SPEAR: SPADE-Net and HuBERT for Enhanced Audio-to-Facial Reconstruction

Xuan-Nam Cao[1,2][a] and Minh-Triet Tran[1,2][b]
[1]*Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam*
[2]*Vietnam National University, Ho Chi Minh City, Vietnam*
fi

Keywords: Generative AI, Audio-to-Face Generation, HuBERT, VAE, SPADE, Optical Flow, Landmark Prediction.

Abstract: Generating talking faces has become an essential area of research due to its broad applications. Previous studies in facial synthesis have faced challenges in maintaining consistency between input landmarks and generated facial images, especially when dealing with complex expressions or pose variations. To address these challenges, this paper proposes a novel generative approach for face synthesis driven by audio, pose, and reference images. The proposed system combines a pretrained Variational Autoencoder (VAE), Transformer encoders, SPADE (Spatially Adaptive Normalization) modules, and optical flow-based warping to generate realistic facial images. The system utilizes HuBERT for audio feature extraction, a pose encoder for capturing pose-driven features, and a reference encoder to provide contextual facial information. The generated face, incorporating audio cues, pose variations, and reference images, is refined through optical flow to align with the driven pose and landmarks, ensuring high fidelity and natural facial animation. Experimental results demonstrate the effectiveness of this system in generating high quality, emotion driven facial animations.

1 INTRODUCTION

Talking face generation has emerged as a prominent research area due to its potential applications in virtual avatars, gaming, and human-computer interaction. The ability to generate realistic and expressive facial animations driven by audio and other cues is crucial for enhancing user engagement and realism in these applications.

Previous studies on facial synthesis have primarily relied on convolutional neural networks (CNNs) (Vougioukas et al., 2020; Eskimez et al., 2021), sometimes combined with LSTM(Hochreiter and Schmidhuber, 1997) architectures (Zhou et al., 2020; Wang et al., 2020a; Song et al., 2022; Cao et al., 2024a), or employed Transformer-based models (Gong et al., 2021; Verma and Berger, 2021; Cao et al., 2023; Cao et al., 2024b) to generate images conditioned on input landmarks. While these approaches have achieved notable progress, they often face challenges in preserving fine-grained consistency between the input landmarks and the synthesized facial images, particularly when dealing with complex expressions or significant pose variations. Global normalization techniques commonly used in these models fail to preserve critical spatial details, leading to noticeable degradation in regions such as the mouth or eyes. Additionally, multi-modal approaches that incorporate audio or pose inputs often face challenges in aligning these inputs with spatial structures, resulting in incoherent or less natural outputs.

In audio-driven facial synthesis, methods like MFCC (Abdul and Al-Talabani, 2022), Wav2Vec Conformer (Baevski et al., 2020) (Gulati et al., 2020), and UniSpeech (Wang et al., 2021) show progress but face notable limitations. MFCC captures basic spectral properties but lacks modeling of complex relationships. Wav2Vec Conformer combines CNNs and transformers but struggles with long-term dependencies and generalization. UniSpeech excels in speech representation but is less effective in aligning speech with facial animations. These methods fail to fully exploit the intricate relationship between audio and facial landmarks, limiting their ability to generate expressive facial animations.

To address these issues, we propose a novel approach for audio-driven face synthesis using pretrained Variational Autoencoders (VAEs) (Kingma and Welling, 2019), Transformer encoders, and

[a] https://orcid.org/0000-0002-3614-7982
[b] https://orcid.org/0000-0003-3046-3041

Cao, X.-N. and Tran, M.-T.
SPEAR: SPADE-Net and HuBERT for Enhanced Audio-to-Facial Reconstruction.
DOI: 10.5220/0013349100003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 1311-1318
ISBN: 978-989-758-737-5; ISSN: 2184-433X

SPADE (Park et al., 2019). Audio features are extracted using HuBERT for rich acoustic representations, while pose and reference encoders capture variations and guide synthesis. Optical flow-based warping ensures alignment with input pose and landmarks, improving naturalness and accuracy. Experiments highlight the method's ability to produce high-quality, emotion-driven facial animations, advancing virtual avatars and related applications.

The main contributions of our research are as follows:

- We demonstrate the effectiveness of HuBERT in landmark prediction, achieving superior performance in both landmark distance (LD) and landmark velocity distance (LVD), surpassing models such as MFCC, Wav2Vec Conformer, and UniSpeech.

