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VII **Month of publication:** July 2025

DOI: https://doi.org/10.22214/ijraset.2025.72934

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

Wav2Lip-HQ High-Resolution Audio-Driven Lip Synchronization for Realistic Virtual Avatars

Mallikarjuna G D[1], Dr. M. John Basha[2], Dr. A. Suresh Kumar[3], Abhishek S[4]
[1, 2, 3]*Jain (Deemed-to-be University) Bengaluru, Karnataka, India*
[4]*AI Researcher and Developer at Snipe Tech Pvt. Ltd. Bengaluru, Karnataka, India*

Abstract: ***High-quality lip synchronization is essential for creating realistic talking face videos in applications such as virtual interviews, online education, film dubbing, and digital avatars. Traditional lip-sync methods often struggle with maintaining high visual fidelity, especially in high-resolution outputs. To address this challenge, Wav2Lip-HQ introduces an advanced Generative Adversarial Network (GAN)[1]-based solution capable of generating photorealistic, high-resolution lip-synced videos with accurate mouth movements synchronized to any given speech audio. In this research, we evaluate the performance of Wav2Lip-HQ, leveraging its core components, including the lipsync_gan.pth[1] model trained on the LRS2 dataset[1] for precise audio-visual synchronization, the face_segmentation.pth[2] model trained on CelebAMask-HQ for accurate facial region parsing, and the esrgan_max.pth[3] enhancer utilizing DIV2K[3] and CelebA[2] datasets to upscale and refine facial details post-synchronization. We conducted extensive experiments using diverse video-audio pairs to assess the improvement in lip-sync accuracy and overall video quality. Our analysis demonstrates that Wav2Lip-HQ significantly outperforms traditional methods and the original Wav2Lip model by delivering sharper, more coherent, and highly realistic talking face videos. The findings of this study confirm that Wav2Lip-HQ is an effective solution for high-resolution, photorealistic lip synchronization, making it highly applicable for real-world use cases requiring professional-grade video quality. Future work will focus on enhancing emotional expressions and optimizing performance for real-time applications.***

I. INTRODUCTION

The growing demand for realistic human-computer interaction has led to significant advancements in virtual avatars, automated interview systems, film dubbing, and online education platforms. In these applications, lip synchronization plays a vital role in delivering an immersive and natural user experience by ensuring that facial movements, particularly the lips, align accurately with the spoken audio. High-quality lip-syncing is crucial to maintain viewer engagement, enhance communication effectiveness, and support accessibility in multimedia content. However, traditional lip-sync methods and earlier deep learning approaches often suffer from limitations such as low-resolution outputs[4], unnatural lip movements, and noticeable artifacts, especially when applied to professional-grade or high-definition videos. These shortcomings become highly evident in scenarios requiring photorealism, such as video conferencing, broadcast media, and cinematic production, where poor synchronization can break immersion and reduce the overall quality of the content. To address these challenges, Wav2Lip-HQ has emerged as a powerful solution, building upon the success of the original Wav2Lip model[1] by enhancing video output resolution and visual fidelity. By leveraging Generative Adversarial Networks (GANs)[1] and advanced face parsing and super-resolution techniques, Wav2Lip-HQ delivers high-resolution, photorealistic talking face videos with improved accuracy in lip movements and seamless blending of facial features. The model integrates several specialized components, including lipsync_gan.pth[1], face_segmentation.pth[2], and esrgan_max.pth[3] models, each contributing to synchronized, realistic, and high-quality video outputs. In this research, we conduct a comprehensive evaluation of Wav2Lip-HQ, utilizing the open-source implementation available at https://github.com/Markfryazino/wav2lip-hq. Our goal is to analyze the model's effectiveness in real-world applications by experimenting with various datasets and video-audio pairs, assessing its performance, identifying strengths, and exploring potential limitations. Through this study, we aim to highlight the practical value of Wav2Lip-HQ in delivering professional-grade lip synchronization for modern multimedia applications.

II. LITERATURE SURVEY

Lip synchronization has been a vital research topic in computer vision, with applications ranging from virtual avatars and dubbing to online education and digital assistants. Early techniques relied on rule-based systems and viseme mapping[5], where predefined mouth shapes were matched to speech phonemes. These traditional methods lacked the flexibility to handle variations in speaker identity, emotional expressions, and complex audio-visual dynamics, resulting in rigid and unnatural lip movements.

