High-Resolution Visual Dubbing: Reproduction and Performance Evaluation of DINet

Jiahao Liang*[1], **Ziyi Liu***[1], **Fangfang Xu***[1], **Jiawei Ou***[1], **Xushi Zhang***[1]†

[1]Department of Computer Science, Xiamen University, Xiamen, China
{liangjiahao, liuziyi, xufangfang, oujiawei, zhangxushi}@stu.xmu.edu.cn

Abstract

The technology of facial visual dubbing in high-resolution videos aims to achieve precise synchronization of lip movements with audio while maintaining facial texture details and consistency in identity. Although progress has been made in previous studies, challenges persist in generating high-quality results and ensuring precise synchronization, particularly in resource-constrained scenarios. This study reproduces and evaluates the Deformation Inpainting Network (DINet), which improves the fidelity and naturalness of generated images by combining spatial deformation with image inpainting techniques. Experimental results show that factors such as data scarcity, limited training time, and insufficient diversity in reference images significantly affect both the image quality and audio-visual synchronization performance of DINet. This research not only validates the potential and limitations of DINet under resource constraints but also suggests possible avenues for improvement, such as enhancing data diversity, optimizing network structures, and incorporating multimodal fusion. These findings provide valuable insights for the practical application and technological optimization of high-resolution facial visual dubbing tasks.

Introduction

In the fields of digital media and film production, realistic facial visual dubbing technology has become a critical component in creating convincing virtual characters and enhancing augmented reality experiences. With the rapid advancements in deep learning technologies, facial visual dubbing using a limited number of samples has emerged as a prominent research focus. The core goal of this technology is to achieve precise synchronization of lip movements with audio, while preserving the identity consistency of the original video characters and maintaining coherent head poses (Chung, Jamaludin, and Zisserman 2017). However, despite significant progress in application areas such as virtual anchors (Wiles, Koepke, and Zisserman 2018), film post-production dubbing (Garrido et al. 2015), and language assistants (Kumar et al. 2019), achieving high-fidelity facial visual dubbing in high-resolution video scenarios remains a formidable challenge. These challenges primarily lie in effectively preserving high-frequency texture details in the mouth region, synchronizing the complex dynamic relationship between lip shapes and audio, and ensuring the naturalness and authenticity of the generated results.

In recent years, with continuous innovations in deep learning algorithms, researchers have proposed various methods to enhance the quality of facial visual dubbing. These methods primarily leverage technologies such as Generative Adversarial Networks (GANs), Recurrent Neural Networks (RNNs), and Spatio-Temporal Convolutional Networks to accurately simulate facial expression changes and lip movements in videos. For example, Vougioukas et al. (Vougioukas, Petridis, and Pantic 2020) introduced an architecture that combines Temporal GANs with Spatio-Temporal Convolutional Networks to generate speech-driven facial animations. This approach incorporates a temporal consistency discriminator, which not only ensures dynamic consistency between frames but also produces natural and smooth lip movements. The GAN evaluates the authenticity of the generated images and their synchronization with the input audio through the discriminator, allowing the model to improve progressively during training and ultimately generate higher-quality facial videos. Similarly, Song et al. (Song et al. 2019) proposed a Conditional Recurrent Generative Network that combines the strengths of RNNs and GANs, focusing on modeling the dynamic changes in facial expressions over time. By integrating a spatio-temporal discriminator, this method guarantees overall coherence and high-resolution detail in the generated videos. Additionally, it enhances lip synchronization accuracy with audio by incorporating a lip-reading discriminator, which proves to be particularly effective in multi-language and multi-accent environments.

In addition to the methods mentioned above, another important research direction is multimodal fusion technology. Researchers have explored integrating multimodal information, such as audio, images, and text, to enhance the naturalness of visual dubbing. For instance, Wiles et al. (Wiles, Koepke, and Zisserman 2018) proposed the X2Face network, which combines image encoding, audio encoding, and pose encoding to generate and control facial animations, ensuring that the generated lip movements are highly consistent with the input multimodal information. This approach

*These authors contributed equally.