- We propose SPADE-Net, a novel architecture aimed at improving the quality of facial landmark prediction and image generation, providing a promising solution for tasks requiring precise face synthesis.

The rest of this paper is organized as follows: Section Section 2 reviews related work. Section 3 details the proposed model. Sections 4 and 5 present experiments and results. Section 6 concludes with future directions.

2 RELATED WORK

2.1 Audio-Based Talking Head Generation

In audio-based facial generation, various deep learning methods have been explored to capture temporal relationships and extract key features from audio signals. LSTM networks (Hochreiter and Schmidhuber, 1997), for example, are effective at modeling sequential data and have been used in systems like Makeittalk (Zhou et al., 2020) to extract content and emotional features from audio. Other studies, such as MEAD (Wang et al., 2020a) and Song et al. (Song et al., 2022), have leveraged LSTMs to capture temporal patterns and emphasize emotional expressions.

CNNs, on the other hand, are useful for capturing local patterns in audio data. Approaches like those from Vougioukas et al. (Vougioukas et al., 2020) and Eskimez et al. (Eskimez et al., 2021) combine CNNs with LSTMs to capture both local features and temporal dependencies, providing more comprehensive feature extraction.

Recently, integrating Transformers with CNNs has gained attention, as seen in studies like AST (Gong et al., 2021) and Verma et al.(Verma and Berger, 2021). This combination allows for capturing both local and long-range dependencies, making it particularly effective for facial landmark prediction tasks.

2.2 Audio Representation Models for Facial Landmark Prediction

HuBERT (Hsu et al., 2021), Wav2Vec Conformer (Baevski et al., 2020)(Gulati et al., 2020), and UniSpeech (Wang et al., 2021) are state-of-the-art models for audio representation learning, each offering distinct advantages in capturing audio features. HuBERT (Hsu et al., 2021) utilizes self-supervised learning to model complex speech patterns, demonstrating superior performance in various tasks, including landmark prediction. Wav2Vec Conformer is designed similarly to Wav2Vec (Baevski et al., 2020), but with the attention block replaced by a conformer block (Gulati et al., 2020). This modification enables Wav2Vec Conformer to combine the advantages of conformer blocks, excelling in capturing both local patterns and long-range dependencies in audio data. UniSpeech (Wang et al., 2021), a multi-task model, integrates acoustic and linguistic features, providing robust representations for diverse speech processing tasks. These models have shown promising results in facial landmark prediction, outperforming traditional methods like MFCC in terms of accuracy and temporal coherence.

3 PROPOSED METHOD

Figure 1 provides an overview of the framework proposed in this work. The input comprises three components: audio, a driving video, and reference landmarks, which can be randomly selected from the same driving video. In Stage 1, the Audio-to-Landmark module 3.1, these inputs are used to predict the facial landmarks. These predicted landmarks are then passed to Stage 2, the Landmark-to-Face module 3.2, which utilizes a pretrained VAE and SPADE-Net to generate the final talking face image.

3.1 Audio to Landmark

3.1.1 Audio Encoder

In our approach, we leverage HuBERT (Hidden-Unit BERT), a self-supervised pre-trained speech representation model, to extract robust audio features. HuBERT processes raw audio waveforms and outputs hidden representations from multiple layers, each