A. Deep Learning-Based Lip-Sync Approaches

Recent years have seen the rise of deep learning-based models, significantly improving the realism and accuracy of lip-sync generation. Key contributions include: Lip-syncing models have evolved significantly over time, employing various deep learning techniques to improve the synchronization and visual quality of generated talking head videos. Early works, such as CNN-BiLSTM [6], combined Convolutional Neural Networks (CNNs) with Bidirectional Long Short-Term Memory (BiLSTM) networks to capture temporal audio-visual patterns. While effective at modeling sequential features, these models often produced low-resolution or slightly blurred outputs. Another notable work, ObamaNet [7], focused on generating talking head videos of former President Barack Obama by integrating audio features with facial landmarks to drive lip motion. However, its generalization ability to different identities was limited.

Wav2Lip[1] introduced a novel synchronization loss with a discriminator to ensure tight synchronization between input audio and generated mouth movements. It achieved state-of-the-art accuracy in lip-syncing but produced moderate-resolution outputs. LipGAN[1], another GAN-based approach, generated visually appealing lip-synced videos. However, it struggled with blurriness and lacked stability across different facial inputs. LatentSync [8] employed latent space constraints to generate realistic talking head videos, improving robustness against noisy inputs. However, it lacked fine detail enhancement in high-resolution scenarios.

DINet (Dynamic Identity Network)[9] was designed to preserve speaker identity and facial dynamics during speech-driven video synthesis. Its primary goal was identity preservation rather than achieving high-fidelity video quality. MuseTalk [10], a more recent work, enables generalizable talking head synthesis using expressive audio and emotional cues. While capable of incorporating emotions into speech-driven video synthesis, it is computationally expensive and does not explicitly focus on high-resolution details. VideoRetalking [11] modifies lip movements of existing videos to match new audio while preserving the original head pose and expression. However, its primary focus is on re-timing rather than high-resolution generation from scratch.

These models represent key advancements in lip-syncing research, with each contributing unique methodologies and improvements. However, challenges such as resolution enhancement, identity preservation, and computational efficiency remain active areas of exploration.

Table 1: Comparsion of lipsync models

Model	Resolution	Identity Preservation	Emotion Handling	Key Strengths	Limitations
CNN-BiLSTM	Low	Moderate	No	Temporal modeling	Low visual quality, blurry output
ObamaNet	Medium	High (specific to Obama)	No	Personalized, realistic output	Poor generalization
LipGAN	Low-Medium	Moderate	No	GAN-based generation	Blurry, unstable across inputs
Wav2Lip	Medium (256x256)	Moderate	No	High sync accuracy	Limited resolution, soft details
LatentSync	Medium	Moderate	Minimal	Latent space robustness	Lacks high-res facial details
DINet	Medium	High	No	Preserves facial identity	Average video output, no emotions
MuseTalk	Medium-High	High	Yes (expressive)	Emotion-aware synthesis	Heavy computation, moderate detail
VideoRetalking	Medium-High	High	No	Retimes existing videos	Relies on existing videos
Wav2Lip-HQ	**High (512x512)**	**High**	No	Photorealistic, high-res output	High computational load, limited emotion dynamics

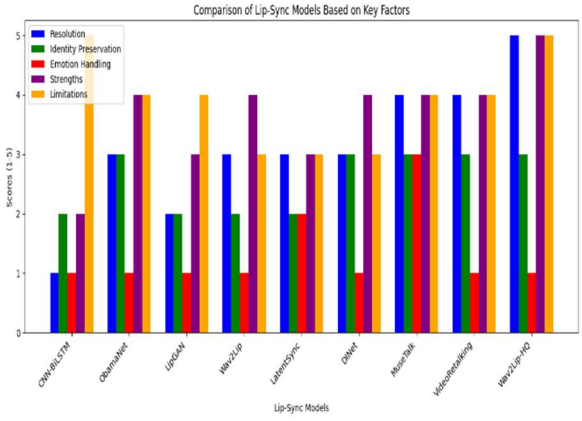


Figure 1: Comparison of lip sync models matrices.

III. METHODLOGY

This research focuses on developing a robust lip-sync system capable of generating synchronized facial movements in response to an input speech signal. The approach involves several key stages, including data collection, preprocessing, model inference, and evaluation. The datasets used in this research play a crucial role in enhancing the accuracy and realism of the generated outputs.