†Corresponding author.

facilitates feature transfer across different speakers, resulting in more diverse and natural facial expressions. However, many of these methods overlook the fine-grained relationship between the mouth region and audio features when learning the complex mapping between audio and lip movements. As a result, they struggle to preserve high-frequency texture details in high-resolution videos, which negatively impacts the overall quality of the generated content.

This study focuses on the Deformation Inpainting Network (DINet), proposed by Zhimeng Zhang et al. (Zhang et al. 2023) in 2023, by reproducing and evaluating its potential for practical application in resource-constrained scenarios. DINet combines spatial deformation and image inpainting techniques to achieve precise synchronization between lip movements and audio, while preserving high-frequency texture details in the face. The core design of DINet consists of two main modules: the deformation module and the inpainting module. The deformation module applies spatial deformation operations to the reference image, aligning the lip shape features with the input audio, while ensuring consistency in head poses with the source image. The inpainting module then blends the deformed feature map with other features of the source image, effectively filling in the texture details of the mouth region and generating a highly detailed and realistic dubbing video.

To evaluate the performance of DINet under resource-constrained conditions, this study designed experiments focusing on the generation quality and audio-visual synchronization of DINet when using small-scale datasets and limited computational resources. The experimental results show that, under these constraints, the image quality and lip synchronization of the generated output from DINet are inferior to those produced by the model trained under ideal conditions. This study provides a detailed assessment of the performance gap through quantitative analysis and, based on the findings, summarizes the potential and limitations of the model in resource-constrained scenarios. These insights are valuable for guiding future improvements and offer an important reference for the practical application and development of high-resolution facial visual dubbing tasks.

Related Work

Our work primarily focuses on the field of face generation, with an emphasis on generating speaking faces. It consists of two main components: Talking Face generation and spatial deformation.

Talking Face Generation

Talking face generation aims to synthesize facial images based on driving audio or text. It encompasses three main directions: one-shot talking face, few-shot face visual dubbing, and person-specific talking face.

One-shot Talking Faces One-shot talking face technology focuses on driving a reference face image to produce synchronized lip movements, realistic facial expressions, and natural head movements.

Few-shot Face Visually Dubbing Few-shot face visual dubbing focuses on modifying the mouth area of the source face based on the driving audio.

Person-specific Talking Face Person-specific talking face techniques require the identity to be present in the training data.

Spatial Deformation

In deep learning-based approaches, spatial deformation can be achieved through two main methods: affine transformation and dense flow. The affine transformation involves calculating the transformation coefficients and then applying the deformation to the image feature map. In contrast, dense flow computes the complete dense flow directly using the network, which is subsequently used to deform the feature map. In our work, we plan to use affine transformations for spatial deformation, as they yield relatively better results.

Proposed Solution

In current high-resolution video facial visual dubbing research, several challenges remain unresolved. Specifically, in resource-constrained scenarios, achieving precise synchronization between lip movements and audio signals while maintaining high generation quality is a key difficulty. Many traditional methods rely on direct generation strategies but often fall short of meeting the required resolution and detail in the generated images. Furthermore, the limited availability of training data restricts the model's generalization ability and applicability, making it difficult for existing methods to perform effectively in complex real-world scenarios.

To address these challenges, Zhimeng Zhang et al. (Zhang et al. 2023) proposed the Deformation Inpainting Network (DINet), specifically designed for high-resolution facial visual dubbing tasks in resource-constrained environments. DINet combines spatial deformation and image inpainting techniques to precisely synchronize lip movements while enhancing the fidelity and naturalness of the generated images. The architecture of DINet consists primarily of two modules: the deformation module and the inpainting module. These modules work in tandem to produce high-quality facial visual dubbing.

The core function of the deformation module is to extract features from both the source image and driving audio, and generate the appropriate spatial deformation operations. Specifically, this module first extracts head pose features from the source image and lip movement features from the driving audio. It then applies affine transformations to adjust the feature map of the reference image, ensuring precise alignment of lip movements with the speech features in the input audio, while maintaining consistency in head pose. The innovation of this module lies in its use of an adaptive affine transformation strategy, which allows for high-precision deformation of both the global structure and local details of the image. This approach eliminates the need to regenerate pixels, thereby preserving the original texture details to the greatest extent possible.

