Occlusion-Insensitive Talking Head Video Generation via Facelet Compensation

Yuhui Deng [1,2*], Yuqin Lu [1,2*], Yangyang Xu [3], Yongwei Nie [1†], Shengfeng He [2†]

[1] South China University of Technology
[2] Singapore Management University
[3] Harbin Institute of Technology (Shenzhen)

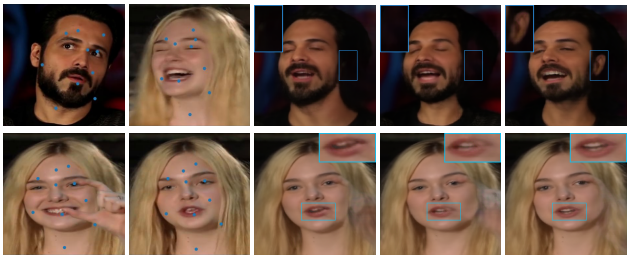
Abstract

Talking head video generation involves animating a still face image using facial motion cues derived from a driving video to replicate target poses and expressions. Traditional methods often rely on the assumption that the relative positions of facial keypoints remain unchanged. However, this assumption fails when keypoints are occluded or when the head is in a profile pose, leading to inconsistencies in identity and blurring in certain facial regions. In this paper, we introduce Occlusion-Insensitive Talking Head Video Generation, a novel approach that eliminates the reliance on spatial correlation of keypoints and instead leverages semantic correlation. Our method transforms facial features into a facelet semantic bank, where each facelet token represents a specific facial semantic. This bank is devoid of spatial information, allowing it to compensate for any invisible or occluded face regions during motion warping. The facelet compensation module then populates the facelet tokens within the initially warped features by learning a correlation matrix between facial semantics and the facelet bank. This approach enables precise compensation for occlusions and pose changes, enhancing the fidelity of the generated videos. Extensive experiments demonstrate that our method achieves state-of-the-art results, preserving source identity, maintaining fine-grained facial details, and capturing nuanced facial expressions with remarkable accuracy.

Introduction

Talking head video generation aims to synthesize human face videos by seamlessly merging the appearance and identity of a source face image with the vivid facial expressions and intricate head movements of a target face. This task is challenging due to the limited appearance information available from the static source image and the considerable pose and facial expression gaps observed between the source and target faces. Despite these challenges, it holds significant potential for applications in industries such as film production, virtual reality, and teleconferencing.

The advent of Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) has led to remarkable advancements in this task, enhancing both quality and robustness. Mainstream approaches (Siarohin et al. 2019b,a, 2021) predominantly follow a self-supervised learning pipeline, aiming to learn accurate motion estimation and expressive representation. These methods often leverage sparse facial keypoints to autonomously learn and characterize a comprehensive global motion flow between the source image and each frame within the driving video sequence.



(a) Source (b) Target (c) MCNet (d) MRFA (e) Ours

Figure 1: Existing methods (c) and (d) assume that each facial element in the source has a matching counterpart in the target, which fails when self-occlusion occurs. We overcome this limitation by introducing a facelet semantic bank that compensates for occluded regions using facelet tokens.

Due to its ill-posed nature, recent approaches have begun incorporating additional information, such as 3D facial prior models (e.g., 3DMM (Blanz and Vetter 1999)) with decoupled expression codes (Ren et al. 2021; Zakharov et al. 2019), depth facial maps (Hong et al. 2022), 3D learned landmarks (Wang, Mallya, and Liu 2021; Gao et al. 2023), or semantic segmentation (Chen et al. 2020) to facilitate the generation of complex facial expressions and head movements. Additionally, some efforts focus on refining motion models for more accurate transformations. For instance, Tao et al. (Tao et al. 2022a, 2024) refined the learning process of global motion flow, while Zhao et al. (Zhao and Zhang 2022) proposed a thin-plate spline motion estimation method to produce a more flexible optical flow.

However, despite precise motion estimation, these global warping-based methods assume that all facial elements can be well-matched between the target and source faces. Consequently, globally warped features often fail to provide adequate facial information where self-occlusion occurs (see Fig.1c and Fig.1d). This results in generated face videos that

*These authors contributed equally.
†Corresponding authors: Yongwei Nie and Shengfeng He.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

exhibit issues such as blurry details, inconsistent identity, and noticeable distortions in certain areas.

To address these challenges, we introduce Occlusion-Insensitive Talking Head Video Generation. The rationale behind our method is that spatially structured information is inevitably and inherently corrupted by any local faults or occlusions. Therefore, we eliminate the reliance on spatial correlation of keypoints/flows and instead leverage semantic correlation. Our method transforms facial features into a facelet semantic bank, where each facelet token represents a specific facial semantic. This bank, devoid of spatial information, adeptly compensates for any invisible or occluded face regions during motion warping. The facelet compensation module then integrates the facelet tokens within the initially warped features by learning a correlation matrix between facial semantics and the facelet bank, ensuring precise compensation for occlusions and pose changes. Additionally, we incorporate a spatially-invariant pose code throughout the generation process, which provides crucial hints for handling extreme head poses. Thus, our solution remains insensitive to any source head poses and occlusion (see Fig. 1e), generating talking head videos with consistent and artifact-free facial appearances.

We conduct extensive experiments on competitive face video benchmarks such as VoxCeleb1 (Nagrani, Chung, and Zisserman 2017) and CelebV (Wu et al. 2018). Experimental results demonstrate the effectiveness of our approach in addressing extreme poses and occlusion. Furthermore, our method significantly outperforms state-of-the-art techniques, as evidenced by both qualitative and quantitative evaluations.