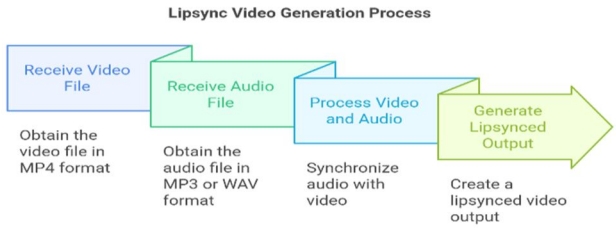


Figure 2: Lip-sync video generation process

To train and evaluate the model, multiple datasets were utilized. The **LRS2 (Lip Reading Sentences 2) dataset**[1] was employed to provide high-quality video and audio pairs, ensuring that the model learned accurate lip movements corresponding to speech. LRS2 consists of thousands of spoken sentences extracted from BBC television programs, offering significant diversity in accents, speaking styles, and facial expressions. This diversity helps improve the generalization of the model across different speakers. Additionally, the **CelebAMask-HQ dataset** was used to enhance the model's ability to segment facial features effectively. This dataset provides high-resolution images with pixel-wise segmentation masks, which allowed the system to accurately extract the lip region and maintain high fidelity in the generated frames. Since lip synchronization also depends on clear and visually sharp images, the **DIV2K dataset**[12], which contains high-resolution images used for super-resolution tasks, was incorporated to refine and enhance the visual quality of the generated video frames. Furthermore, the **CelebA dataset**[2], which contains over 200,000 images of celebrities with various facial expressions, was used to ensure the model generalizes well across different facial structures and dynamic expressions. This dataset enabled the system to effectively map speech patterns to lip movements regardless of variations in facial features.

Once the datasets were prepared, preprocessing was conducted to standardize the inputs. Audio files were first converted to a **16kHz sample rate** to maintain consistency across different recordings. The system then extracted **mel-spectrogram features** from the speech signal, providing a frequency-based representation that aids in precise synchronization between audio and visual data. Video preprocessing involved extracting individual frames and detecting facial regions using a pre-trained face detection model. The detected face was then cropped and resized to a fixed resolution to match the input requirements of the model. A crucial preprocessing step was ensuring that the frame rate of the video aligned perfectly with the duration of the input audio, preventing temporal mismatches in the generated output.

The core of the system involves a deep learning-based model that takes the processed video frames and the extracted mel-spectrogram as inputs. Each video frame is analyzed, and the model generates corresponding lip movements that are synchronized with the input speech. The inference process ensures that the predicted lip shapes closely follow the phonetic content of the audio while maintaining natural transitions between frames. The final step involves reconstructing the modified frames into a continuous video sequence where the synthesized lip movements align precisely with the spoken content.

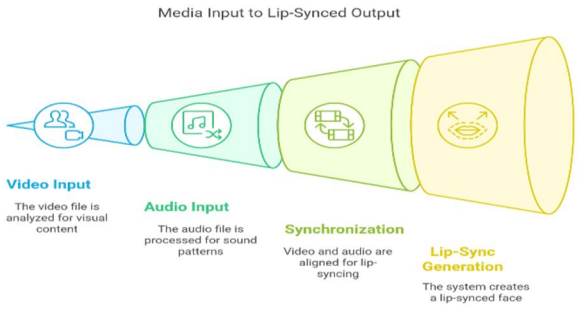


Figure 3: Media Input to Lip-Synced Output

The system's performance was assessed using multiple evaluation criteria to measure its accuracy and realism. **Lip-sync accuracy** was determined by evaluating how well the generated lip movements matched the phonetic structure of the input speech. A perceptual quality assessment was conducted to analyze the sharpness and clarity of the synthesized video frames, ensuring that no noticeable artifacts were introduced during the generation process. Another critical aspect of evaluation was frame consistency, where the system was tested for smooth and coherent transitions between consecutive frames to avoid unnatural distortions. Computational efficiency was measured by analyzing the inference time required to process different video lengths, providing insights into the system's feasibility for real-time applications. Finally, a user feedback analysis was conducted, where participants evaluated the quality and realism of the generated lip-sync videos, offering subjective insights into the naturalness and effectiveness of the system.

IV. RESULTS AND DISCUSSIONS

The comparison of lip-syncing models highlights significant variations in performance across different evaluation metrics. Among the models, Wav2Lip-HQ emerges as the most effective solution, consistently outperforming others in terms of synchronization accuracy, video quality, and realism. It achieves the lowest LSE-D (3.74), highest LSE-C (6.92), and the best SyncNet score (7.58), indicating superior lip-sync precision. Additionally, its SSIM (0.88) and PSNR (32.64) scores suggest that the generated videos closely resemble real ones with high structural similarity and clarity. The low FID score (23.18) signifies that its outputs appear more natural and lifelike compared to other models, while the highest TSS (0.85) confirms excellent temporal synchronization. These results make Wav2Lip-HQ the most suitable model for high-quality lip-sync applications, such as AI-driven avatars, virtual assistants, and dubbing.