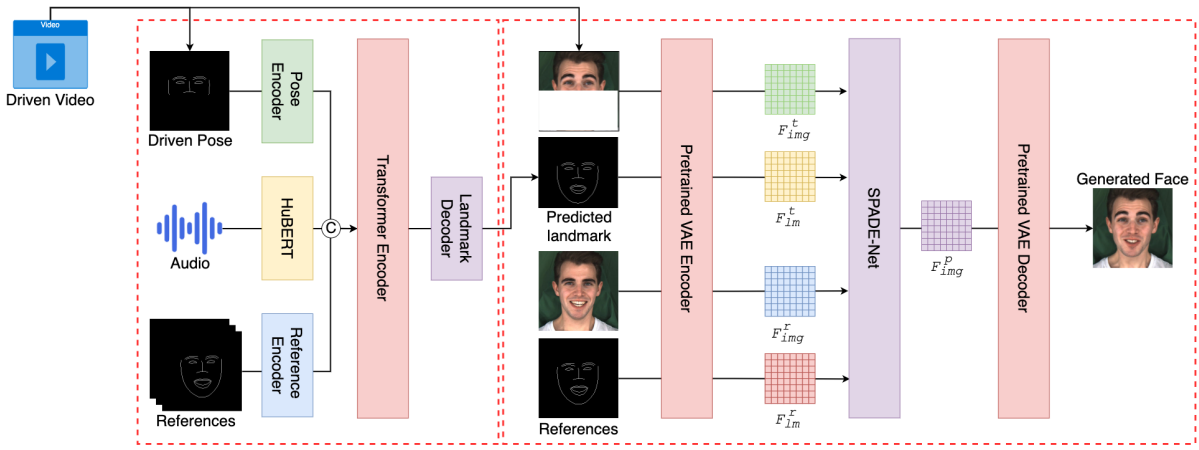

Figure 1: General visualization of the proposed framework.

capturing different levels of acoustic and linguistic information. Specifically, we utilize the outputs from the 6th, 12th, and 24th hidden layers of HuBERT. Each layer produces a feature vector of size $(1024,)$, representing the latent features extracted from the audio at different hierarchical levels. These three feature vectors are concatenated along the channel dimension, resulting in a combined feature representation of shape $(3, 1024)$.

The choice of the 6th, 12th, and 24th layers is intentional, as it reflects the hierarchical nature of HuBERT's learned representations. The lower layers (e.g., the 6th) capture low-level acoustic features such as phonemes and prosody, essential for preserving speech rhythm and intonation. The middle layers (e.g., the 12th) focus on intermediate representations, balancing acoustic and semantic information, while the higher layers (e.g., the 24th) encode high-level linguistic features and contextual understanding, which are critical for tasks involving semantic comprehension and emotion.

This concatenated feature captures both low-level acoustic characteristics and high-level semantic information, providing a rich representation of the input speech signal. Once the combined feature $(3, 1024)$ is obtained, it is passed through a 1D Convolutional Neural Network (1D-CNN) for further processing. The input tensor is reshaped to $(B \times T, 3, 1024)$, where B is the batch size and T is the number of time frames.

The 1D-CNN encoder consists of multiple stacked convolutional layers with residual connections to enhance feature extraction and maintain stable gradients. As the network deepens, the temporal resolution is reduced, while the feature channels are increased, capturing increasingly abstract representations.

3.1.2 Pose and Reference Landmark Encoder

The Pose and Reference Landmark Encoder utilizes the same architecture as the Landmark Encoder in our previously published SpeechSyncNet (Cao et al., 2023), which is designed to capture both spatial and temporal features effectively. For the pose input, we use a tensor with the shape $(B, T_p, 2, 74)$, where B is the batch size, T_p is the number of pose frames, and 74 represents the key landmark points corresponding to the upper half of the face. These pose landmarks play a crucial role in guiding the prediction process by providing essential orientation cues, ensuring that the model can generate coherent and natural facial movements.

Similarly, for the reference input, the encoder processes a tensor with the shape $(B, T_r, 2, 131)$, where T_r denotes the number of reference frames, and 131 represents the full set of facial landmark points. The reference landmarks serve as a foundation for refining predictions and maintaining consistency with the target facial expressions and geometry.

The Landmark Encoder in SpeechSyncNet is specifically designed to capture both spatial relationships between landmarks and temporal dependencies across frames. By leveraging global average pooling and 1D convolutions, the encoder ensures that both local and global features are preserved. This dual capability allows the encoder to effectively process dynamic facial expressions and subtle pose variations, enabling high-quality and temporally coherent landmark predictions. The integration of both pose and reference landmarks ensures that the model not only maintains spatial accuracy but also produces smooth and contextually appropriate facial movements over time.

3.1.3 Landmark Decoder

The Landmark Decoder generates accurate facial landmark predictions by incorporating audio, pose, and reference landmark information. Position embeddings are applied to both the audio and pose features to capture temporal and spatial relationships, ensuring the model recognizes frame order.

These combined features are then passed through a Transformer Encoder, which uses multi-head self-attention and feed-forward layers to capture long-range dependencies. This allows the model to integrate information from both audio and pose modalities effectively, ensuring coherent landmark trajectories across multiple frames.

The final output of the Transformer Encoder is processed through a linear projection layer to produce the output shape $(B, T_p, 2, 57)$, where B is the batch size, T_p is the number of pose frames, 2 represents the x and y coordinates, and 57 corresponds to the predicted landmarks for the lower half of the face, ensuring accurate and expressive facial movements.