Related Work

Image Animation involves generating a realistic image where the content is driven by a source image and influenced by a target video. Image animation can be categorized into subject-dependent and subject-agnostic approaches.

Subject-dependent methods are trained on a specific subject and can only animate that particular individual (Bansal et al. 2018; Suwajanakorn, Seitz, and Kemelmacher-Shlizerman 2017; Thies, Zollhöfer, and Nießner 2019; Wu et al. 2018; Vlasic et al. 2006; Thies et al. 2016). For instance, Face2Face (Thies et al. 2016) tracks facial expressions in both the source and target videos through a dense photometric consistency measure, employing fast and efficient deformation transfer between the source and target faces. With the advancement of neural radiance fields (NeRF) (Mildenhall et al. 2021), methods like AD-NeRF (Guo et al. 2021) and NVP (Thies et al. 2020) have utilized NeRF for generating talking head videos driven by audio input.

In contrast, subject-agnostic methods are less complex and more general, making them applicable to a wider range of subjects (Siarohin et al. 2018, 2019a,b; Wang et al. 2022; Tao et al. 2022a). Monkey-Net (Siarohin et al. 2019a) was the first to use sparse learned keypoints for estimating optical flow, enabling the animation of various objects. FOMM (Siarohin et al. 2019b) extends Monkey-Net by incorporating local affine transformation. MRAA (Siarohin et al. 2021) proposed region-based affine motion representations, while MTIA (Tao et al. 2022a) and DAM (Tao et al. 2022b) constrained the affine motion model with stronger priors. LIA (Wang et al. 2022) introduced a set of learned direction vectors as motion representations, animating images by linear navigation in the latent space. However, these methods typically assume a locally rigid object and are primarily applied to human bodies, resulting in limited performance when specifically applied to human faces.

Talking Head Video Generation achieved significant developments in recent years, thanks to the advent of large face video datasets. This field is typically divided into two strategies: image-driven and audio-driven generation.

In image-driven methods, researchers strive to capture the expression of a given driving image and combine it with the facial identity from a specified source image. Keypoint-based methods (Siarohin et al. 2018, 2019a, 2021, 2019b; Tao et al. 2022a) learn a facial motion model to characterize the motion transformations between two sets of keypoints. Several methods (Ren et al. 2021; Doukas, Zafeiriou, and Sharmanska 2021; Li et al. 2017) leverage 3D information, such as 3DMM and 3D mesh, to extract expression and identity codes from a given face. For example, SAFA (Wang, Zhang, and Li 2021) and HeadGAN (Doukas, Zafeiriou, and Sharmanska 2021) use a 3D morphable model for expression transfer.

Some warping-based approaches (Tao et al. 2022a; Zhao and Zhang 2022; Tao et al. 2024) focus on identifying improved local keypoints and corresponding affine representations or models for motion transformation parameters. TPSM (Zhao and Zhang 2022) suggests employing local thin-plate-spline motions as an improvement over affine motions, MRFA (Tao et al. 2024) introduces a motion refinement module to compensate for affine motions, and FNeVR (Zeng et al. 2022) introduces a 3D face volume rendering module to enhance affine motions for image rendering. Additionally, DaGAN and DaGAN++ (Hong et al. 2022; Hong, Shen, and Xu 2023) present a depth estimator to leverage face depth information for generation, while MCNet (Hong and Xu 2023) proposes a global facial refinement network to compensate for warped source facial features.

Audio-driven talking head generation (Chen et al. 2019; Ji et al. 2022; Prajwal et al. 2020) primarily focuses on producing lip motion from the audio sequence. This task is relatively easier due to the natural disentanglement of the driving audio and the source content. MakeitTalk (Zhou et al. 2020) and ATVGnet (Chen et al. 2019) disentangle content and identity information from audio sequences to generate talking faces. A recent method (Zhong et al. 2023) employs a two-stage framework consisting of audio-to-landmark and landmark-to-video rendering procedures.

In this work, our focus lies on warping-based image-driven talking head generation. Unlike previous approaches, we aim to resolve the problem of occlusion sensitivity in existing methods, resulting in higher-quality generation with enhanced local detail and robustness to large head pose shifts.

Method

The proposed framework comprises three principal components: a motion estimator conditioned on a spatially-invariant pose code, a facelet semantic bank module, and a facelet compensation network. An overview of our approach is depicted

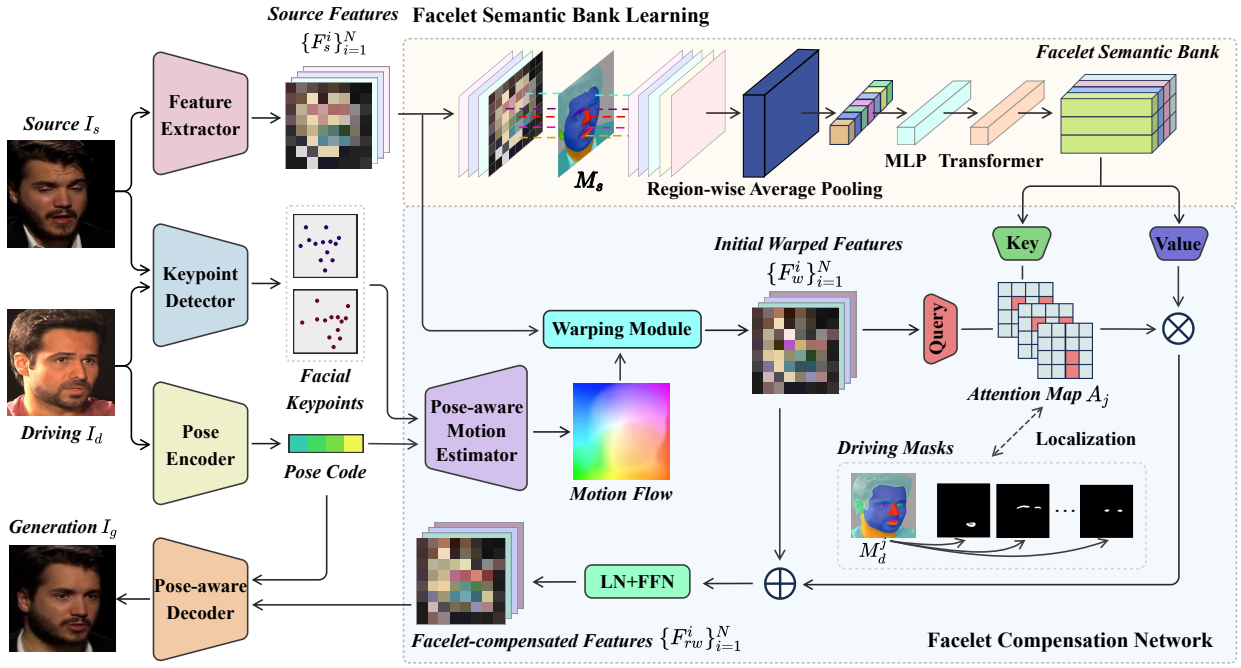


Figure 2: Overview of our pipeline. Our approach generates a facelet semantic bank and leverages its spatial invariance to compensate for occlusions that occur during the generation process.

in Fig. 2. Initially, we estimate the motion flow between the source and driving face, conditioned on the target facial pose and the keypoints of the source and driving frames. Leveraging the estimated motion flow, we warp the encoded features in each layer, producing the initial warped features, which might be corrupted by occlusion. Simultaneously, we learn a facelet semantic bank from the multi-scale source features, capturing different facial semantics. The initially warped features are then refined through query-based compensation from the facelet semantic bank at each resolution within the facelet compensation network. Finally, our decoder synthesizes the final animation frame by frame using all the refined feature maps. In the following paragraphs, we elaborate on each of these components.

Spatially-Invariant Pose Code

We first follow existing methods (Siarohin et al. 2018; Xu et al. 2022, 2020; Siarohin et al. 2019b,a; Zhao and Zhang 2022) by utilizing a keypoint estimator (Newell, Yang, and Deng 2016) to predict a set of sparse facial keypoints $\{k_{l,n}\}_{n=1}^{K}$ for both the source and target faces, where $l \in \{s, d\}$, $k \in \mathbb{R}^{2\times K}$, and K is the total number of keypoints. Alongside the keypoints, we predict their affine matrices $\{A_{l,n}\}_{n=1}^{K}$, where $A \in \mathbb{R}^{2\times 2\times K}$. The local motion flow $\mathcal{F}_n$ is then derived using the following equation:

$$\mathcal{F}_n(z) = k_{s,n} + A_{s,n}(A_{d,n})^{-1}(z - k_{d,n}), \tag{1}$$

where $z \in \mathbb{R}^2$ indicates a location in the image. In addition to the K part local motion flows, an extra motion field is introduced to preserve static background information:

$$\mathcal{F}_0(z) = z. \tag{2}$$

Given that local motion flows may not adequately capture global motion, especially in challenging cases where the source face is partially self-occluded (e.g., a profile face with a large angle), we introduce a spatially-invariant pose code to better handle these scenarios. To achieve this, we encode the target spatially-invariant pose code p_t using an encoder E_p, which is a pre-trained 3D shape model (Feng et al. 2021) designed to extract pose information while being invariant to spatial details of the facial appearance. The spatially-invariant pose code captures the global orientation and pose of the target face without being influenced by the specific texture or appearance of the face. By isolating the pose information, the pose code enables our model to accurately understand and replicate the head's orientation and movement even when parts of the face are occluded. This ensures that the generated motion flow can correctly guide the facial animation, maintaining the correct pose even under challenging conditions.

Using the spatially-invariant pose code and facial keypoints, we adopt a facial motion mask estimator to capture global and local head motion information between the source and target faces. Using the predicted motion mask $M_n \in \mathbb{R}^{H\times W\times(K+1)}$, the final motion flow between the driving and source images is derived as follows:

$$\mathcal{F}(z) = \textstyle\sum_{n=0}^{K} M_n \mathcal{F}_n(z). \tag{3}$$

As shown in Fig. 2, we then utilize the final motion flow $\mathcal{F}$ to warp the multi-scale appearance features F_s^i, where i is the i_{th} scale feature, to produce an initial warped feature map. An occlusion map O_i predicted by the motion mask estimator is used to mask out regions of the initial warped feature map that require inpainting due to variations in head rotations.

This process yields the warped source image feature F_w^i as follows:

$$F_w^i = O_i \times \mathtt{W}(F_s^i, \mathcal{F}), \tag{4}$$

where $\mathtt{W}[\cdot,\cdot]$ is the warping operation. Through this approach, the warped features $\{F_w^i\}_{i=1}^N$, with N representing the number of feature scales, more effectively retain the identity of the source image while ensuring consistency in facial pose and motion between the source and target faces.

Facelet Semantic Bank

Despite accurate motion estimation, the final generated face videos often exhibit blurred local details and inconsistency between facial areas. This issue arises from the limitations of the global warped representation, particularly in scenarios where the source face is partially self-occluded or displays delicate expression variations. To address this, our initial goal is to construct a facelet semantic bank to ensure consistency under occlusion.

As illustrated in the top-right of Fig. 2, given an image I_s, we employ an off-the-shelf face parsing model (Yu et al. 2018) to extract a segmentation mask M_s that encompasses regions such as the eyes, ears, lips, and brows. For each scale of feature F_s^i, we resize the facial mask M to the same spatial size and apply the region-wise average pooling operation ($\mathcal{RAP}$) on F_s^i. We then aggregate and average each region's features and feed them into an MLP network ($\mathcal{MLP}$) to embed an initial facelet bank $F_{fb} \in \mathbb{R}^{N\times C\times 512}$ of the source image:

$$M_s^j = (\lfloor M_s \rfloor == j), \forall j \in \{1, 2, \ldots, C\}, \tag{5}$$

$$F_{fb}^i = \mathcal{MLP}(\mathtt{Concat}(\mathcal{RAP}(F_s^i \odot (M_s^j)))), \tag{6}$$

where C is the number of regions, $\odot$ is the Hadamard product, $\mathtt{Concat}[\cdot,\cdot]$ denotes the concatenation operation, and $\lfloor M_s \rfloor$ denotes the resized segmentation mask with the same spatial size as F_s^i.

Due to the lack of interaction among the projected facial local features, simply using the initial facelet semantic bank cannot effectively preserve accuracy in appearance fidelity. To address this, we feed the initial facelet bank into a transformer layer. In this layer, F_{fb}^i is linearly projected to the query, key, and value features: $Q_{F_{fb}^i}$, $K_{F_{fb}^i}$, and $V_{F_{fb}^i}$. For clarity, we temporarily omit the subscript and define Multi-head Self-Attention (MSA) as follows:

$$\text{head}^i = \mathtt{Softmax}\left(\frac{QW_Q^i(KW_K^i)^T}{\sqrt{d_k}}\right)VW_V^i, \tag{7}$$

$$\mathtt{MSA}(Q,K,V) = \mathtt{Concat}\left(\text{head}^1, \ldots, \text{head}^h\right)W_O, \tag{8}$$

$$\hat{F}_{fb}^i = F_{fb}^i + \mathtt{FFN}\left(\mathtt{LN}\left(\mathtt{MSA}\left(F_{fb}^i\right)\right)\right), \tag{9}$$

where W_O, W_Q^i, W_K^i, and W_V^i represent learned parameters used for feature projections. d_k denotes the hidden dimension of the projection subspace, while h signifies the total number of heads in each MSA layer. FFN and LN refer to the feed-forward network, comprised of linear transformations, ReLU activation, and layer normalization. Thus, we obtain the ultimate enriched facelet semantic bank, denoted as $\hat{F}_{fb}^i$, endowed with the capability for identity interaction, ensuring better preservation of appearance fidelity and consistency.

Facelet Semantic Compensation

The obtained facelet semantic bank is spatially invariant and consists solely of semantic information, making it ideal for recovering occluded regions. For example, an occluded right eye can be reconstructed using the facelet token for the left eye. Specifically, we use the facelet semantic bank to produce the key and value features, $K_{\hat{F}_{fb}^i}$ and $V_{\hat{F}_{fb}^i}$ respectively, through two dynamic layers. This approach ensures that the key and value features encapsulate fine-grained facial details and are tailored for occlusion recovery.

First, we flatten and project F_w^i as a query feature $Q_{F_w^i}$. The attention map is then computed as:

$$A = \mathtt{Softmax}\left(\frac{Q_{F_w^i}(K_{\hat{F}_{fb}^i})^T}{\sqrt{d_k}}\right), \tag{10}$$

where A denotes the attention map, indicating the relevance of the facelet token to each pixel. Finally, the refined warped feature maps F_{rw}^i are obtained as follows:

$$F_{rw}^i = F_w^i + AV_{\hat{F}_{fb}^i}. \tag{11}$$

Ideally, the attention maps of each facial region token should focus solely on their respective facial feature areas, avoiding the blending of irrelevant facial features and preventing their spread across the entire facial image. Motivated by (Chefer et al. 2023), we localize the cross-attention map by utilizing the target segmentation mask. Let M_d^j represent the segmentation mask of the target face image in the j-th facial region, and A_j denote the cross-attention map of the j-th facelet token. We enforce the cross-attention map A_j to closely match the segmentation mask M_d^j.

We employ a localization loss $\mathcal{L}_{loc}$ to minimize the distance between the cross-attention map and the segmentation mask, formulated as a balanced $\mathcal{L}_1$ loss:

$$\mathcal{L}_{loc} = \frac{1}{C}\sum_{j=1}^{C}\left(\mu(A_j\lfloor\bar{M}_d^j\rfloor) - \mu(A_j\lfloor M_d^j\rfloor)\right), \tag{12}$$

where $\mu(\cdot)$ denotes the mean operation, $\bar{M}_d^j = 1 - M_d^j$, and C denotes the number of target face regions. As a result, our facelet compensation module effectively addresses the issue of inadequate preservation during occlusion, ensuring high fidelity and consistency in the generated facial features.

Talking Head Generation

We have now obtained the warped and compensated feature maps $\{F_{rw}^i\}_{i=1}^N$. To further enhance the generation of facial details, our approach extends facelet compensation across multiple layers. To improve the ability to generate content that does not exist in the source, we introduce a facial pose-aware decoder. The process begins by treating F_{rw}^1 as F_g^1. Subsequently, F_g^2 is generated from F_g^1 through an upsample convolution block and an AdaIN block, conditioned on the target spatially-invariant pose code p_t. At each level, the compensated feature map F_{rw}^i is concatenated with the generated feature F_g^i to produce the next level feature F_g^{i+1}. Finally, the last layer applies a sigmoid function to generate the final facial image I_g, ensuring high fidelity and consistency with the target facial pose and expressions.

Methods	Same-identity Reenactment on VoxCeleb1						Cross-identity Reenactment on VoxCeleb1			Cross-identity Reenactment on CelebV		
	SSIM↑	$\mathcal{L}_1$↓	PSNR↑	LPIPS↓	AKD↓	AED↓	CSIM↑	ARD↓	AUH↓	CSIM↑	ARD↓	AUH↓
FOMM	0.851	0.0415	23.95	0.169	1.384	0.129	0.745	2.416	0.243	0.695	2.497	0.252
MRAA	0.860	0.0391	24.41	0.163	1.377	0.126	0.749	2.938	0.297	0.699	2.891	0.301
DaGAN	0.862	0.0383	24.69	0.164	1.368	0.123	0.751	2.583	0.241	0.689	2.583	0.232
TPSM	0.879	0.0351	25.62	0.150	1.217	0.121	0.759	2.256	0.236	0.707	2.341	0.222
MRFA	0.889	0.0329	26.26	0.143	1.177	0.106	0.732	2.219	0.232	0.684	2.276	0.215
FNeVR	0.876	0.0367	25.35	0.154	1.246	0.119	0.748	2.319	0.237	0.691	2.422	0.236
MCNet	0.882	0.0347	25.83	0.148	1.243	0.110	0.743	2.206	0.243	0.681	2.366	0.235
Ours	**0.896**	**0.0306**	**26.88**	**0.136**	**1.156**	**0.098**	**0.779**	**2.124**	**0.217**	**0.711**	**2.265**	**0.201**

Table 1: Comparisons with state-of-the-art methods on same-identity reenactment on VoxCeleb1 dataset and cross-identity reenactment on VoxCeleb1 and CelebV datasets.

Objective Function

We train our method using a combination of three losses. First, following FOMM (Siarohin et al. 2019b), we use the perceptual loss $\mathcal{L}_p$ to align the human-perceptual similarity between the generated frame and the driving frame. Second, we introduce the equivariance loss $\mathcal{L}_{eq}$ to ensure the stability of the Facial Pose-aware Motion Estimator. Additionally, we incorporate the localization loss $\mathcal{L}_{loc}$ (Eq. 12) to learn the facelet semantic compensation code. The overall loss function is formulated as follows:

$$\mathcal{L} = \lambda_p \mathcal{L}_p + \lambda_{eq} \mathcal{L}_{eq} + \lambda_{loc} \mathcal{L}_{loc}, \tag{13}$$

where λ_p, λ_{eq}, and λ_{loc} are the weights for the respective loss functions.

Experiments

We follow existing works (Siarohin et al. 2019b, 2021) to employ Hourglass network architecture for keypoint estimation (Newell, Yang, and Deng 2016). Our facial mask estimator comprises three `Downsample ResBlocks`, followed by an `AdaIN ResBlock` and three `Upsample ResBlocks`. Similarly, the encoder and facial pose-aware decoder consist of four `Downsample` and `AdaIN Upsample ResBlocks`. We set the number of keypoints K as 10, in line with established methodologies (Siarohin et al. 2019b; Zhao and Zhang 2022). Our model is trained for 100 epochs using two RTX 4090 GPUs in an end-to-end training manner, taking up to approximately 5 days in total. The Adam optimizer (Kingma and Ba 2015) is adopted with learning rate with $2e^{-4}$, $\beta_1 = 0.5$ and $\beta_2 = 0.999$.

Experimental Setting

Datasets. In our experiments, we use two commonly used datasets for the validation of talking head generation: Voxceleb1 (Nagrani, Chung, and Zisserman 2017) and CelebV dataset (Wu et al. 2018). Voxceleb1 consists of videos of different celebrities talking in different scenarios, collected from YouTube and cropped to a resolution of 256×256, with various lengths ranging from 64 to 1024 frames. CelebV contains 200K images of celebrity faces, each annotated with 98 landmarks. We follow the same data pre-processing protocol and train-test split strategy in (Siarohin et al. 2019b; Hong et al. 2022) for evaluation. To assess the model's generalization ability, we adopt the test set sampling strategy from DaGAN (Hong et al. 2022), training our model on VoxCeleb1 while evaluating it on CelebV.

Metrics. We assess the reconstruction quality using structured similarity (SSIM), L1 distance ($\mathcal{L}_1$), Peak Signal-to-Nosie Ratio (PSNR), and LPIPS (Zhang et al. 2018). To evaluate the pose quality of the generated videos, we follow previous studies (Siarohin et al. 2019b; Zhao and Zhang 2022; Hong et al. 2022) that compute the Average Keypoint Distance (AKD) between the generated and ground-truth video frames. The identity preservation capability is measured by the Average Euclidean Distance (AED) between identity features. Additionally, for cross-identity video face reenactment, we use the cross-identity similarity (CSIM) (Wang, Mallya, and Liu 2021) to evaluate the quality of identity preservation. The average rotation distance (ARD) (Doukas, Zafeiriou, and Sharmanska 2021) and the facial action unit hamming distance (AUH) (Doukas, Zafeiriou, and Sharmanska 2021) are respectively used to measure errors of head poses and facial expressions. For comparison with them, we adopt the motion model (Tao et al. 2024) to compute our motion flow, while we also employ the motion models (Siarohin et al. 2019b; Zhao and Zhang 2022) in the section of model ablations.

Competitors. We compare our method with a variety of state-of-the-art methods, FOMM (Siarohin et al. 2019b), MRAA (Siarohin et al. 2021), DaGAN (Hong et al. 2022), TPSM (Zhao and Zhang 2022), FNeVR (Zeng et al. 2022), MRFA (Tao et al. 2024), and MCNet (Hong and Xu 2023). All reported results are obtained by evaluating these methods with their official code implementation and publicly available pre-trained models.

Quantitative Comparison

Comparison with SOTAs. For same-identity face reenactment, we conduct the comparisons on VoxCeleb1 dataset, in which the source and the driving images share the same identity, and the quantitative results are reported in the left part of Table 1. We can see that our method consistently outperforms other approaches across all metrics. Particularly noteworthy are the substantial improvements in reconstruction metrics such as SSIM, $\mathcal{L}_1$, PSNR, and LPIPS, which underscore our method's proficiency in capturing finer facial details and achieving superior visual quality. That should attributed to the introduction of facelet semantic compensation, as it compensates the initially warped feature with the learned facelet semantic bank, which effectively leverages

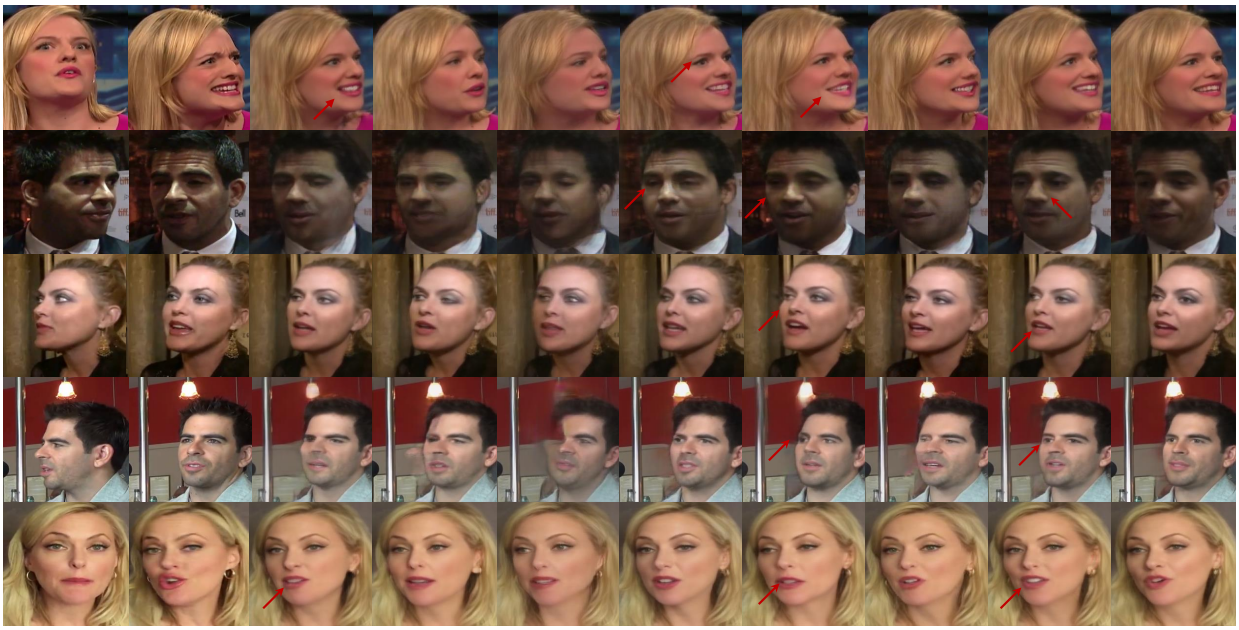


Figure 3: Comparison of same-identity video reenactment results on the VoxCeleb1 dataset. Our method consistently outperforms multiple state-of-the-art approaches. Please zoom in for more details.

Methods	Same-identity Reenactment					
	SSIM ↑	PSNR ↑	LPIPS ↓	$\mathcal{L}_1$ ↓	AKD ↓	AED ↓
FOMM	0.791	21.20	0.245	0.0606	1.600	0.246
MRAA	0.779	20.98	0.253	0.0613	2.042	0.271
DaGAN	0.795	21.34	0.249	0.0635	1.805	0.253
TPSM	0.829	22.63	0.216	0.0542	1.430	0.220
MRFA	0.823	22.69	0.220	0.0519	1.381	0.218
FNeVR	0.812	21.99	0.230	0.0567	1.456	0.253
MCNet	0.821	22.64	0.222	0.0524	1.427	0.212
Ours	**0.837**	**23.28**	**0.202**	**0.0477**	**1.326**	**0.207**

Table 2: Comparisons with state-of-the-art methods on a set of images with self-occlusion occurring on source face.

Variants	SSIM↑	$\mathcal{L}_1$↓	PSNR↑	LPIPS↓	AKD↓	AED↓
Baseline	0.889	0.0329	26.26	0.143	1.177	0.106
+ PAM	0.890	0.0326	26.40	0.141	1.172	0.111
+ PAD	0.891	0.0322	26.44	0.141	1.168	0.108
+ IR	0.895	0.0311	26.79	0.137	1.160	0.099
+ $\mathcal{L}_{loc}$	**0.896**	**0.0306**	**26.88**	**0.136**	**1.156**	**0.098**
FOMM	0.851	0.0415	23.95	0.169	1.384	0.129
FOMM ++	0.881	0.0341	25.86	0.147	1.209	0.114
TPSM	0.879	0.0351	25.62	0.150	1.217	0.121
TPSM ++	0.890	0.0336	26.35	0.142	1.174	0.108

Table 3: Quantitative comparison of various variants.

semantic facial information to refine facial regions where self-occlusion occurs. Furthermore, our method achieves superior facial motion quality as evidenced by the highest AKD score, showcasing the effectiveness of our facial pose-aware motion estimator and facial pose-aware decoder in representing spatially-invariant pose and capturing facial motion. Leveraging our facelet semantic bank, we also achieve the best identity preservation, as indicated by the highest score on the AED metric. We also conduct cross-identity face reenactment experiments on both the VoxCeleb1 and CelebV datasets. This task involves using source and driving images from different identities. As shown in the right part of Table 1, our method achieves the best ARD and AUH metric scores, which indicates the superior facial expression and accurate head pose quality of our method. Additionally, the highest CSIM score illustrates our method's ability to maintain high appearance fidelity during face animation. This capability is further supported by the qualitative results (Fig. 3 and Fig. 4).

Occlusion Evaluation. To further assess the efficacy of our proposed methodology under challenging scenarios involving self-occlusion in the source face image, we constructed a specialized benchmark dataset tailored for self-reenactment tasks. This dataset consists of 1,300 image pairs, each containing a source image with self-occlusion, manually curated from the VoxCeleb1 test set. The comparative analysis, presented in Table 2, indicates that our approach outperforms existing techniques in terms of both image quality and the accuracy of head pose and expression transfer fidelity.

Qualitative Comparison

We conducted experiments on both the same-identity video reconstruction and cross-identity face reenactment similar to prior studies (Siarohin et al. 2019b; Hong et al. 2022; Hong and Xu 2023). The qualitative outcomes are visually depicted in Fig. 3 and Fig. 4. Notably, our method consistently demonstrates enhanced robustness against extreme changes in head pose (illustrated in 1_{st}, 2_{nd}, 4_{th} samples of

Figure 4: Cross-identity face reenactment on VoxCeleb1 (first three rows) and CelebV (last two rows). Our method consistently achieves higher fidelity than existing techniques. Please zoom in for details.

Fig. 3 and 2_{nd}, 5_{th} rows in Fig. 4), delicate expression variations (see 3_{rd}, 5_{th} samples in Fig. 3 and 2_{nd}, 4_{th} samples of Fig. 4), and challenging occasions where source face is partially occluded (3_{rd}, 4_{th} samples of Fig. 3 and 1_{st}, 3_{rd} rows in Fig. 4), etc. Furthermore, our approach outperforms competing methodologies in maintaining appearance fidelity and capturing diverse facial expressions with intricate details, notably within fine-grained facial regions such as the eyes, lips, and ears, while circumventing prevalent artifacts. Our method can effectively learn the occlusion-insensitive facelet semantic bank for compensation and enhance facial pose retargeting, thereby manifesting a notable advancement over existing approaches that often exhibit constraints in maintaining identity consistency and facial local details.

Ablation Studies

We conducted an ablation study to evaluate the effectiveness of each component in our model. We began by removing the pose retargeting strategy to assess the performance of the pose-aware motion estimator and the proposed pose-aware decoder. Next, we evaluated the facelet compensation approach by incrementally adding Pose-Aware Motion (PAM), Pose-Aware Decoder (PAD), and Identity Refinement (IR). Finally, we introduced the $\mathcal{L}_{loc}$ loss during training to validate its impact. Quantitative results for all variants are presented in Table 3. Both the Pose-Aware Motion and Pose-Aware Decoder significantly improve the $\mathcal{L}_1$ and AKD metrics, leading to more accurate head pose and global motion estimation. Adding Identity Refinement further enhances all metrics, particularly SSIM, $\mathcal{L}_1$, AED, and PSNR, reflecting improved identity consistency and overall visual quality, demonstrating the effectiveness of our facelet compensation approach. Including $\mathcal{L}_{loc}$ helps mitigate identity blending among facial features, resulting in superior performance across all tasks.

Moreover, we integrated our full components into FOMM (Siarohin et al. 2019b) and TPSM (Zhao and Zhang 2022), leading to the variants FOMM++ and TPSM++. As shown in Table 3, these integrations substantially improve all metrics, confirming the adaptability, transferability, and generalization capability of our proposed method.

Conclusion and Discussions

In this paper, we introduce an occlusion-insensitive method for talking head video generation. Unlike traditional approaches that rely on keypoint correlations for motion flow, our method uses semantic correlations to address artifacts from global warping, especially in occluded areas. We estimate motion flow to warp source features, creating an initial warped feature, which is then refined using a facelet semantic bank. A decoder processes this refined feature to generate the final animated faces. Experiments in same-identity and cross-identity reenactment demonstrate that our method excels in preserving identity, fine details, and facial expressions.

Limitations. One limitation of our method is that facial shapes may experience minor adjustments to better align with the driving frame. This is due to keypoints capturing facial shape information, which can lead to subtle deformations beyond expressions, particularly in cross-identity animation. While relative motion transfer (Siarohin et al. 2019b, 2021) can help mitigate this effect, it may result in slightly less precise motion portrayal. Alternatively, incorporating prior models to capture facial shape nuances could further enhance our method's accuracy.

Acknowledgments

This work is supported by the Guangdong Natural Science Funds for Distinguished Young Scholars (Grant 2023B1515020097), the AI Singapore Programme under the National Research Foundation Singapore (Grant AISG3-GV-2023-011), and the Lee Kong Chian Fellowships.

References

Bansal, A.; Ma, S.; Ramanan, D.; and Sheikh, Y. 2018. Recycle-gan: Unsupervised video retargeting. In *ECCV*, 119–135.

Blanz, V.; and Vetter, T. 1999. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, 187–194.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM TOG*, 42(4): 1–10.

Chen, L.; Maddox, R. K.; Duan, Z.; and Xu, C. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, 7832–7841.

Chen, Z.; Wang, C.; Yuan, B.; and Tao, D. 2020. Puppeteergan: Arbitrary portrait animation with semantic-aware appearance transformation. In *CVPR*, 13518–13527.

Doukas, M. C.; Zafeiriou, S.; and Sharmanska, V. 2021. Headgan: One-shot neural head synthesis and editing. In *ICCV*, 14398–14407.

Feng, Y.; Feng, H.; Black, M. J.; and Bolkart, T. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *ACM TOG*, 40(4): 1–13.

Gao, Y.; Zhou, Y.; Wang, J.; Li, X.; Ming, X.; and Lu, Y. 2023. High-fidelity and freely controllable talking head video generation. In *CVPR*, 5609–5619.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. volume 27.

Guo, Y.; Chen, K.; Liang, S.; Liu, Y.-J.; Bao, H.; and Zhang, J. 2021. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *ICCV*, 5784–5794.

Hong, F.-T.; Shen, L.; and Xu, D. 2023. DaGAN++: Depth-Aware Generative Adversarial Network for Talking Head Video Generation.

Hong, F.-T.; and Xu, D. 2023. Implicit identity representation conditioned memory compensation network for talking head video generation. In *ICCV*, 23062–23072.

Hong, F.-T.; Zhang, L.; Shen, L.; and Xu, D. 2022. Depth-aware generative adversarial network for talking head video generation. In *CVPR*, 3397–3406.

Ji, X.; Zhou, H.; Wang, K.; Wu, Q.; Wu, W.; Xu, F.; and Cao, X. 2022. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *SIGGRAPH*, 1–10.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Li, T.; Bolkart, T.; Black, M. J.; Li, H.; and Romero, J. 2017. Learning a model of facial shape and expression from 4D scans. *ACM TOG*, 36(6): 194–1.

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

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. VoxCeleb: a large-scale speaker identification dataset. In *InterSpeech*.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *ECCV*, 483–499. Springer.

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 *ACM Multimedia*, 484–492.

Ren, Y.; Li, G.; Chen, Y.; Li, T. H.; and Liu, S. 2021. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *ICCV*, 13759–13768.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019a. Animating arbitrary objects via deep motion transfer. In *CVPR*, 2377–2386.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019b. First Order Motion Model for Image Animation. In *NeurIPS*.

Siarohin, A.; Sangineto, E.; Lathuiliere, S.; and Sebe, N. 2018. Deformable gans for pose-based human image generation. In *CVPR*, 3408–3416.

Siarohin, A.; Woodford, O. J.; Ren, J.; Chai, M.; and Tulyakov, S. 2021. Motion representations for articulated animation. In *CVPR*, 13653–13662.

Suwajanakorn, S.; Seitz, S. M.; and Kemelmacher-Shlizerman, I. 2017. Synthesizing obama: learning lip sync from audio. *ACM TOG*, 36(4): 1–13.

Tao, J.; Gu, S.; Li, W.; and Duan, L. 2024. Learning Motion Refinement for Unsupervised Face Animation. In *NeurIPS*, volume 36.

Tao, J.; Wang, B.; Ge, T.; Jiang, Y.; Li, W.; and Duan, L. 2022a. Motion Transformer for Unsupervised Image Animation. In *ECCV*, 702–719.

Tao, J.; Wang, B.; Xu, B.; Ge, T.; Jiang, Y.; Li, W.; and Duan, L. 2022b. Structure-Aware Motion Transfer with Deformable Anchor Model. In *CVPR*, 3637–3646.

Thies, J.; Elgharib, M.; Tewari, A.; Theobalt, C.; and Nießner, M. 2020. Neural voice puppetry: Audio-driven facial reenactment. In *ECCV*, 716–731.

Thies, J.; Zollhöfer, M.; and Nießner, M. 2019. Deferred neural rendering: Image synthesis using neural textures. *ACM TOG*, 38(4): 1–12.

Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; and Nießner, M. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2387–2395.

Vlasic, D.; Brand, M.; Pfister, H.; and Popovic, J. 2006. Face transfer with multilinear models. In *SIGGRAPH*, 24–es.

Wang, Q.; Zhang, L.; and Li, B. 2021. SAFA: Structure Aware Face Animation. In *3DV*, 679–688.

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 10039–10049.

Wang, Y.; Yang, D.; Bremond, F.; and Dantcheva, A. 2022. Latent Image Animator: Learning to Animate Images via Latent Space Navigation. In *ICLR*.

Wu, W.; Zhang, Y.; Li, C.; Qian, C.; and Loy, C. C. 2018. Reenactgan: Learning to reenact faces via boundary transfer. In *ECCV*, 603–619.

Xu, C.; Li, K.; Luo, X.; Xu, X.; He, S.; and Zhang, K. 2022. Fully deformable network for multiview face image synthesis.

Xu, X.; Li, K.; Xu, C.; and He, S. 2020. GDFace: Gated deformation for multi-view face image synthesis. volume 34, 12532–12540.

Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 325–341.

Zakharov, E.; Shysheya, A.; Burkov, E.; and Lempitsky, V. 2019. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, 9459–9468.

Zeng, B.; Liu, B.; Li, H.; Liu, X.; Liu, J.; Chen, D.; Peng, W.; and Zhang, B. 2022. FNeVR: Neural Volume Rendering for Face Animation. In *NeurIPS*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 586–595.

Zhao, J.; and Zhang, H. 2022. Thin-plate spline motion model for image animation. In *CVPR*, 3657–3666.

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 *CVPR*, 9729–9738.

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