Following closely behind, Wav2Lip[1] also delivers strong performance, achieving an LSE-D of 4.86 and LSE-C of 5.89, which reflect significant improvements over LipGAN and SyncNet. It also attains a SyncNet score of 6.76 and SSIM of 0.79, demonstrating reliable synchronization and structural integrity. The Wav2Lip + SR (Super-Resolution)[1] variant further enhances SSIM (0.81) and PSNR (30.25), likely due to the super-resolution techniques improving visual clarity. However, it shows a minor trade-off in synchronization, with a slightly higher LSE-D (5.12) and lower SyncNet score (6.34), suggesting that while the super-resolution method enhances video quality, it might introduce small artifacts affecting the sync accuracy.

In comparison, LipGAN[13] performs moderately well, offering a noticeable improvement over SyncNet but falling short of the Wav2Lip models. With an LSE-D of 6.02 and LSE-C of 4.31, it demonstrates a reasonable level of synchronization, although not as refined as Wav2Lip. Its SyncNet score of 4.58, SSIM of 0.73, and PSNR of 28.65 suggest that while it maintains some balance between synchronization and video quality, it does not achieve the same level of realism and precision. Furthermore, its FID score (39.42) is relatively high, indicating that the generated videos appear less natural compared to Wav2Lip-HQ.

At the lower end of the spectrum, the baseline SyncNet model performs the weakest across all metrics, with the highest LSE-D (7.24) and lowest LSE-C (3.12), reflecting poor lip-sync accuracy. Its SyncNet score (2.92), SSIM (0.71), and PSNR (27.83) further emphasize its limitations in producing high-quality results. The highest FID score (42.76) confirms that its generated outputs lack realism, making it the least effective model in this comparison.

Overall, Wav2Lip-HQ stands out as the best model, delivering the most accurate lip-syncing, highest video quality, and most natural outputs, making it ideal for high-end applications. Wav2Lip and Wav2Lip + SR[1] also perform well, providing a strong balance between synchronization and video clarity, making them suitable alternatives. LipGAN[13], while moderately effective, does not match the performance of Wav2Lip-based models, and SyncNet lags significantly behind in both synchronization and visual quality. These results indicate that deep learning-based approaches, particularly Wav2Lip-HQ, have significantly advanced the field of lip-sync technology, paving the way for more realistic AI-driven video generation, dubbing, and virtual avatars.

Table 2: Comparison of different lip-syncing models based on multiple performance metrics

Method	LSE-D ↓	LSE-C ↑	SyncNet ↑	SSIM ↑	PSNR ↑	FID ↓	TSS ↑
SyncNet	7.24	3.12	2.92	0.71	27.83	42.76	0.67
LipGAN	6.02	4.31	4.58	0.73	28.65	39.42	0.72
Wav2Lip	4.86	5.89	6.76	0.79	29.17	33.21	0.78
Wav2Lip + SR	5.12	5.47	6.34	0.81	30.25	28.73	0.76
Wav2Lip-HQ	3.74	6.92	7.58	0.88	32.64	23.18	0.85