3.2 Landmark to Face

The input to the pretrained VAE Encoder comprises four elements: the predicted landmarks I^t_{lm}, the target image I^t_{img}, the reference landmarks I^r_{lm}, and the reference image I^r_{img}. Each input is processed independently, producing a latent feature with a shape of $(4, 32, 32)$. Notably, both the predicted and reference landmarks are first rendered as images before being input into the VAE, and the target image I^t_{img} is masked in the lower half to obscure the mouth region. These latent features are subsequently passed through SPADE-Net, which predicts the final latent representation $\mathcal{F}^p_{img} \in \mathbf{R}^{4\times32\times32}$ for the synthesized image. Finally, the pretrained VAE Decoder uses this predicted latent feature to reconstruct the facial image I^p_{img}, ensuring consistency between the predicted landmarks and the generated face.

3.2.1 Pretrained Variational Autoencoder

The model utilizes a pretrained Variational Autoencoder (VAE) (Kingma and Welling, 2019), specifically the AutoencoderKL architecture (Kingma and Welling, 2022), which incorporates KL divergence loss as a core component. Fine-tuned for high-quality facial image generation, this VAE uses a probabilistic encoder to transform input images into a latent space that captures both global and fine-grained features. The input to the encoder is a tensor with shape $(3, 256, 256)$, representing the facial image in RGB format. The latent space, shaped $(4, 32, 32)$, encapsulates the essential characteristics of the input image. Once encoded, the compressed latent vector is passed through the decoder, which reconstructs the image into an output tensor with the same shape $(3, 256, 256)$, corresponding to the generated facial image.

$$\mathcal{F}^t_{img}, \mathcal{F}^t_{lm}, \mathcal{F}^r_{img}\mathcal{F}^r_{lm}, = \mathcal{E}(\mathcal{M}(I^t_{img}), I^t_{lm}, I^r_{img}I^r_{lm}) \quad (1)$$

where $\mathcal{E}$ represents the VAE Encoder and $\mathcal{M}$ performs the masking operation.

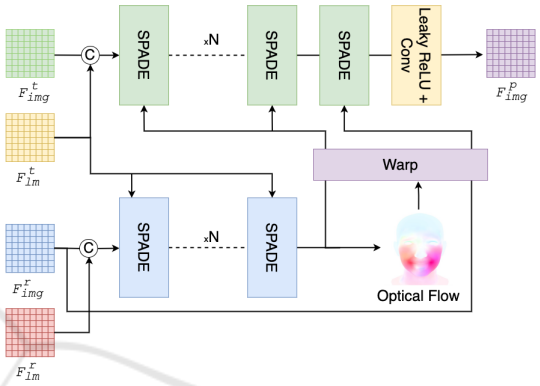

Figure 2: The architecture of the SPADE-Net.

3.2.2 SPADE-Net

Inspired by the IP_LAP framework (Zhong et al., 2023), SPADE-Net is designed with two primary modules, each composed of a series of consecutive SPADE (Spatially Adaptive Denormalization) layers (Park et al., 2019). These modules work together to predict the target image's features in the latent space. The architecture of SPADE-Net is illustrated in Figure 2, highlighting how the sequential layers enable precise feature prediction for image generation.

In the first module $\mathcal{G}_1$, the reference image feature $\mathcal{F}^r_{img}$ and the reference landmark feature $\mathcal{F}^r_{lm}$ are passed through N x SPADE layers. The target landmarks are used as conditioning inputs, enabling the model to learn the optical flow $\mathcal{F}_o \in \mathbf{R}^{2\times32\times32}$ between the reference and target facial configurations. By leveraging the target landmarks, SPADE-Net effectively captures the spatial transformation needed to align the reference and target faces, modeling facial motion and structure accurately.

The output of the SPADE layer sequence and the reference image feature are then warped using the learned optical flow, producing the warping SPADE feature $\mathcal{F}^s_{warp}$ and the warping reference feature $\mathcal{F}^r_{warp}$. These warped features serve as inputs to the second module.

$$\mathcal{F}^s_{warp}, \mathcal{F}^r_{warp} = \mathcal{G}_1(\mathcal{F}^r_{img}, \mathcal{F}^r_{lm}, \mathcal{F}^t_{lm}) \quad (2)$$

In the second module $\mathcal{G}_2$, the target mask image feature $\mathcal{F}^t_{img}$ and the target landmark feature $\mathcal{F}^t_{lm}$ are passed through N x SPADE layers. The warped feature $\mathcal{F}^s_{warp}$ derived from the first module is used as an additional conditioning input for the SPADE layers, guiding the generation of the target face image.

A critical aspect of this architecture is the warping of the reference image's latent features using the predicted optical flow. The warped latent feature $\mathcal{F}^r_{warp}$ is then used as a condition for the final SPADE layer, ensuring that essential characteristics of the reference face, such as its overall structure and facial features, are preserved in the generated output. This warping process ensures consistency and realism in the final synthesized facial image, even as the face undergoes transformations driven by the target landmarks.

$$\mathcal{F}^p_{img} = \mathcal{G}_2(\mathcal{F}^t_{img}, \mathcal{F}^t_{lm}, \mathcal{F}^s_{warp}, \mathcal{F}^r_{warp}) \tag{3}$$

$$I^p_{img} = VAE\_Decoder(\mathcal{F}^p_{img}) \tag{4}$$

4 EXPERIMENTS

4.1 Datasets

We evaluate the model using the MEAD dataset (Wang et al., 2020b), a benchmark for talking face generation. MEAD features 281,400 high-resolution video clips (1920×1080) from 60 actors, annotated with diverse emotional expressions, making it ideal for testing emotion recognition and synthesis.

4.2 Loss Function

The audio-to-landmark module is trained using two loss functions: Mean Squared Error (MSE) loss and Landmark Velocity loss. MSE loss $\mathcal{L}_{\text{mse}}$ minimizes the difference between predicted and ground truth landmark positions, ensuring spatial accuracy. Landmark Velocity loss $\mathcal{L}_{\text{v}}$ promotes temporal consistency by capturing smooth transitions between successive landmark predictions.

$$\mathcal{L}_{\text{mse}} = \frac{1}{F}\sum_{i=1}^{F} \|\hat{\text{y}}_i - \text{y}_i\|_2^2 \tag{5}$$

$$\mathcal{L}_{\text{v}} = \frac{1}{F-1}\sum_{f=1}^{F-1}\frac{1}{N}\sum_{i=1}^{N} \left\| (\hat{\text{y}}_{f+1,i} - \hat{\text{y}}_{f,i}) - (\text{y}^{\text{gt}}_{f+1,i} - \text{y}^{\text{gt}}_{f,i}) \right\|_2^2 \tag{6}$$

$$\mathcal{L}_{\text{a2lm}} = \mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{v}} \tag{7}$$

where: $\mathbf{y}_i$, $\hat{\mathbf{y}}_i$ denote the ground truth and predicted landmarks for the i-th sample, respectively. F denotes the number of frames and N the number of points.

In the landmark-to-face generation stage, we use multiple loss functions to enhance the quality of generated faces. Reconstruction loss $\mathcal{L}_r$ minimizes the MSE between the generated and ground truth faces to ensure structural similarity. Perceptual loss $\mathcal{L}_p$, computed in the latent space of a pretrained network, captures high-level semantic details for perceptual realism. GAN loss $\mathcal{L}_{gan}$ encourages realistic-looking faces by differentiating real and generated images. Additionally, a mouth-specific loss $\mathcal{L}_{mouth}$ focuses on the MSE in the mouth region, improving accuracy in reconstructing occluded or degraded mouth features. Together, these losses ensure high visual fidelity in talking face generation.

$$\mathcal{L}_{\text{lm2f}} = \lambda_r \mathcal{L}_{\text{r}} + \lambda_p \mathcal{L}_{\text{p}} + \lambda_g \mathcal{L}_{\text{gan}} + \lambda_m \mathcal{L}_{\text{mouth}} \tag{8}$$

where $\lambda_r, \lambda_p, \lambda_g, \lambda_m$ are the weighting factors for the reconstruction loss, perceptual loss, GAN loss, and attention loss, respectively.

4.3 Experimental Setup

The MEAD dataset was split into training, validation, and testing sets at an 80%-10%-10% ratio. The Audio-to-Landmark model was trained for 200 epochs, and the Landmark-to-Face model for 300 epochs, with video data processed at 25 FPS. Training used a batch size of 32, a learning rate of 1.0×10^{-4}, the Adam optimizer, and the ExponentialLR scheduler for adaptive learning rate adjustment. Loss weights were set as $\lambda_r = \lambda_g = 1.0$, $\lambda_p = 4.0$, and $\lambda_m = 2.5$.

5 RESULTS

5.1 Experimental Metrics

The Landmark Distance (LD) metric, introduced by (Chen et al., 2018), measures the average Euclidean distance between predicted and reference landmarks, normalized by face width. For our task, we calculate LD using only landmarks in the lower half of the face, excluding the upper region, to focus on this specific area. The equation is represented as follows:

$$\text{LD} = \frac{1}{F}\sum_{f=1}^{F}\frac{1}{N}\sum_{i=1}^{N}\frac{\|\text{l}^{\text{pred}}_{f,i} - \text{l}^{\text{ref}}_{f,i}\|_2}{w_{\text{face}}} \tag{9}$$

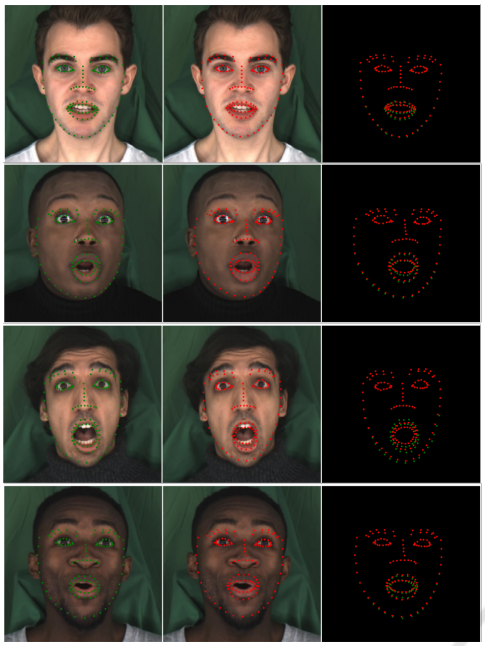

Figure 3: Visualization of groundtruth landmarks (green) and predicted landmarks (red).

where F denotes the number of frames, N denotes the number of points, and w_{face} is the width of the face.

The Landmark Velocity Difference (LVD) metric quantifies the dynamics of landmark motion by calculating the difference in landmark positions between consecutive frames. The LVD is defined using the same equation as in Equation 6.

Furthermore, to evaluate the quality of the generated images, we employed three key metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM)(Larkin, 2015), and Fréchet Inception Distance (FID)(Heusel et al., 2018).

5.2 Performance Comparision

We evaluate our methods against several approaches in talking face generation on the MEAD dataset to ensure a fair comparison. The benchmark methods on the MEAD dataset include Chen et al. (Chen et al., 2018), MEAD (Wang et al., 2020b), EVP (Ji et al., 2021), Song et al. (Song et al., 2022), SpeechSyncNet (Cao et al., 2023), Trans APL (Cao et al., 2024b), EAPC (Cao et al., 2024a), Sinha et. al. (Sinha et al., 2022), and IP_LAP (Zhong et al., 2023)

5.3 Quatitative

Table 1 compares model performance on the MEAD dataset using Landmark Distance (LD) and Landmark Velocity Distance (LVD). Our model achieves the best LD score of 1.84, significantly outperforming Chen et al. (3.27) and Trans APL (3.58). Although SpeechSyncNet has a slightly better LVD score (0.97 vs. 1.28), our model's superior LD highlights its accuracy in landmark prediction, making it highly effective for precision-critical real-time applications.

Table 1: Comparison of landmark prediction performance across different models on the MEAD dataset.

Model (MEAD dataset)	LD↓	LVD↓
Chen et al. (Chen et al., 2018)	3.27	2.09
MEAD (Wang et al., 2020b)	2.52	2.28
EVP (Ji et al., 2021)	2.45	1.78
Song et al. (Song et al., 2022)	2.54	1.99
EAPC (Cao et al., 2024a)	2.01	1.25
Trans APL (Cao et al., 2024b)	3.58	1.03
SpeechSyncNet (Cao et al., 2023)	1.92	**0.97**
Our	**1.84**	1.28

Table 2: Performance of different models on the MEAD dataset based on PSNR, SSIM, and FID.

Model (MEAD dataset)	PSNR↑	SSIM↑	FID↓
MEAD	28.61	0.68	25.52
EVP	29.53	0.71	**7.99**
Sinha et. al.	30.06	0.77	35.41
IP_LAP	**32.91**	**0.93**	27.87
Our	30.12	0.90	31.36

Table 2 compares our model with others on the MEAD dataset using PSNR, SSIM, and FID metrics. Our model achieves a PSNR of 30.12 and SSIM of 0.90, ranking second to the IP_LAP model (PSNR: 32.91, SSIM: 0.93). While our FID score of 31.36 is higher than EVP's 7.99, indicating less realism, our model balances structural similarity and image quality effectively, showcasing its ability to generate facial images that preserve key features with competitive performance.

5.4 Qualitative Visualization

Figure 3 shows that the mouth shape generated from predicted landmarks in our method closely matches the ground truth. The smooth landmark flow accurately captures natural head movements, demonstrating strong temporal coherence in the sequences. This highlights the effectiveness of our approach in maintaining landmark consistency over time.

Additionally, Figure 4 visually demonstrates how using driven landmarks and reference facial images enables effective inpainting of occluded facial areas, even within a compact latent space. This allows for accurate reconstruction of facial features, yielding re-

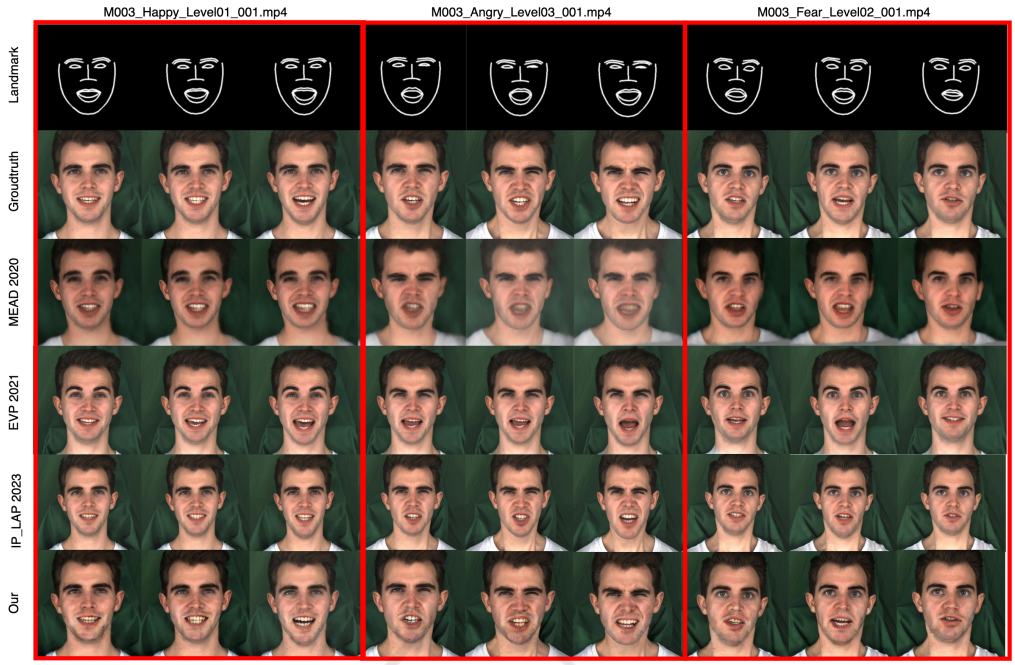

Figure 4: Visual quality comparison between generated faces and groundtruth facial images.

alistic lip sync, head movements, and emotional expressions. While the inpainting quality is high, slight variations around the mouth region suggest areas for further refinement.

5.5 Ablation Study

To evaluate HuBERT's effectiveness, experiments were conducted with four models: HuBERT, MFCC, Wav2Vec Conformer, and UniSpeech. These models represent different audio feature extraction techniques, each with a distinct approach. The evaluation focused on their ability to predict landmarks and generate facial images.

Table 3: Comparison of LD and LVD performance across different models.

Model	LD↓	LVD↓
HuBERT	**1.8394**	**1.2813**
MFCC	1.9684	1.5855
Wav2Vec Conformer	2.1627	1.5911
UniSpeech	2.3756	1.8629

Table 3 shows that HuBERT outperforms all other models in both landmark distance (LD) and landmark velocity distance (LVD), achieving scores of 1.8394 and 1.2813, respectively. In comparison, MFCC performs slightly worse with LD and LVD values of 1.9684 and 1.5855, respectively. Wav2Vec Conformer and UniSpeech exhibit even higher values, with Wav2Vec Conformer at 2.1627 (LD) and 1.5911 (LVD), and UniSpeech at 2.3756 (LD) and 1.8629 (LVD). These results emphasize HuBERT as the most effective model for this task. Additionally, Figure 5 illustrates experimental results across multiple faces, demonstrating the method's generalizability.

6 CONCLUSION

In conclusion, this work highlights the effectiveness of HuBERT for landmark prediction, outperforming models like MFCC, Wav2Vec Conformer, and UniSpeech in both landmark distance and velocity. We also introduce SPADE-Net, a novel architecture that enhances facial landmark prediction and image generation, providing a robust solution for precise face synthesis. The integration of HuBERT and SPADE-Net marks a significant advancement in improving the reliability and realism of facial reconstruction from audio-driven landmarks.

ACKNOWLEDGEMENTS

This research is supported by research funding from Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City.

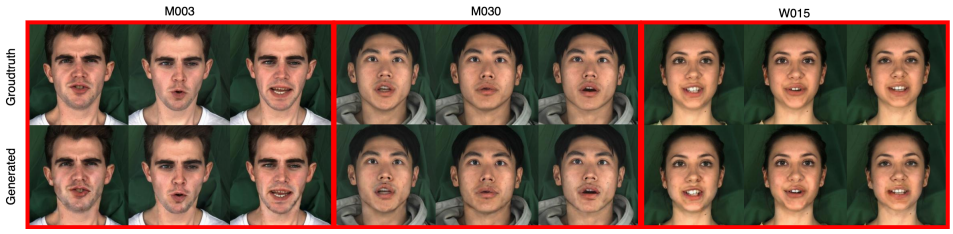

Figure 5: Visualization of face construction in difference persons.

REFERENCES

Abdul, Z. K. and Al-Talabani, A. K. (2022). Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, 10:122136–122158.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations.

Cao, X.-N., Trinh, Q.-H., Do-Nguyen, Q.-A., Ho, V.-S., Dang, H.-T., and Tran, M.-T. (2024a). Eapc: Emotion and audio prior control framework for the emotional and temporal talking face generation. In *ICAART (2)*, pages 520–530.

Cao, X.-N., Trinh, Q.-H., Ho, V.-S., and Tran, M.-T. (2023). Speechsyncnet: Speech to talking landmark via the fusion of prior frame landmark and the audio. In *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE.

Cao, X.-N., Trinh, Q.-H., and Tran, M.-T. (2024b). Transapl: Transformer model for audio and prior landmark fusion for talking landmark generation. In *ICMV*.

Chen, L., Li, Z., Maddox, R. K., Duan, Z., and Xu, C. (2018). Lip movements generation at a glance. In *Proceedings of the European conference on computer vision (ECCV)*, pages 520–535.

Eskimez, S. E., Zhang, Y., and Duan, Z. (2021). Speech driven talking face generation from a single image and an emotion condition.

Gong, Y., Chung, Y.-A., and Glass, J. (2021). Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2018). Gans trained by a two time-scale update rule converge to a local nash equilibrium.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units.

Ji, X., Zhou, H., Wang, K., Wu, W., Loy, C. C., Cao, X., and Xu, F. (2021). Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089.

Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *CoRR*, abs/1906.02691.

Kingma, D. P. and Welling, M. (2022). Auto-encoding variational bayes.

Larkin, K. G. (2015). Structural similarity index ssimplified: Is there really a simpler concept at the heart of image quality measurement?

Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization.

Sinha, S., Biswas, S., Yadav, R., and Bhowmick, B. (2022). Emotion-controllable generalized talking face generation.

Song, L., Wu, W., Qian, C., He, R., and Loy, C. C. (2022). Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security*, 17:585–598.

Verma, P. and Berger, J. (2021). Audio transformers:transformer architectures for large scale audio understanding. adieu convolutions.

Vougioukas, K., Petridis, S., and Pantic, M. (2020). Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128:1398–1413.

Wang, C., Wu, Y., Qian, Y., Kumatani, K., Liu, S., Wei, F., Zeng, M., and Huang, X. (2021). Unispeech: Unified speech representation learning with labeled and unlabeled data.

Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., and Loy, C. C. (2020a). Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*.

Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., and Loy, C. C. (2020b). Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*.

Zhong, W., Fang, C., Cai, Y., Wei, P., Zhao, G., Lin, L., and Li, G. (2023). Identity-preserving talking face generation with landmark and appearance priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

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