The inpainting module further processes the deformed feature map to generate a complete and natural dubbing facial image. Through a series of convolutional layers, it seamlessly blends the deformed feature map with other features from the source image, particularly focusing on filling in the details of the mouth region. This ensures precise synchronization between the lip movements and the audio-driven input, while also enhancing the realism of texture details, thereby significantly improving the quality and naturalness of the generated image.

The innovation of DINet lies in two key aspects: First, the deformation module employs an adaptive affine transformation strategy. By calculating affine coefficients specific to feature channels, it performs high-precision spatial deformation operations, offering greater efficiency and stability compared to traditional dense optical flow methods, particularly in high-resolution image processing tasks. Second, by integrating the deformation and inpainting modules, DINet overcomes the limitations of conventional direct generation methods, significantly enhancing the synchronization accuracy of lip movements and preserving facial texture details.

Compared to existing methods, DINet demonstrates significant advantages. For instance, the Wav2Lip method (Prajwal et al. 2020) often neglects texture details in the mouth region, resulting in blurry lip movements. In contrast, DINet directly transfers high-frequency textures from the source image to the mouth region, producing clearer and more realistic lip movements. The MakeitTalk method (Zhou et al. 2020), due to limitations in facial landmark accuracy, frequently generates distorted lip movements. In comparison, DINet enhances the accuracy of lip motion by directly learning the mapping between audio and lip images. Furthermore, unlike the PC-AVS method (Zhou et al. 2021), which generates pixels directly, DINet avoids this approach and instead combines deformation and inpainting, effectively improving resolution and preserving texture details in the generated images.

In practical applications, DINet is particularly well-suited for resource-constrained scenarios due to its flexibility. For example, when trained on small-scale datasets, the deformation module of DINet optimizes the use of limited reference image information, while the inpainting module significantly enhances the overall quality of the generated images through feature fusion. To evaluate DINet's performance in such environments, experiments were conducted to assess its generation quality and audio-visual synchronization under small-scale datasets and limited computational resources. A quantitative comparison with existing methods was performed, analyzing DINet's potential in these conditions and identifying future directions for optimization.

Experiments

The aim of this experiment is to validate the performance of DINet under small-scale datasets and limited computational resources, and to compare it with the model trained by the original authors using the original English dataset under ideal conditions. The study analyzes the impact of resource constraints and dataset switching on the generation quality of DINet and evaluates its practical application potential in resource-constrained scenarios. Finally, the limitations of DINet are summarized, and future optimization directions are proposed. The experimental design includes data sources and preprocessing, experimental environment and training strategies, as well as the evaluation process for model performance.

Data Sources and Pre-processing

To simulate resource-constrained scenarios, this experiment strictly controlled the scale of data usage. The experimental data comes from the Chinese Mandarin Lip Reading dataset (CMLR), which consists of 102,072 spoken sentences recorded by 11 speakers from national news programs between June 2009 and June 2018. Each sentence contains no more than 29 Chinese characters and excludes English letters, Arabic numerals, and rare punctuation marks. This dataset is primarily used for visual speech recognition research. To simulate data scarcity, a subset of video clips was randomly selected from the dataset to construct the training and testing sets. The training dataset consists of 100 video clips, each approximately 4 seconds long with a frame rate of 25 frames per second, totaling around 10,000 frames of images. To further reduce data diversity, the video clips in the training set were limited to specific subject identities and facial expression variations, in order to simulate data scarcity in real-world applications. The testing dataset consists of 10 video clips, matching the training set in length, with approximately 975 frames of images. The testing set features slight differences in identity and facial expressions compared to the training set, allowing for an evaluation of the model's performance in data extrapolation.

During the data preprocessing stage, we performed uniform frame rate resampling on all the videos and used the OpenFace tool to extract and align the 68 facial key points for each frame. The facial region of each frame was cropped to a resolution of 416×320, and the mouth region was further extracted, with the resolution adjusted to 256×256. The audio features were extracted using a pre-trained DeepSpeech model, resulting in 29-dimensional deep audio features, ensuring that they corresponded to the time steps of the image frames. After these preprocessing steps, both the video and audio data were standardized in terms of spatio-temporal features, providing normalized input data for model training.