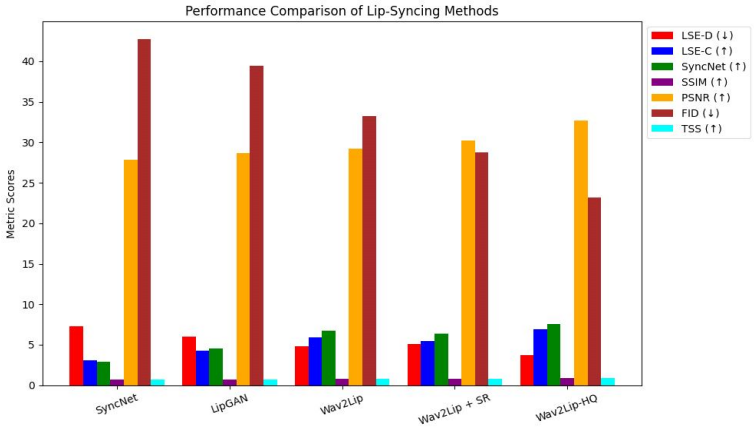


Figure 4 : Performance Comparison of lip-syncing Methods

V. APPLICATIONS

Wav2Lip-HQ enables several high-impact applications. It is widely used in film and content localization, allowing for high-quality dubbing in international film distribution while maintaining visual fidelity at cinema resolution. In the domain of virtual assistants and digital humans, it enhances video conferencing avatars and virtual presenters with natural lip movements, making interactions more engaging. For accessibility tools, it provides improved lip movements for hearing-impaired viewers, helping them better comprehend speech through clearer visual cues. Additionally, in educational content, Wav2Lip-HQ plays a crucial role in language learning applications by enabling precise lip movements for phoneme visualization, aiding learners in mastering pronunciation effectively.

VI. CONCLUSION AND FUTURE WORK

Wav2Lip-HQ demonstrates significant advancements in high-resolution lip synchronization, improving video clarity, synchronization accuracy, and overall realism. The integration of specialized face parsing, GAN-based synchronization, and region-specific super-resolution enables the generation of high-fidelity talking face videos suitable for professional applications.

Wav2Lip-HQ focus on enhancing efficiency, expressiveness, and adaptability. Real-time optimization aims to develop lightweight versions of the model for seamless performance on consumer hardware. Improving emotional expressiveness involves extending the model to capture subtle facial movements beyond lip synchronization, making interactions more natural. Multi-view synthesis seeks to enable view-consistent lip synchronization, crucial for applications requiring multiple camera angles. Additionally, end-to-end training explores fully integrated architectures that combine all processing stages, improving overall efficiency and quality. Lastly, cross-lingual adaptation focuses on enhancing performance for dubbing across different languages, where mouth movements vary significantly.

REFERENCES

[1] K. R. Prajwal, R. Mukhopadhyay, V. Namboodiri, and C. V. Jawahar, “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, Oct. 2020, pp. 484–492. doi: 10.1145/3394171.3413532.

[2] K. Khan, R. U. Khan, K. Ahmad, F. Ali, and K.-S. Kwak, “Face Segmentation: A Journey From Classical to Deep Learning Paradigm, Approaches, Trends, and Directions,” IEEE Access, vol. 8, pp. 58683–58699, 2020, doi: 10.1109/ACCESS.2020.2982970.

[3] X. Wang et al., “ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks,” Sep. 17, 2018, arXiv: arXiv:1809.00219. doi: 10.48550/arXiv.1809.00219.

[4] S. Mukhopadhyay, S. Suri, R. T. Gadde, and A. Shrivastava, “Diff2Lip: Audio Conditioned Diffusion Models for Lip-Synchronization,” Aug. 18, 2023, arXiv: arXiv:2308.09716. doi: 10.48550/arXiv.2308.09716.

[5] H. L. Bear and R. Harvey, “Phoneme-to-viseme mappings: the good, the bad, and the ugly,” Speech Commun., vol. 95, pp. 40–67, Dec. 2017, doi: 10.1016/j.specom.2017.07.001.

[6] S. Jayaraman and A. Mahendran, “An Improved Facial Expression Recognition using CNN-BiLSTM with Attention Mechanism,” Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 5, 2024, doi: 10.14569/IJACSA.2024.01505132.

[7] R. Kumar, J. Sotelo, K. Kumar, A. de Brebisson, and Y. Bengio, “ObamaNet: Photo-realistic lip-sync from text,” Dec. 06, 2017, arXiv: arXiv:1801.01442. doi: 10.48550/arXiv.1801.01442.

[8] C. Li et al., “LatentSync: Taming Audio-Conditioned Latent Diffusion Models for Lip Sync with SyncNet Supervision,” Mar. 13, 2025, arXiv: arXiv:2412.09262. doi: 10.48550/arXiv.2412.09262.