Experimental Setup and Training Strategy

The experiment was conducted on a workstation equipped with an NVIDIA RTX 3090 24GB GPU, utilizing the PyTorch framework for model training and inference. To better simulate resource-constrained scenarios, limitations were imposed on parameters such as training time and batch size, and adjustments were made to the optimization strategy.

The training time was approximately 5 hours, completing a total of 60 training epochs. Although this training period is significantly shorter than the training time used by the original authors under ideal conditions, it provides practical reference value for small-scale datasets and high-resolution input tasks in resource-constrained environments.

Due to memory limitations, the batch size was set to 10. While this setting may introduce instability in gradient estimation, it was effectively managed and controlled through the optimization strategy. The Adam optimizer was used with an initial learning rate of 0.0001, and a dynamic learning rate adjustment strategy was applied. This strategy gradually decreased the learning rate based on the model's training progress to stabilize the convergence process.

In each training step, the model's input data consisted of one source image frame, the corresponding audio features for a 5-frame time window, and 5 reference image frames. This multi-source information fusion design was intended to mitigate the potential decline in generation quality due to the insufficiency of single-frame information. During the training process, we strictly adhered to the constraints of the resource-limited scenario, closely monitoring DINet's convergence performance and generation quality under the conditions of limited training time and small sample sizes.

Evaluation Metrics and Testing Procedure

To comprehensively assess the performance of the Deformation Inpainting Network (DINet) under resource-constrained conditions, this study employed a range of quantitative metrics and analyzed the model's generation results through a standardized testing process. The primary evaluation metrics encompassed two dimensions: visual quality and audio-visual synchronization performance, which were used to measure DINet's ability to generate detailed images and achieve accurate temporal alignment.

The visual quality was evaluated using three metrics: Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS). SSIM and PSNR primarily measure low-level structural and pixel-level errors, focusing on the retention of image details and overall structural consistency. In contrast, LPIPS focuses on high-level semantic differences, assessing the perceptual quality of the generated images.

The evaluation of audio-visual synchronization performance was conducted using two metrics: Lip Sync Error Distance (LSE-D) and Lip Sync Confidence (LSE-C). LSE-D measures the temporal alignment between the generated mouth movements and the input audio signal, with smaller values indicating better synchronization. LSE-C reflects the confidence in the synchronization of mouth movements with the audio pronunciation, with higher values indicating greater reliability in the model's generated results.

In the testing process, the model generated corresponding results for each frame in the test set, which were then compared with those of the DINet model trained by the original authors under ideal conditions, based on the metrics mentioned above. Through quantitative data analysis, we were able to visually assess the impact of resource limitations on DINet's performance. Additionally, to gain a more comprehensive understanding of the model's strengths and weaknesses, the study also examined video examples to observe the dynamic consistency and mouth shape detail in the generated results.

Experimental Results and Analysis

The experimental results are summarized in Table 1, which presents the main quantitative metrics for DINet under different training conditions, including SSIM, PSNR, LPIPS, LSE-D, and LSE-C. Significant differences are observed between the model trained under ideal conditions by the original authors and the model trained under resource-constrained conditions in this experiment. These differences clearly highlight the impact of resource limitations on the model's performance.

As shown in Table 1, resource limitations had a significant impact on DINet's generation performance. Firstly, in terms of image quality, the model trained by the original authors performed excellently on the SSIM and PSNR metrics, generating images with clear structures and rich details. This clearly highlights the importance of sufficient training data and computational resources for the model to effectively learn facial details and overall structure. In contrast, the model trained under resource-constrained conditions in this experiment demonstrated poorer performance in retaining details and pixel-level accuracy, particularly evidenced by a significant decrease in the PSNR metric.

Secondly, the significant increase in the LPIPS metric indicates that, under resource-constrained conditions, the perceptual quality of the images generated by DINet, in terms of high-level semantics, is lower. This gap is primarily due to the insufficient diversity in the training data, which made it challenging for the model to learn a rich distribution of facial features, leading to generated images that lack naturalness and consistency.

Finally, in terms of audio-visual synchronization performance, both the LSE-D and LSE-C metrics show a clear degradation trend. The LSE-D value for the model trained under resource-constrained conditions is significantly higher than that of the original authors' model, indicating a substantial deviation in the temporal alignment between the generated mouth movements and the audio signal. The significant decrease in the LSE-C value further suggests that the model's confidence in synchronizing mouth movements with the audio pronunciation is insufficient. These results highlight the challenge of effectively capturing the complex mapping between audio and mouth movements when training data and computational resources are limited.

Conclusion and the Impact of Resource Constraints

The experimental results indicate that resource constraints not only limited DINet's performance during the training phase but also had a significant impact during the generation phase. Specifically, the lack of sufficient training data and diversity directly constrained the model's generalization ability. When the test set included unseen facial expression combinations or mouth shape variations, the generated results tended to be distorted, often exhibiting a blurry, white haze(as shown in Figure: 1). This issue is tentatively attributed to the small training dataset and limited number of training iterations, which contributed to suboptimal model performance. Additionally, the limited training time

Table 1: Comparison of DINet Performance Metrics

Metrics	SSIM ↑	PSNR ↑	LPIPS ↓	LSE-D ↓	LSE-C ↑
DINet trained by the original authors	0.9425	30.0082	0.0289	8.3771	6.8416
DINet in this experiment (small-scale)	0.7436	20.3419	0.1473	12.7513	4.2351

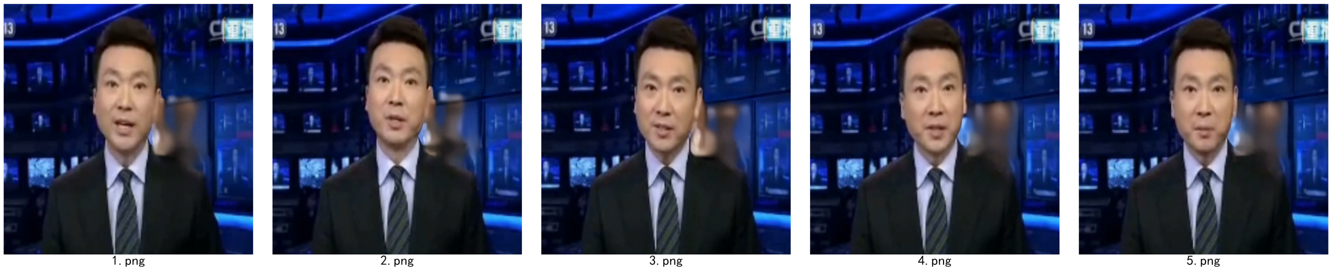

Figure 1: The model generates distorted results with a patch of white haze.

and small-batch training strategy led to insufficient parameter learning, which further affected the model's convergence speed and the quality of the generated output.

On the other hand, the limitation in the number of reference images also has a significant impact on the feature deformation process. The spatial transformation operations employed in the design of DINet rely heavily on the information provided by the reference images. However, when the distribution of reference images is insufficient, the generated results may suffer from detail loss or local inconsistencies. These limitations are particularly evident in the dynamic representation of lip movements and the coherence of head poses.

In summary, the impact of resource limitations on the performance of DINet affects the entire generation process, leading to issues such as degraded image quality, reduced synchronization accuracy, and insufficient dynamic details. The experimental results further highlight the importance of data diversity, adequate training time, and sufficient hardware resources for high-resolution facial visual dubbing technology. These findings also offer clear directions for future model optimization and algorithm design.

Conclusion

This study reproduces and evaluates the performance of the Deformation Inpainting Network (DINet) under resource-constrained conditions, with a focus on its application in high-resolution facial visual dubbing tasks. By combining spatial deformation and image inpainting techniques, DINet seeks to enhance the synchronization of lip movements with audio while preserving facial texture details. Although the network faces limitations in generation quality and audio-visual synchronization when tested with a small dataset and limited computational resources, its innovative design offers a novel approach to tackling high-fidelity dubbing tasks in such constrained environments.

The experimental results validate DINet's practical performance in resource-constrained scenarios, while also highlighting key factors that impact its effectiveness. The findings reveal that insufficient data, limited training time, and a lack of diversity in reference images significantly hinder the model's performance in areas such as image quality, lip movement synchronization, and facial detail preservation. These challenges underscore the limitations DINet faces in real-world applications and offer clear directions for future optimization efforts.

To fully realize DINet's potential and advance its application in practical scenarios, future research should focus on optimizing several key areas. First, the diversity and volume of data directly impact the model's ability to generalize and the quality of its output. Therefore, constructing high-quality datasets or applying data augmentation techniques can enable the model to better learn the complex relationship between audio and lip movements, ultimately improving the accuracy and naturalness of the generated results. Second, optimizing training strategies and efficiently utilizing hardware resources are essential for enhancing model performance. Extending training time, increasing batch size, or adopting distributed training methods can help mitigate resource constraints, while strategic management of hardware resources can boost performance and reduce computational costs.

Additionally, leveraging transfer learning and pre-trained models presents an effective solution to resource constraints. By utilizing high-quality pre-trained models or applying transfer learning, DINet's performance on small-scale datasets can be significantly improved, while also accelerating its convergence. This approach is particularly advantageous for practical applications with limited training data. Moreover, optimizing the network architecture is essential. Designing more lightweight and efficient architectures can reduce the reliance on computational resources while enhancing the representation of details and dynamic consistency in the generated images, making the model more suitable for real-time applications.

Future research could explore the potential of multimodal fusion techniques. By integrating audio, image, and text modalities, DINet's ability to generate natural lip movements and preserve facial details could be further enhanced. This approach could also extend DINet's applicability to multiple languages and scenarios, broadening its potential for a wider range of applications.

In conclusion, this study has validated DINet's performance under resource-constrained conditions and outlined key areas for optimization. The design concept of DINet offers valuable implications for high-resolution facial visual dubbing, presenting new approaches to tackle the challenges of high-fidelity dubbing and providing insights for future research and applications. With ongoing optimization of data, algorithms, and hardware, DINet is poised to deliver greater value in practical scenarios, paving the way for breakthroughs in high-resolution video generation.

References

Chung, J. S.; Jamaludin, A.; and Zisserman, A. 2017. You said that? *arXiv preprint arXiv:1705.02966*.

Garrido, P.; Valgaerts, L.; Wu, C.; Beeler, T.; Thompson, W.; Casas, D.; and Theobalt, C. 2015. VDub: Modifying Face Video of Actors for Plausible Visual Alignment to a Dubbed Audio Track. *Computer Graphics Forum*, 34(2): 193–204.

Kumar, R.; Sotelo, J.; Kumar, K.; Sotocinal, P.; and Courville, A. 2019. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. 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*, 484–492.

Song, Y.; Zhu, J.; Li, D.; Wang, A.; and Qi, H. 2019. Talking face generation by conditional recurrent adversarial network. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, 919–925.

Vougioukas, K.; Petridis, S.; and Pantic, M. 2020. Realistic Speech-Driven Facial Animation with GANs and Temporal Convolutional Networks. *International Journal of Computer Vision (IJCV)*, 128(3): 555–576.

Wiles, O.; Koepke, A. S.; and Zisserman, A. 2018. X2Face: A network for controlling face generation using images, audio, and pose codes. In *European Conference on Computer Vision (ECCV)*, 670–686.

Zhang, Z.; Hu, W.; Deng, C.; Fan, T.; Lv, T.; and Ding, Y. 2023. DINet: Deformation Inpainting Network for Realistic Face Visually Dubbing on High Resolution Video. *Journal Name*, Volume(Issue): page numbers.

Zhou, H.; Sun, Y.; Wu, W.; Loy, C. C.; Wang, X.; and Liu, Z. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4176–4186.

Zhou, Y.; Han, X.; Shechtman, E.; Echevarria, J.; Kalogerakis, E.; and Li, D. 2020. Makelttalk: speaker-aware talkinghead animation. *ACM Transactions on Graphics (TOG)*, 39(6): 1–15.