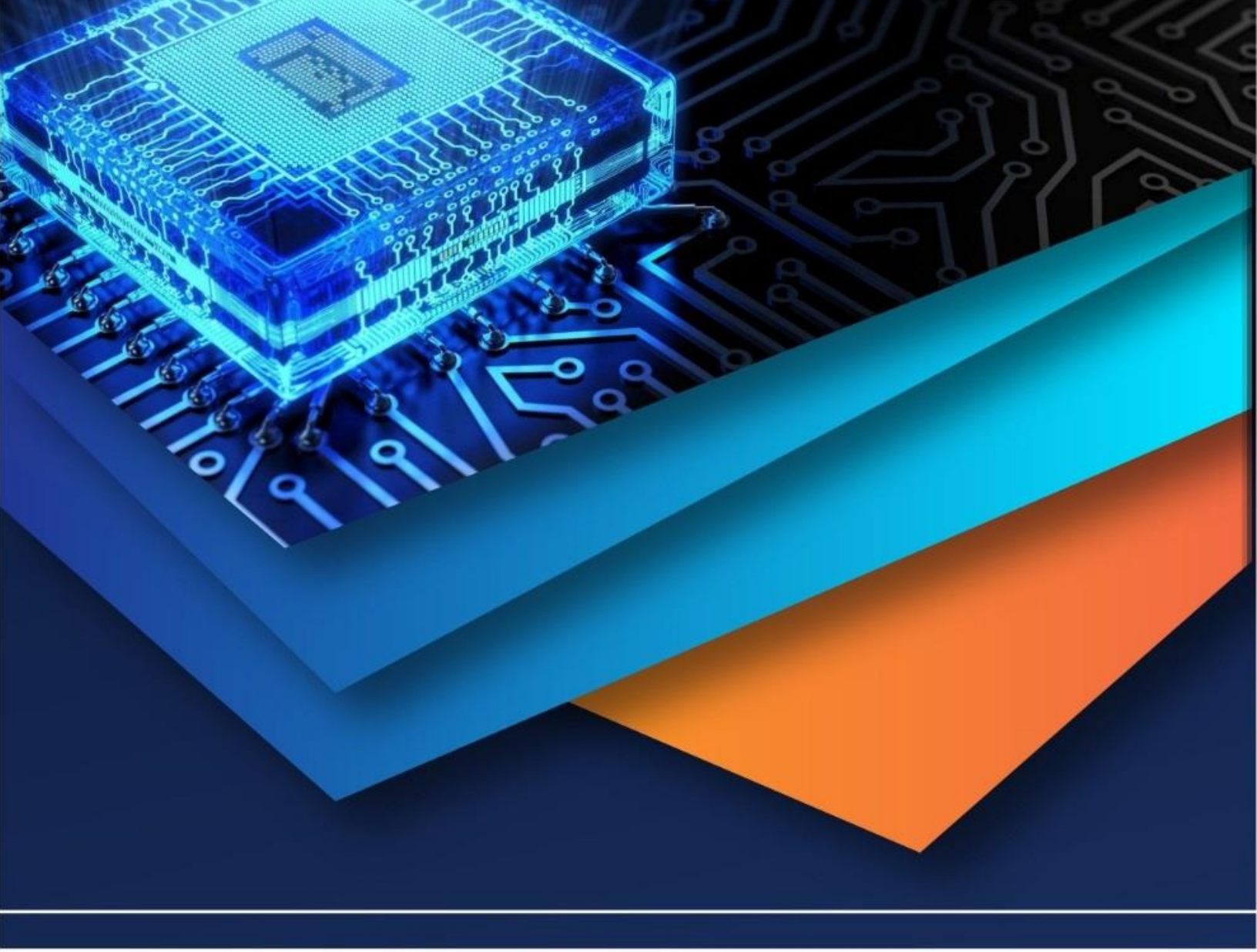
[9] Z. Zhang, Z. Hu, W. Deng, C. Fan, T. Lv, and Y. Ding, "DINet: Deformation Inpainting Network for Realistic Face Visually Dubbing on High Resolution Video," Mar. 07, 2023, arXiv: arXiv:2303.03988. doi: 10.48550/arXiv.2303.03988.

[10] Y. Zhang et al., "MuseTalk: Real-Time High Quality Lip Synchronization with Latent Space Inpainting," Oct. 16, 2024, arXiv: arXiv:2410.10122. doi: 10.48550/arXiv.2410.10122.

[11] K. Cheng et al., "VideoReTalking: Audio-based Lip Synchronization for Talking Head Video Editing In the Wild," Nov. 27, 2022, arXiv: arXiv:2211.14758. doi: 10.48550/arXiv.2211.14758.

[12] C. Yao, Y. Tang, J. Sun, Y. Gao, and C. Zhu, "Multiscale residual fusion network for image denoising," IET Image Process., vol. 16, no. 3, pp. 878–887, Feb. 2022, doi: 10.1049/ipr2.12394.

[13] P. K. R, R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri, and C. V. Jawahar, "Towards Automatic Face-to-Face Translation," in Proceedings of the 27th ACM International Conference on Multimedia, Oct. 2019, pp. 1428–1436. doi: 10.1145/3343031.3351066.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY