

From Pixels to Portraits: A Comprehensive Survey of Talking Head Generation Techniques and Applications

Shreyank N Gowda^a, Dheeraj Pandey^a, Shashank Narayana Gowda^a

^a*Videosyncracy, Bangalore, India*

Abstract

Recent advancements in deep learning and computer vision have led to a surge of interest in generating realistic talking heads. This paper presents a comprehensive survey of state-of-the-art methods for talking head generation. We systematically categorises them into four main approaches: image-driven, audio-driven, video-driven and others (including neural radiance fields (NeRF), and 3D-based methods). We provide an in-depth analysis of each method, highlighting their unique contributions, strengths, and limitations. Furthermore, we thoroughly compare publicly available models, evaluating them on key aspects such as inference time and human-rated quality of the generated outputs. Our aim is to provide a clear and concise overview of the current landscape in talking head generation, elucidating the relationships between different approaches and identifying promising directions for future research. This survey will serve as a valuable reference for researchers and practitioners interested in this rapidly evolving field.

Keywords: Talking head, facial animation, lip movement, facial expression

1. Introduction

In recent years, deep learning and deep neural networks [1, 2, 3] have revolutionized the field of computer vision and enabled transformative breakthroughs in areas like talking head generation. Deep learning utilizes neural networks with many layers that can learn hierarchical representations of data and model complex functions. The rise of deep learning since the early 2010s has allowed computer vision models to achieve state-of-the-art results on a diverse range of tasks. One captivating application empowered by deep neural networks is talking head generation - the synthesis of realistic and expressive

human faces that convincingly articulate speech. This survey paper provides an overview of the state-of-the-art techniques and methodologies employed in talking head generation, exploring the underlying algorithms, datasets, and evaluation metrics. Talking head generation has gained significant attention in recent years as a mesmerizing domain that highlights the capabilities of deep learning.

Early approaches in talking head generation focused on rule-based techniques, generating lip movements from audio or text inputs [4, 5]. However, the rise of deep learning and the availability of large-scale face datasets revolutionized the field[6, 7, 8, 9, 10, 11]. Data-driven methods utilizing neural networks emerged as the dominant paradigm, enabling the synthesis of natural and synchronized facial movements. In fact, the rapid advancement of computing power has sparked significant interest in deep learning-based talking-head generation, leading to the flourishing development of this field.

Two key deep learning techniques that have driven advancements in talking head generation are generative adversarial networks (GANs) [12] and attention mechanisms [13]. GANs have led to significant progress in computer vision tasks like image synthesis [14, 15, 16], image-to-image translation [17, 18, 19], synthetic feature generation [20, 21, 22] and augmentation techniques [23, 24, 25] among others. Attention mechanisms have enhanced computer vision models in areas like object detection [26, 27, 28], image captioning [29, 30, 31], video captioning [32, 33, 34] and action recognition [35, 36, 37] among others by enabling focused processing of salient regions. GANs leverage generative and discriminative networks to model complex data distributions, while attention allows models to focus on salient parts of the input. Conditional adversarial networks [38] (cGANs) have played a significant role in talking head generation. These models leverage generative and discriminative networks to capture the complex relationship between audio or text inputs and facial expressions, resulting in highly realistic and expressive faces. Attention mechanisms have also been employed [39, 40, 8] to enhance the quality and realism of generated talking heads, allowing models to focus on specific facial regions and generate fine-grained details.

Evaluation metrics are crucial for assessing the quality and perceptual realism of generated talking heads. While evaluation metrics such as peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) have been commonly used to assess the quality [41] of generated talking heads, they may not capture the perceptual realism and naturalness of the synthesized faces accurately. These metrics focus primarily on pixel-level similarity

and do not account for higher-level visual cues such as facial expressions, lip movements, and overall coherence in speech animation. Additionally, subjective evaluation through human perceptual studies provides valuable insights into the perceptual quality and believability of the generated talking heads, but it can be subjective and time-consuming. Therefore, there is a need for more comprehensive and perceptually grounded evaluation metrics that can effectively measure the quality and authenticity of generated talking heads in a more holistic manner.

In this survey paper, we systematically categorize the approaches in talking head generation into four main categories: image-driven, audio-driven, video-driven, and others (including neural radiance fields (NeRF) and 3D-based methods). Each approach is analyzed in-depth, emphasizing their unique contributions, strengths, and limitations. Furthermore, we conduct a comprehensive comparison of publicly available models, evaluating them based on key aspects such as inference time and human-rated quality of the generated outputs. Our objective is to provide a clear and concise overview of the current landscape in talking head generation, shedding light on the relationships between different approaches and identifying promising directions for future research.

The evaluation of inference time is crucial as it directly impacts the real-time applicability and usability of talking head generation systems. Faster inference times enable seamless integration of these systems into various applications, such as virtual reality, video conferencing, and gaming. Additionally, assessing the human-rated quality of the generated outputs is vital for ensuring that the synthesized talking heads meet the perceptual expectations of human viewers. This subjective evaluation helps gauge the level of realism, naturalness, and believability of the generated faces, ultimately determining their effectiveness in engaging and communicating with human users. By considering both inference time and human-rated quality, we can gain comprehensive insights into the performance and usability of talking head generation models in practical scenarios.

This survey will serve as a valuable reference for researchers and practitioners interested in this rapidly evolving field, fostering further advancements and innovations in talking head generation.

2. What Comprises a Good Talking Head Generation?

There has been prior work [42] that studied the concept of what makes a good talking head generation along with a survey about the methods. However, the survey is over three years old and there has been a wealth on new research in the domain. Along with this, there is no human evaluation on the methods in the survey, something that we believe is critical to properly evaluate how good the output from a model is. None the less, we summarize what makes a good talking head generation according to the paper.

The authors note that while talking-head video generation has progressed significantly, the evaluation of these methods presents several challenges. Many current evaluation approaches use human subjects, which can be cumbersome, unrepeatable, and may hinder the evolution of new research. To address these issues, the authors have designed a benchmark for evaluating talking-head video generation methods, with standardized dataset pre-processing strategies.

Four desired properties for a good synthesized talking-head video are proposed: preserving the subject’s original identity, maintaining lip-sync at a semantic level, keeping high visual quality, and containing spontaneous motions. The authors propose new metrics or select the most appropriate ones to evaluate these properties. These include:

- Identity Preserving: Using cosine similarity between embedding vectors of ArcFace to measure identity mismatch.
- Visual Quality: Using SSIM and FID to evaluate visual quality at an image-level, and CPBD to judge the sharpness of the synthesized video.
- Lip-sync: The video should maintain synchronization at a semantic level. This means the movement of the lips in the video should match the speech that is heard.
- Natural Motion: The video should contain spontaneous, natural movements. This refers to the smoothness and naturalness of the head and facial movements in the video.

2.1. Fundamentals in Talking Head Generation

The main idea is to animate a static source image (the person’s face) to match the dynamics of a driving video or audio. This is often achieved

through a combination of deep learning techniques, including convolutional neural networks (CNNs), generative adversarial networks (GANs), and recurrent neural networks (RNNs), often aided by attention mechanisms.

Let's denote the source image as I_s , the driving video as V_d , and the generated video as V_g . The process can be mathematically described as:

$$V_g = f(I_s, V_d; \theta) \quad (1)$$

where f is a generative function parameterized by θ , which represents the parameters of our model.

For a more detailed discussion, let's break this process into a few key steps: encoding, transformation, and decoding.

Encoding. The source image I_s and frames of the driving video V_d are passed through an encoder network E , which extracts deep feature representations. This can be formulated as:

$$F_s = E(I_s; \theta_E), \quad F_d = E(V_d; \theta_E) \quad (2)$$

where F_s and F_d are the deep feature representations of the source image and driving video, respectively. θ_E denotes the parameters of the encoder.

Transformation. Next, a transformation function T maps the source image features F_s and driving video features F_d to a new set of features F_t that retains the identity from F_s and the expressions from F_d . This can be expressed as:

$$F_t = T(F_s, F_d; \theta_T) \quad (3)$$

where θ_T are the parameters of the transformation function.

Decoding Finally, a decoder function D is used to generate the output video frames V_g from the transformed features F_t . This can be described as:

$$V_g = D(F_t; \theta_D) \quad (4)$$

where θ_D are the parameters of the decoder.

Loss Functions. Training these models involves optimizing a combination of loss functions. Commonly used loss functions include:

Reconstruction loss: This ensures that the generated video resembles the source image in terms of identity.

$$L_{\text{rec}} = \|I_s - D(E(I_s; \theta_E); \theta_D)\|_2 \quad (5)$$

Adversarial loss: This ensures that the generated video frames are indistinguishable from real video frames.

$$L_{\text{adv}} = \mathbb{E}V_d \sim p_{\text{data}}(V) [\log D(V_d)] + \mathbb{E}I_s \sim p_{\text{gen}}(I) [\log(1 - D(I_s))] \quad (6)$$

where D is a discriminator network, and $p_{\text{data}}(V)$ and $p_{\text{gen}}(I)$ represent the probability distributions of the real data and the generated images, respectively.

Perceptual loss: This loss ensures that the generated video preserves the content and style of the source image and the driving video. It's often calculated as the L2 distance between the high-level features extracted from a pre-trained CNN (e.g., VGG-16).

$$L_{\text{perc}} = \|\phi(I_s) - \phi(D(E(I_s; \theta_E); \theta_D))\|_2 \quad (7)$$

where $\phi(\cdot)$ denotes the feature extraction function.

Temporal coherence loss: This loss ensures the smoothness of the generated video frames. One example of such a loss is optical flow consistency. The final loss function is a weighted sum of these individual loss terms:

$$L_{\text{total}} = \lambda_{\text{rec}} L_{\text{rec}} + \lambda_{\text{adv}} L_{\text{adv}} + \lambda_{\text{perc}} L_{\text{perc}} \quad (8)$$

where λ_{rec} , λ_{adv} , and λ_{perc} are the weights assigned to each loss.

This is a broad overview of the process and has been described in Figure 1 where the driving media is a video. However, there are many variations and improvements over this basic structure, such as using attention mechanisms to focus on certain regions of the face, using different types of encoders and decoders, leveraging keypoints to guide the animation process, and more. The choice of architecture and techniques depends largely on the specific requirements of the application.

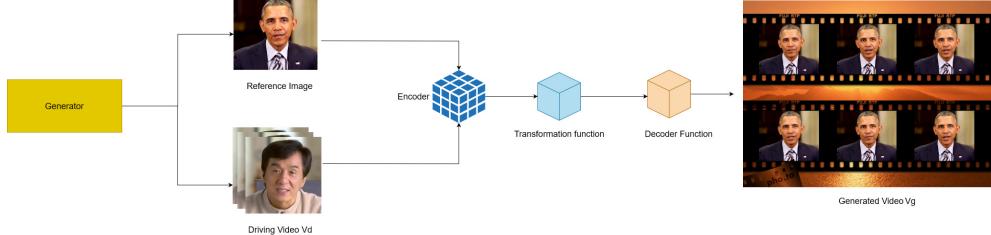


Figure 1: An example pipeline of a talking head generation model. In this figure, the driving media is that of a video.

3. Categorising Methods

Talking head generation methods can be classified according to the driving media, leading to several prominent categories. Visual-driven methods encompass a wide range of techniques that leverage both still images and sequential frames as primary inputs. These methods utilize sophisticated machine learning algorithms to animate the visual content, simulating speech, varying expressions, and capturing nuanced movements. In contrast, audio-driven techniques take audio inputs to synthesize corresponding facial motions, specifically focusing on lip-syncing and emotional expressions that align with the vocal input. Finally, a less common but increasingly explored category includes methods like NeRF-based (Neural Radiance Fields) and 3D-aware techniques. These advanced models employ 3D geometry understanding and complex light interactions to create remarkably immersive and lifelike animations, offering substantial potential for further research and development.¹

3.1. Visual-driven

We further breakdown visual-driven approaches to broadly 5 groups as shown in Figure 2. This grouping allows us to easily understand future possible directions to work on based on type of research being performed in this domain. Note that this grouping does not necessarily mean that an approach does not fit into one of the sub groups, but more that it is our belief that it fits better in the chosen sub group.

¹<https://github.com/harlanhong/awesome-talking-head-generation>

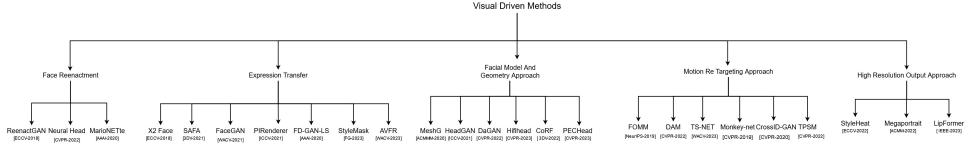


Figure 2: Categorization of visual-driven talking head generation approaches. This broad grouping highlights the diversity of techniques and future research directions in synthesizing talking heads from audio. This categorization provides an overview of the landscape and evolution of audio-driven talking head generation.

3.1.1. Face Reenactment Approaches

The task of talking head generation was initially formulated as a real-time facial reenactment of a monocular target video sequence [43]. They propose a method that captures the facial expressions of a subject in real-time, and transfers these expressions to a target subject in a video, effectively creating a facial reenactment. This is achieved using a dense photometric consistency measure, a fast model fitting algorithm, and a novel re-rendering process.

ReenactGAN [44], can transfer facial movements and expressions from any individual’s monocular video input to another person’s video. Rather than transferring directly in the pixel space, which could lead to structural artifacts, it initially maps the source face onto a boundary latent space. A transformer is then used to adjust the source face’s boundary to match the target’s boundary, and finally, a target-specific decoder generates the reenacted target face.

NeuralHead [45] proposes a method for creating realistic talking head models with just a few images of a target person. The model uses adversarial learning techniques to produce high-quality animated sequences that maintain the unique details of the person, such as their facial expressions and characteristics. They train a few-shot meta-learning based approach to tackle the problem of talking head generation in a limited labelled setting.

MarioNETte [46] introduces a novel approach for few-shot face reenactment that preserves the identity of unseen target individuals. The method utilizes a generative adversarial network (GAN) with a modified architecture that includes a conditional module and an identity preservation module. This allows for the transfer of facial expressions from a source actor to an unseen target actor while preserving the unique identity characteristics of the target. By incorporating additional identity loss and identity augmentation techniques, the approach improves the fidelity of the reenacted faces and

maintains the individuality of the target subjects.

3.1.2. Expression Transfer and Attribute Manipulation Approaches

X2Face [47] is a deep learning model, that can control a source face, specified by one or more frames, using another face in a driving frame to produce a generated frame with the identity of the source frame but the pose and expression of the face in the driving frame. The network is trained fully self-supervised using a large collection of video data. Additionally, the authors demonstrate that the generation process can be driven by other modalities, such as audio or pose codes, without any further training of the network.

SAFA [48] focuses on capturing and preserving the structural details of the face to improve the fidelity of the animation. SAFA combines a 3D facial shape model with a deep learning framework to learn and manipulate the facial expressions. By incorporating a shape-based regularization term and an expression-based objective, the approach ensures accurate and consistent facial animation. Additionally, SAFA leverages a hierarchical structure to model the spatial relationships between facial components, enabling more coherent and natural-looking animations.

FACEGAN [49] employs a two-stage pipeline consisting of an attribute predictor and a reenactment generator. The attribute predictor learns to extract and classify facial attributes, while the reenactment generator generates reenacted faces based on the source expressions and target attributes. By conditioning on both the source and target attributes, FACEGAN enables fine-grained control over the reenactment process, allowing for attribute-specific manipulation.

PIRenderer [50] combines the power of generative adversarial networks (GANs) with a semantic decomposition module to achieve fine-grained control over various aspects of the generated portraits. By separating the latent code into semantic attributes, such as pose, expression, and lighting, the method allows for independent manipulation and adjustment of these attributes during the image generation process.

FD-GAN-LS [51] focuses on face/head reenactment, aiming to transfer the facial pose of a target face to a source face. They introduce a pipeline to discover latent directions in the GAN space that control facial pose and expression variations, leveraging a 3D shape model that captures disentangled directions for pose, identity, and expression. The method allows reenactment of real-world faces, supports one-shot reenactment with a single source

image, and enables cross-person reenactment.

StyleMask [52] addresses the problem of neural face reenactment, aiming to transfer the target’s pose and expressions to a source image while preserving the source’s identity characteristics, even when the faces belong to different identities. They propose a framework that disentangles identity and pose using unpaired facial images and the style space of StyleGAN2.

AVFR [53] transfers head motion from a driving video to animate a source image using a dense motion field generated by learnable keypoints. The quality of lip sync is enhanced by incorporating audio as an additional input, allowing the network to focus on the mouth region. Face segmentation and face mesh priors are employed to improve the structural accuracy of the reconstructed faces. Additionally, a carefully designed identity-aware generator module enhances visual quality by generating high-quality outputs with fine-grained details, using the source image and warped motion features as input.

3.1.3. 3-D Facial Models and Geometry Approaches

MeshG [54] presents a novel approach for one-shot face reenactment that uses mesh-guided graph convolutional networks. The method aims to transfer the facial expressions of a source person to a target person in a single input image. By leveraging a mesh representation of the face, the approach extracts local and global features and applies graph convolutional networks to model the relationships between facial regions.

HeadGAN [10] employs a 3D face representation to condition synthesis, thereby separating facial identity from expression. It also uses audio features to enhance mouth movements. The utilization of 3D face representation allows HeadGAN to function as a real-time reenactment system, an efficient tool for facial video compression and reconstruction, a facial expression editing method, and a novel view synthesis system, including face frontalisation.

DaGAN [8] introduces a self-supervised method to automatically obtain 3D facial geometry (or depth) from face videos without needing costly 3D annotation data. They use the dense depth maps they have learned to estimate sparse facial keypoints that capture essential movements of the human head. The depth information is also used to develop a 3D-aware cross-modal (appearance and depth) attention mechanism to guide the creation of motion fields that warp source image representations.

HifiHead [55] is a high-fidelity neural talking head synthesis method that effectively preserves the appearance of the source image and provides flexible

control over motion aspects such as pose, expression, and gaze. The method utilizes 3D morphable face models (3DMM) parameters derived from a driving image or user input to achieve desired motion effects. By extracting 3D face descriptors and hierarchical representations, the authors guide the generator to faithfully capture appearance and shape information. They introduce a coarse-to-fine style-based generator comprising feature alignment and refinement (FAR) blocks, which efficiently update dense flow fields and enhance RGB outputs.

CoRF [56] focuses on 3D-aware synthesis of face dynamics, specifically addressing the control of generative models to produce non-rigid motion, such as facial expression changes, while maintaining 3D-awareness. The proposed approach achieves motion control by embedding motion features in the latent motion space of a style-based generator. To ensure consistency in various attributes like lighting, texture, shapes, and identity, a face parsing network, a head regressor, and an identity encoder are utilized.

PECHead [57] addresses challenges in talking head generation, including unexpected deformation, limited attribute manipulation, and flickering artifacts. It utilizes self-supervised learned landmarks and 3D face model-based landmarks to model motion, while a motion-aware multi-scale feature alignment module transfers motion without face distortion. Additionally, a feature context adaptation and propagation module enhances the smoothness of the synthesized talking head videos.

3.1.4. Motion Re-targeting Approaches

The "First Order Motion Model" (FOMM) [58] is a novel approach to image animation that generates a video sequence by animating an object in a source image according to the motion of a driving video. The framework decouples appearance and motion information through a self-supervised formulation, using a set of self-learned keypoints and their local affine transformations to support complex motions. The model further incorporates an occlusion-aware generator network to handle occlusions and combines the appearance extracted from the source image with the motion derived from the driving video.

The Deformable Anchor Model (DAM) [59] automatically discovers the motion structure of arbitrary objects without prior knowledge. It utilizes motion anchors and a latent root anchor to capture appearance, motion, and structural information. DAM can be extended hierarchically and is learned in an unsupervised manner, enforcing correspondences between anchors to

effectively capture and preserve structural information in motion transfer.

TS-Net [60] is a dual-branch network for video motion retargeting. It combines a warp-based transformation branch and a warp-free synthesis branch to preserve identity and handle occlusion. A mask-aware similarity module is introduced to improve efficiency.

Monkey-net [61] proposes a novel approach for animating diverse objects using deep motion transfer techniques. The method combines a deep learning framework with an optimization process to transfer the motion characteristics from a source object to a target object, enabling realistic and coherent animations. By leveraging pre-trained networks and a motion warping algorithm, the approach allows for the animation of various objects without requiring explicit 3D models or rigging.

CrossID-GAN [62] introduces a novel approach for cross-identity face reenactment by learning identity-invariant motion representations. They focus on addressing the challenge of preserving the unique identity of the target individual while transferring facial expressions from a source actor. By leveraging a motion-guided warping network, the approach learns to separate the motion information from the identity information in the source video. This allows for the generation of reenactment results that accurately reflect the expressions of the source actor while maintaining the identity of the target actor.

TPSM [6] proposes an end-to-end unsupervised motion transfer framework to address the challenge of large pose gaps between objects in source and driving images. It introduces thin-plate spline motion estimation for flexible optical flow and leverages multi-resolution occlusion masks for realistic restoration of missing regions. The method employs auxiliary loss functions to encourage high-quality image generation and successfully animates various objects, such as talking faces, human bodies, and pixel animations.

3.1.5. High Resolution Output Approaches

StyleHEAT [63] leverages the latent feature space of a pre-trained StyleGAN to overcome the resolution limitations of training datasets. Their unified framework enables high-resolution video generation, disentangled control through video or audio input, and flexible face editing. The framework achieves a resolution of 1024×1024 , surpassing the training dataset's lower resolution, and incorporates motion generation modules, a calibration network, and domain loss for refinement.

MegaPortrait [64] focuses on advancing neural head avatar technology to

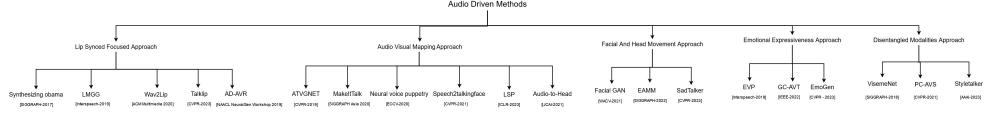


Figure 3: Categorization of audio-driven talking head generation approaches. This broad grouping highlights the diversity of techniques and future research directions in synthesizing talking heads from audio. This categorization provides an overview of the landscape and evolution of audio-driven talking head generation.

achieve megapixel resolution, specifically addressing the challenging task of cross-driving synthesis. They propose new neural architectures and training methods that utilize both medium-resolution video data and high-resolution image data to achieve high-quality rendered images and generalization to new views and motions.

LipFormer [65] is a Transformer-based framework that utilizes a pre-trained codebook of high-quality face images as a prior for capturing fine facial details. LipFormer simplifies the task by focusing on finding appropriate lip-codes to characterize lip variations during talking. An adaptive face warping module is introduced to handle pose variations and aid in accurate lip-code prediction.

3.2. Audio-driven

We further breakdown audio-driven approaches to broadly 5 groups as shown in Figure 3. This grouping allows us to easily understand future possible directions to work on based on type of research being performed in this domain. Note that this grouping does not necessarily mean that an approach does not fit into one of the sub groups, but more that it is our belief that it fits better in the chosen sub group.

3.2.1. Lip-sync Focused Approaches

“Lip Reading in the Wild” [66] proposed a deep learning-based approach combining convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for robust lip reading. They introduced the LRW dataset containing over 500,000 video clips and employed a two-stream architecture to process spatial and motion information.

Synthesizing-Obama [67] presents a deep learning-based method for generating realistic lip sync animations of Barack Obama speaking. The model learns the mapping between audio and visual features using a large dataset

of Obama’s speeches. The generated lip sync animations are compared with real footage, demonstrating promising results.

LMGG [68] introduces a novel method for generating realistic lip movements. The researchers propose a visual speech synthesis model that leverages a variational autoencoder (VAE) architecture. The model learns to map textual inputs to lip movements by encoding and decoding the facial expressions. The approach utilizes a compact and efficient representation of lip movements, enabling fast and accurate synthesis.

Wav2Lip [69] is a lip-syncing algorithm that can generate realistic lip movements based on an input audio waveform and a static image of a person. The model utilizes a two-stage process, where it first predicts a coarse lip shape from the audio, and then refines it using a fine-grained lip shape prediction network. It also employs a face landmark detection model to accurately align the lip shapes with the input image. By combining audio-visual information, Wav2Lip achieves impressive lip-syncing results, even in challenging scenarios with non-native speakers or low-quality audio.

TalkLip [70] proposes using a lip-reading expert to enhance the clarity of generated lip regions, penalizing incorrect generation results. To offset data scarcity, this lip-reading expert is trained in an audio-visual self-supervised way. Additionally, novel contrastive learning is used to improve lip-speech synchronization, and a transformer is deployed to encode audio and video synchronously. The team also suggests a new evaluation strategy using two different lip-reading experts to measure the intelligibility of the created videos.

3.2.2. Audio-Visual Mapping Approaches

AD-AVR [71] proposes an adversarial learning framework that disentangles audio and visual representations to capture the correspondence between them. By combining a visual generator, audio generator, and discriminator, the model learns to generate synchronized facial movements based on audio input. The disentangled representation enables control over facial expressions and lip movements independently.

ATVGNet [72] utilizes a hierarchical structure with multiple stages to handle different aspects of the generation process. It incorporates a cross-modal learning framework, allowing the model to leverage both audio and visual information to enhance facial expressions and lip movements. To further improve the generated results, a dynamic pixel-wise loss is introduced, which adapts the loss function based on the difficulty of synthesizing different

regions of the face.

MakeItTalk [73] leverages speaker embeddings, which encode unique characteristics of a specific speaker, to synthesize realistic facial movements and lip-syncing. The model incorporates a two-stage architecture that separately handles the speaker and lip motion generation. By explicitly considering the speaker information, the generated animations exhibit speaker-specific characteristics, including speech style, facial expressions, and idiosyncrasies.

NeuralVoicePuppetry [74] aims to manipulate the facial movements of a target person in a video by using a different source audio. The model consists of two components: a lip-sync network and a facial reenactment network. The lip-sync network predicts accurate lip movements from the source audio, while the facial reenactment network transfers these movements onto the target person’s face.

Speech2TalkingFace [75] extracts identity-relevant and identity-irrelevant information from a speech clip to animate the speaker’s face. It employs a style-based generative framework, using contrastive learning to map the speaker’s identity and speech content to visual representation spaces. By leveraging class centroids, they strengthen the identity space and use curriculum learning to balance information from different latent spaces.

LSP [76] uses a deep neural network to extract audio features and projects these onto the target person’s speech space. It then learns facial dynamics from these features, predicting head and upper body movements. The final stage uses these predictions to generate feature maps, which, in conjunction with a set of candidate images, are fed into an image-to-image translation network to create photorealistic animations. The system renders high-fidelity details, like wrinkles and teeth, and allows for explicit control over head poses.

Audio2Head [11] proposes a method addressing two main challenges: generating natural head movements that match speech prosody, and maintaining the speaker’s appearance despite large head movements. They use a motion-aware recurrent neural network to predict rigid 6D head movements and then generate detailed facial movements. A keypoint-based dense motion field representation is employed to depict all image movements arising from audio. The dense motion fields are generated from input audio, head poses, and a reference image. Finally, an image generation network creates photorealistic talking-head videos.

3.2.3. Facial and Head Movement Approaches

RhythmicHead [77] utilizes a GAN based framework combined with a recurrent neural network (RNN) to generate both facial expressions and synchronized head movements. By introducing a rhythmic motion module, the model generates head motions that follow natural patterns and rhythms observed in real human speakers.

FACIAL [78] introduces a method to generate realistic talking face videos synchronized with an audio input. The technique involves a Generative Adversarial Network (FACIAL-GAN) that models facial attributes, and methods to encode contextual and phonetic information. The result is photo-realistic video frames with synchronized lip movements, head poses, and realistic eye blinks using a joint explicit and implicit attribute learning framework.

EAMM [79] uses an Audio2Facial-Dynamics module to produce talking faces from audio-driven key-point motion, while an Implicit Emotion Displacement Learner represents emotion-related facial dynamics as linear additions to the motion representations. These combined elements allow the EAMM to generate realistic talking faces for any subject.

SadTalker [7] is a system for generating talking head videos that addresses issues like unnatural head movement, distorted expressions, and identity changes. The system uses 3D motion coefficients derived from audio to implicitly modulate a 3D-aware face render. To achieve realistic motion, they model connections between audio and different types of motion coefficients individually. They use ExpNet to learn facial expressions from audio and PoseVAE to create various head motion styles. The final video is synthesized by mapping the generated 3D motion coefficients to the unsupervised 3D keypoints space of the proposed face render.

3.2.4. Emotional Expressiveness Approaches

EVP [80] employs Generative Adversarial Networks (GANs) to decode and learn correlations between emotional audio cues and corresponding facial expressions. A dedicated emotion encoder is used to isolate emotional content from speech and map it onto the synthesized video portraits. This allows the generation of video portraits that accurately reflect the emotional tone of the input audio.

Granularly Controlled Audio-Visual Talking Heads (GC-AVT) [81] system generates more expressive talking heads in virtual human models. GC-AVT goes beyond existing methods, which typically focus on lip-sync and

head motion, to also incorporate emotional expressions. It disassembles the driving image into three components: a cropped mouth for lip-sync, a masked head for learning pose, and the upper face for generating expressions. These components work together, and their encoded features are balanced through reconstruction training.

EmoGen [82] prioritizes emotional expressions, an aspect often overlooked in prior work. This focus allows the generation of more realistic and engaging videos. The framework is also conditioned on a categorical emotion, enabling it to adapt to arbitrary identities, emotions, and languages, across six defined emotional states.

3.2.5. Disentangled Modalities Approaches

VisemeNet [83] is a deep learning model that predicts viseme sequences directly from audio input. The model allows animators to easily control and modify the generated lip movements through a set of intuitive parameters. By decoupling the animation control from audio processing, VisemeNet offers flexibility and empowers animators to create visually appealing speech animations.

PC-AVS [84] is a framework that separately encodes audio, pose, and identity features, maintaining their implicit modularity, which is critical for achieving flexibility in pose control. The technique uses Generative Adversarial Networks (GANs) and leverages disentangled representation learning to generate highly realistic results.

StyleTalker [85] uses a single reference image and produces a video of a talking person with synchronized lip shapes, realistic head poses, and eye blinks. The system uses a pre-trained image generator and encoder to estimate latent codes reflective of the given audio. Unique components of this model include a contrastive lip-sync discriminator, a conditional sequential variational autoencoder, and an auto-regressive prior augmented with normalizing flow.

3.3. Others

In this Section, we group other methods that we do not believe fit into the general framework used to group methods before.

Write-a-speaker [86] introduces a new framework for generating realistic talking-head videos based on text input. The framework incorporates contextual sentiments, speech rhythm, and pauses to synthesize accurate facial expressions and head motions. It consists of a speaker-independent stage

and a speaker-specific stage. In the speaker-independent stage, three parallel networks generate animation parameters for the mouth, upper face, and head from texts. The speaker-specific stage utilizes a 3D face model guided attention network to tailor videos for different individuals. A high-accuracy motion capture dataset is used to establish authentic correspondences between visual motions and audio, enabling end-to-end training of the network.

NerFACE [87] introduces dynamic neural radiance fields, a method for modeling the appearance and dynamics of a human face. This is crucial for applications like telepresence in augmented reality or virtual reality, where accurate reproduction of appearance and viewpoint is necessary. Unlike existing approaches that explicitly model geometry and material properties or rely solely on images, the proposed method uses an implicit representation based on scene representation networks. To handle facial dynamics, a low-dimensional morphable model is combined with the scene representation network, enabling control over pose and expressions.

HeadNERF [88] learns from monocular RGB portrait videos and encompasses surface geometry and appearance. It consists of a morphable model for coarse shape and expressions, as well as two feed-forward networks for predicting vertex offsets and view- and expression-dependent textures. The proposed method accurately extrapolates to unseen poses and viewpoints, generates natural expressions, and provides detailed textures.

IMAvatar (Implicit Morphable avatar) [89] is a method for learning implicit head avatars from monocular videos. It addresses the limitations of traditional 3D morphable face models (3DMMs) and neural volumetric representations. IMAvatar utilizes learned blendshapes and skinning fields to represent expression- and pose-related deformations, allowing for fine-grained control. It employs ray marching and iterative root-finding techniques for surface intersection computation and introduces a novel analytical gradient formulation for end-to-end training.

ROME (Realistic One-shot Mesh-based human head avatars) [90] is a system that creates realistic human head avatars from a single photograph. The model estimates a person-specific head mesh and neural texture, capturing both photometric and geometric details. The avatars are rigged and can be rendered using a neural network trained alongside the mesh and texture estimators on a dataset of real-world videos.

HiDe-NeRF [91] is a method for high-fidelity and free-view talking-head synthesis. Existing methods relying on 2D representations suffer from face distortion with large head rotations, while recent approaches utilizing 3D

structural representations or implicit neural rendering lack fidelity in identity and expression. HiDe-NeRF addresses these limitations by employing Deformable Neural Radiance Fields to represent the 3D dynamic scene with a canonical appearance field and an implicit deformation field. To enhance identity expressiveness, a generalized appearance module is designed, leveraging multi-scale volume features. Additionally, a lightweight deformation module is proposed to improve expression preciseness by decoupling pose and expression for precise modeling.

TalkCLIP [92] is an expression-controllable one-shot talking head method that specifies facial expressions through natural language descriptions. A text-video paired talking head dataset is constructed, with alternative descriptions containing emotion annotations and fine-grained facial action unit (AU) annotations. The proposed CLIP-based style encoder aligns text embeddings with speaking style representations, leveraging the rich textual knowledge encoded by CLIP.

4. Datasets and Evaluation Metrics

4.1. Datasets

The success of talking head generation heavily relies on the utilization of datasets, which are crucial for producing realistic and compelling results. Generating a convincing animated face that synchronizes with desired speech inputs necessitates a system that can effectively animate facial movements. Optimal selection of datasets plays a pivotal role in ensuring the quality and effectiveness of talking head generation models. By incorporating high-quality datasets that encompass a wide spectrum of facial expressions, poses, lighting scenarios, and demographic variations, the training process becomes more robust, leading to improved fidelity and efficiency in generating talking heads. Additionally, the inclusion of datasets emphasizing the importance of head movements adds an essential dimension to the realism and authenticity of the generated content.

With that in mind, the most commonly used datasets are VoxCeleb1 [93], VoxCeleb2 [94], MEAD [95], CREMA-D [96] and LRW [66]. We also list other used datasets such as GRID [97], MSP-Improv [98], ObamaSet [67], LRW-1000 [99], FaceForensics++ [100] and HDTF [101]. We show the year the dataset was released, the number of hours of data available, the number of unique speakers in the dataset and the number of sentences spoken. We also show if there is any head movement present along with the source of the

video (if it was recorded in a lab or if it is in the wild). All these details are shown in Table 1.

Name	Year	Hours	Speaker	Sentence	Head Movement	Source
GRID	2006	27.5	33	33k	No	Lab
Crema-D	2014	11.1	91	12	Yes	Lab
MSP-Improv	2016	18	12	652	Yes	Lab
LRW	2017	173	1k+	539k	No	Wild
ObamaSet	2017	14	1	N/A	Yes	Wild
VoxCeleb1	2017	352	1.2k	153.5K	Yes	Wild
VoxCeleb2	2018	2.4k	6.1k+	1.1M	Yes	Wild
LRW-1000	2019	57	2k+	718k	Yes	Wild
FaceForensics++	2019	5.7	1k	1k+	Yes	Wild
MEAD	2020	39	60	20	Yes	Lab
HDTF	2021	15.8	362	10k	Yes	Wild

Table 1: Details of typically used datasets. Details include year of release, number of hours, number of unique speakers, number of sentences, presence of head movement and source of recording.

5. Evaluating Quality

Evaluation metrics play a crucial role in assessing the performance and quality of talking head generation models. These metrics provide objective measures that help researchers and developers understand the strengths and weaknesses of different models. While evaluating talking head generation models, several metrics are commonly employed to quantify the visual fidelity, realism, identity preservation, lip synchronization, mouth shape, head motions, and alignment with audio.

In the area of talking face generation, we find two main challenges. First, it is tough to create faces that appear to be talking naturally. Second, it's also difficult to correctly judge how well these created faces are doing. We have some ways to measure this, but they have their own set of problems. For instance, if we use people to rate the quality, the results can be different every time and hard to repeat. On the other hand, if we use numbers to rate the generated videos using some quantifiable metric, some methods might not work well or even give conflicting results.

It's very important to make specific ways to measure how well the talking face models are doing. We can divide these ways of measuring into groups.

Some are based on people's opinions (subjective), and some are based on facts and numbers (objective). Further, some can be expressed in words (qualitative) and others in numbers (quantitative).

- Peak Signal-to-Noise Ratio (PSNR): PSNR is a commonly used metric to measure the similarity between the generated video frames and the ground truth frames. It calculates the ratio of the peak power of the signal (video frame) to the mean square error (MSE) between the generated and ground truth frames.

$$PSNR = 10 \log_{10} \left(\frac{MAX^2}{MSE} \right) \quad (9)$$

where MAX represents the maximum possible pixel value and MSE is the mean square error.

- Structural Similarity Index (SSIM): SSIM assesses the structural similarity between the generated and ground truth frames by considering luminance, contrast, and structural information. It compares the local patterns of the images and produces a value between 0 and 1, where 1 indicates perfect similarity. SSIM is more perceptually meaningful than PSNR as it accounts for the human visual system's characteristics.
- Fréchet Inception Distance (FID): FID measures the distance between the feature representations of the generated and real video frames extracted from a pre-trained Inception network. It evaluates both the visual quality and diversity of the generated samples.

$$FID = \| \mu_{real} - \mu_{gen} \|_2^2 + \text{Tr} \left(\Sigma_{real} + \Sigma_{gen} - 2 (\Sigma_{real} \Sigma_{gen})^{1/2} \right) \quad (10)$$

where μ_{real} and μ_{gen} are the mean feature vectors of the real and generated frames, respectively, and Σ_{real} and Σ_{gen} are their covariance matrices.

- Inception Score (IS): IS assesses the quality and diversity of the generated images by leveraging the output of a pre-trained Inception network. It measures the classifiability of the generated samples and the diversity of the predicted class labels.

- Cumulative Probability Blur Detection (CPBD): CPBD is a metric used to assess image quality by detecting blur. It calculates the cumulative probability of blur detection scores for the generated images.
- Cosine Similarity (CSIM): CSIM is employed to evaluate identity preservation in the frames that have been generated. It measures the cosine similarity between the identity embeddings of the source images and the generated frames.
- Lip Shape Evaluation - Distance Score (LSE-D): LSE-D evaluates the perceptual differences in lip shape between the generated frames and a reference method (e.g., Wav2Lip). It represents the distance score between the lip shapes.
- Lip Shape Evaluation - Confidence Score (LSE-C): LSE-C assesses the confidence in the generated lip shapes. It represents the confidence score associated with the lip shape generation.
- Standard Deviation of Head Motion Feature Embeddings: This metric assesses the diversity of the generated head motions. The standard deviation of the head motion feature embeddings, extracted from the generated frames using Hopenet, is calculated.
- Beat Align Score: The Beat Align Score is used to evaluate the alignment between the audio and the generated head motions. It represents a score that quantifies the synchronization of beats between the audio and the generated head motions.

5.1. Evaluation Using Existing Metrics

We consider video-driven models and compare some methods on a wide range of metrics. We see that no single model does the best on all metrics. This highlights a problem with the existing evaluation metrics. These results are listed in Table 2.

5.2. Comparing Visual Quality of Methods

To visually investigate how these methods perform, we pick random images of celebrities off the internet and test these models.

First, we use images of Barack Obama, Lebron James, Chris Hemsworth and Son Heung-Min, all popular celebrities from different countries. We test

Method	$\mathcal{L}_1 \downarrow$	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	AKD \downarrow	AED \downarrow
Bilayer [102]	0.1753	0.5733	12.802	0.3201	13.83	0.0564
PIRender [50]	0.0574	0.2225	21.154	0.6564	2.249	0.0321
FOMM [58]	0.0451	0.1479	23.422	0.7521	1.456	0.0247
Face vid2vid [103]	0.0456	0.1395	23.279	0.7487	1.615	0.0258
MRAA [104]	0.0511	0.2620	30.892	0.7807	1.796	0.0213
DaGAN [8]	0.0468	0.1465	23.449	0.7564	1.546	0.0257
TPSM [6]	0.0527	0.2540	30.632	0.7822	1.703	0.0210
DaGAN++ [9]	0.0469	0.2440	31.121	0.8015	1.675	0.0195

Table 2: Comparing some video-driven recent SOTA models using different evaluation metrics. We see there is no single model that does significantly well on all metrics raising a question mark on the current evaluation metrics.

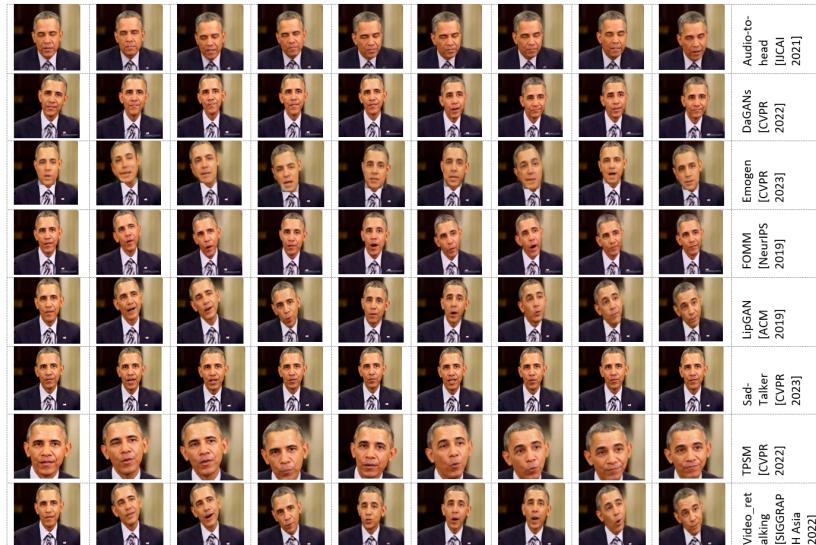


Figure 4: Comparison of images using Obama as the source image.

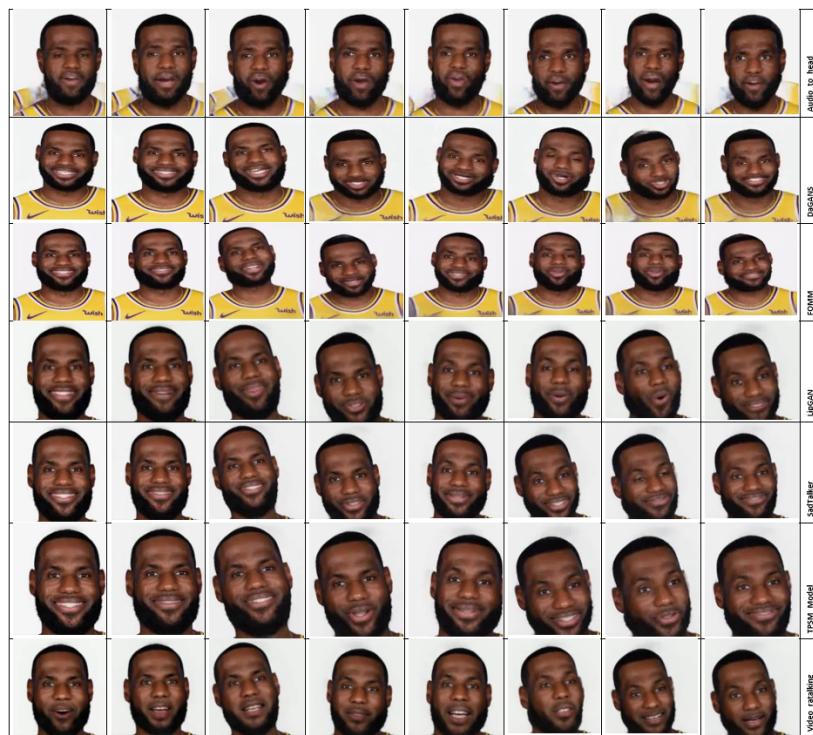


Figure 5: Comparison of images using LeBron James as the source image.



Figure 6: Comparison of images using Chris Hemsworth as the source image.



Figure 7: Comparison of images using Son as the source image.

Method	Year	Inference Time	GPU Memory	Rating
FOMM	2019	34 Sec	2 GB	3.45
LipGAN	2019	23 Sec	10.9 GB	2.76
Audio2Head	2021	20 Sec	3.9 GB	1.69
DaGANs	2022	22 Sec	4.6 GB	3.27
Video Retalking	2022	40 Sec	4.4 GB	3.84
TPSM	2022	19 Sec	1.9 GB	4.15
EmoGen	2023	25 Sec	2.8 GB	2.24
SadTalker	2023	4 min 50 Sec	4.1 GB	3.72

Table 3: Comparing some recent SOTA models on hardware requirements and visual quality perception using annotators (corresponds to column listed as Rating. We also list the GPU memory required and time needed for inference. All outputs are using a Tesla T4 GPU.

TPSM [6], Sad Talker [7], LipGAN [69], EmoGen [82] among others. The outputs for Obama can be seen in Figure 4 and for Lebron in Figure 5.

We see similar conclusions for all inputs. The worst output is seen with Audio2Head which is expected as the model is the oldest among the ones listed. Though FOMM was also released in the same year, FOMM has excellent video outputs in comparison to other approaches. Among the metrics listed in Table 3 DaGAN does best. However, visually it performs much worse than other methods with the listed outputs. We add a table comparing these methods on not just visual quality, but also on the memory requirements, inference time etc. The best outputs seem to be when the face is cropped and only facial landmarks are edited as is the case with TPSM, SadTalker and Video Re-Talking.

Based on the results, we see that despite TPSM appearing before more recent papers such as EmoGen and SadTalker, it is in fact the best performing model in terms of speed, memory and human annotator quality for the samples we generated. It is helped by the fact that this uses a reference video to generate head movements in comparison to say SadTalker which uses the audio. However, on the metrics that these papers are evaluated on TPSM performs worse. Which leads to the question what next?

6. What Next?

The recent progress in talking head generation has opened up intriguing possibilities across various application domains. However, as with any rapidly advancing technology, there are important ethical considerations and

potential challenges that must be proactively addressed. As research in this field continues to accelerate, it is critical that the community works to develop these systems responsibly and align them with human values. In the sections that follow, we survey some of the promising application areas for talking head models as well as key issues around ethics, societal impact, and areas needing further research. While there is still much work to be done, talking head generation remains an exciting field with huge potential to enable new forms of communication, creativity, and human-computer interaction. With care and foresight, researchers can ensure these systems are designed to augment human capabilities in a trustworthy and socially beneficial manner.

6.1. Application Areas

Talking head models have demonstrated potential for a wide range of applications:

- **Digital avatars and virtual assistants** - Talking heads can be used to create realistic digital avatars and virtual assistants/companions with facial expressions and lip sync. These have applications in gaming, animated films, virtual reality, and human-computer interaction.
- **Video conferencing and live streaming** - Talking heads can capture a person's face and reanimate it in real time for video chat or live streaming. This can enable eye contact, facial cues, and enhanced communication in remote interactions.
- **Synthetic media/content generation** - Talking heads enable generating synthetic footage of a person for use in films, ads, podcasts, audiobooks, and other media. This has creative applications for content generation.
- **Dubbing/translation of video content** - Talking head models can be used to dub speech in a person's original voice and facial expressions for foreign language translation in movies, TV, online lectures, etc.
- **Accessibility applications** - Talking heads hold promise for converting text/speech to sign language animations to aid the deaf and hard-of-hearing communities.

- **Telepresence robots** - Animated talking heads can give more personality and connection to telepresence robot interactions for remote tours, learning, healthcare, etc.
- **Digital deception** - While concerning, talking heads also enable creation of deepfakes and synthetic media for potentially unethical/illegal aims.

In summary, talking head generation has far-reaching potential across many fields, from creative industries to human-computer interaction and accessibility tools. While exciting, researchers must be mindful of the ethical implications as this technology is deployed.

6.2. Ethical and Societal Considerations

While talking head models unlock new creative possibilities, there are several ethical and societal concerns that warrant consideration:

- **Deepfakes and misinformation** - Talking heads make it possible to generate photorealistic videos of public figures or private citizens saying or doing things they never actually said or did. This enables creation of highly deceptive misinformation and maliciously false content at scale. Without proper safeguards, talking heads could be co-opted to sway elections, harm reputations, or mislead the public through fake news and doctored footage. Ongoing research into deepfake detection methods and watermarking authentic content is important to counteract these risks.
- **Privacy concerns** – Talking head models rely heavily on access to large datasets of personal images, videos, and voice recordings to achieve realistic results. Collecting and commercializing people's biometric data without proper consent raises significant privacy issues. More openness, transparency and providing individuals control over the use of their likeness is important for public trust. Companies and researchers have a duty to only use personal data in ethical ways that respect people's dignity.
- **Regulatory and legal issues** - Synthetic media and deepfakes exist in a legal gray area today. Policymakers are grappling with difficult questions about implementing new laws and regulations to address this

fast-moving technology. Issues like attributing manipulated footage, adapting copyright and defamation laws, and allowing legal use cases while restricting harmful ones remain open ended. Constructive public discourse between policymakers, researchers, tech companies and citizens will be key to developing fair, balanced and future-proof regulations.

- **Bias and representation** - ML biases in training data can lead talking head models to poorly represent minorities and marginalized groups. Diversity and inclusion efforts, and evaluating how systems perform across skin tones, genders, ages and ethnicities is vital to mitigate exclusion or discrimination. Democratizing access to personalized talking heads also remains an open research area.
- **Authenticity and consent** - Talking heads enable creating footage of deceased icons or public figures without consent. Synthesized videos of private citizens or unauthorized use of someone's likeness also raise ethical flags around truthfulness and individual rights. Further work is needed on authentication methods and watermarking to preserve the sanctity of genuine footage.

Overall, talking heads have profound societal impacts spanning information authenticity, privacy, ethics and legal policy issues. A robust public discussion weighing benefits against risks, and multi-disciplinary research into making these systems transparent, fair and human-centric will be key as applications continue proliferating.

6.3. Challenges and Future Direction

While rapid progress has been made, talking head generation still faces some key challenges requiring further research:

- **Improving model fidelity** - More work is needed to achieve highly realistic and natural talking head videos, including better facial movements, expressions, gaze control and pose matching.
- **Enabling controllable generation** - Current systems lack fine-grained control over synthesized footage. Improving control over attributes like speech, gaze, pose, expressions and lighting in a granular fashion remains an open problem.

- **Mitigating data bias** - Talking heads often perpetuate and amplify societal biases present in training data. Developing bias mitigation techniques and representative datasets is important for fairness.
- **High computational cost** - State-of-the-art talking head models require extensive computing resources to train and run. Reducing compute needs would improve access and widen applications.
- **Validating authenticity** - Robust digital watermarking, media forensics, and other authentication techniques need more development to reliably verify original vs synthesized footage.
- **Multimodal inputs** - Most systems currently rely on audio or text prompts. Enabling control via modalities like gestures, gaze and emotions is an intriguing area for future work.

To address these gaps, interdisciplinary efforts drawing from computer vision, graphics, HCI, ethics and other domains will be essential. Overall, responsible steering of research priorities and spurring positive applications while proactively minimizing harms will be key as talking heads continue proliferating.

7. Conclusion

In this survey, we have presented a comprehensive overview of the current state of talking head generation, systematically examining the main approaches of image-driven, audio-driven, video-driven, and 3D methods. As summarized in the abstract, our goal was to analyze the unique innovations, strengths, and limitations of different techniques for synthesizing artificial talking heads. We also empirically compared publicly available models across key aspects like speed and output quality.

Our analysis reveals remarkable progress, with recent video-driven methods approaching photorealistic results on several metrics. However, there remain areas needing improvement, including model robustness, control, and mitigating societal risks. We hope this survey provides a solid reference for practitioners seeking to leverage talking heads, while also highlighting open challenges to help guide responsible future research. While head synthesis capabilities have reached impressive levels of sophistication, there is still much

to be done before these technologies can be deployed pervasively in alignment with human values. Through proactive efforts across disciplines, the research community can steer these systems to augment human creativity in fair, ethical and enriching ways.

References

- [1] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [2] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.
- [3] S. N. Gowda, C. Yuan, Colornet: Investigating the importance of color spaces for image classification, in: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14, Springer, 2019, pp. 581–596.
- [4] C. Bregler, M. Covell, M. Slaney, Video rewrite: Driving visual speech with audio, in: Proceedings of the 24th annual conference on Computer graphics and interactive techniques, 1997, pp. 353–360.
- [5] L. Xie, Z.-Q. Liu, Realistic mouth-synching for speech-driven talking face using articulatory modelling, *IEEE Transactions on Multimedia* 9 (3) (2007) 500–510.
- [6] J. Zhao, H. Zhang, Thin-plate spline motion model for image animation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3657–3666.
- [7] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, F. Wang, Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8652–8661.

- [8] F.-T. Hong, L. Zhang, L. Shen, D. Xu, Depth-aware generative adversarial network for talking head video generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3397–3406.
- [9] F.-T. Hong, L. Shen, D. Xu, Dagan++: Depth-aware generative adversarial network for talking head video generation, arXiv preprint arXiv:2305.06225 (2023).
- [10] M. C. Doukas, S. Zafeiriou, V. Sharmanska, Headgan: One-shot neural head synthesis and editing, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 14398–14407.
- [11] S. Wang, L. Li, Y. Ding, C. Fan, X. Yu, Audio2head: Audio-driven one-shot talking-head generation with natural head motion, arXiv preprint arXiv:2107.09293 (2021).
- [12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27 (2014).
- [13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
- [14] T. Park, M.-Y. Liu, T.-C. Wang, J.-Y. Zhu, Semantic image synthesis with spatially-adaptive normalization, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 2337–2346.
- [15] Y. Choi, Y. Uh, J. Yoo, J.-W. Ha, Stargan v2: Diverse image synthesis for multiple domains, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 8188–8197.
- [16] L. Wang, W. Chen, W. Yang, F. Bi, F. R. Yu, A state-of-the-art review on image synthesis with generative adversarial networks, IEEE Access 8 (2020) 63514–63537.
- [17] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image

- translation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8789–8797.
- [18] K. Ko, T. Yeom, M. Lee, Superstargan: Generative adversarial networks for image-to-image translation in large-scale domains, *Neural Networks* 162 (2023) 330–339.
 - [19] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.
 - [20] A. Figueira, B. Vaz, Survey on synthetic data generation, evaluation methods and gans, *Mathematics* 10 (15) (2022) 2733.
 - [21] S. N. Gowda, Synthetic sample selection for generalized zero-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 58–67.
 - [22] Y. Xian, T. Lorenz, B. Schiele, Z. Akata, Feature generating networks for zero-shot learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5542–5551.
 - [23] F. H. K. D. S. Tanaka, C. Aranha, Data augmentation using gans, arXiv preprint arXiv:1904.09135 (2019).
 - [24] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, S. Narayanan, Data augmentation using gans for speech emotion recognition., in: Interspeech, 2019, pp. 171–175.
 - [25] S. N. Gowda, M. Rohrbach, F. Keller, L. Sevilla-Lara, Learn2augment: learning to composite videos for data augmentation in action recognition, in: European conference on computer vision, Springer, 2022, pp. 242–259.
 - [26] X. Dai, Y. Chen, J. Yang, P. Zhang, L. Yuan, L. Zhang, Dynamic detr: End-to-end object detection with dynamic attention, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2988–2997.

- [27] H. Li, G. Chen, G. Li, Y. Yu, Motion guided attention for video salient object detection, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 7274–7283.
- [28] X. Dai, Y. Chen, J. Yang, P. Zhang, L. Yuan, L. Zhang, Dynamic detr: End-to-end object detection with dynamic attention, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2988–2997.
- [29] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic attention, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4651–4659.
- [30] L. Huang, W. Wang, J. Chen, X.-Y. Wei, Attention on attention for image captioning, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 4634–4643.
- [31] J. Lu, C. Xiong, D. Parikh, R. Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 375–383.
- [32] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, Q. Dai, Stat: Spatial-temporal attention mechanism for video captioning, *IEEE transactions on multimedia* 22 (1) (2019) 229–241.
- [33] K. Lin, L. Li, C.-C. Lin, F. Ahmed, Z. Gan, Z. Liu, Y. Lu, L. Wang, Swinbert: End-to-end transformers with sparse attention for video captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17949–17958.
- [34] L. Gao, Z. Guo, H. Zhang, X. Xu, H. T. Shen, Video captioning with attention-based lstm and semantic consistency, *IEEE Transactions on Multimedia* 19 (9) (2017) 2045–2055.
- [35] S. Song, C. Lan, J. Xing, W. Zeng, J. Liu, An end-to-end spatio-temporal attention model for human action recognition from skeleton data, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 31, 2017.

- [36] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, M. Chiaberge, Action transformer: A self-attention model for short-time pose-based human action recognition, *Pattern Recognition* 124 (2022) 108487.
- [37] S. N. Gowda, M. Rohrbach, L. Sevilla-Lara, Smart frame selection for action recognition, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 2021, pp. 1451–1459.
- [38] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784 (2014).
- [39] L. Li, S. Wang, Z. Zhang, Y. Ding, Y. Zheng, X. Yu, C. Fan, Write-a-speaker: Text-based emotional and rhythmic talking-head generation, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 2021, pp. 1911–1920.
- [40] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, C. Xu, Talking-head generation with rhythmic head motion, in: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX*, Springer, 2020, pp. 35–51.
- [41] A. Hore, D. Ziou, Image quality metrics: Psnr vs. ssim, in: *2010 20th international conference on pattern recognition*, IEEE, 2010, pp. 2366–2369.
- [42] L. Chen, G. Cui, Z. Kou, H. Zheng, C. Xu, What comprises a good talking-head video generation?: A survey and benchmark, arXiv preprint arXiv:2005.03201 (2020).
- [43] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Nießner, Face2face: Real-time face capture and reenactment of rgb videos, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395.
- [44] W. Wu, Y. Zhang, C. Li, C. Qian, C. C. Loy, Reenactgan: Learning to reenact faces via boundary transfer, in: *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 603–619.
- [45] E. Zakharov, A. Shysheya, E. Burkov, V. Lempitsky, Few-shot adversarial learning of realistic neural talking head models, in: *Proceedings*

- of the IEEE/CVF international conference on computer vision, 2019, pp. 9459–9468.
- [46] S. Ha, M. Kersner, B. Kim, S. Seo, D. Kim, Marionette: Few-shot face reenactment preserving identity of unseen targets, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 34, 2020, pp. 10893–10900.
 - [47] O. Wiles, A. Koepke, A. Zisserman, X2face: A network for controlling face generation using images, audio, and pose codes, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 670–686.
 - [48] Q. Wang, L. Zhang, B. Li, Safa: Structure aware face animation, in: 2021 International Conference on 3D Vision (3DV), IEEE, 2021, pp. 679–688.
 - [49] S. Tripathy, J. Kannala, E. Rahtu, Facegan: Facial attribute controllable reenactment gan, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 1329–1338.
 - [50] Y. Ren, G. Li, Y. Chen, T. H. Li, S. Liu, Pirenderer: Controllable portrait image generation via semantic neural rendering, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13759–13768.
 - [51] S. Bounareli, V. Argyriou, G. Tzimiropoulos, Finding directions in gan’s latent space for neural face reenactment, arXiv preprint arXiv:2202.00046 (2022).
 - [52] S. Bounareli, C. Tzelepis, V. Argyriou, I. Patras, G. Tzimiropoulos, Stylemask: Disentangling the style space of stylegan2 for neural face reenactment, in: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), IEEE, 2023, pp. 1–8.
 - [53] M. Agarwal, R. Mukhopadhyay, V. P. Namboodiri, C. Jawahar, Audio-visual face reenactment, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 5178–5187.

- [54] G. Yao, Y. Yuan, T. Shao, K. Zhou, Mesh guided one-shot face reenactment using graph convolutional networks, in: Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 1773–1781.
- [55] Y. Wang, X. Chen, J. Zhu, W. Chu, Y. Tai, C. Wang, J. Li, Y. Wu, F. Huang, R. Ji, Hififace: 3d shape and semantic prior guided high fidelity face swapping, arXiv preprint arXiv:2106.09965 (2021).
- [56] P. Zhuang, L. Ma, S. Koyejo, A. Schwing, Controllable radiance fields for dynamic face synthesis, in: 2022 International Conference on 3D Vision (3DV), IEEE, 2022, pp. 1–11.
- [57] Y. Gao, Y. Zhou, J. Wang, X. Li, X. Ming, Y. Lu, High-fidelity and freely controllable talking head video generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5609–5619.
- [58] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, N. Sebe, First order motion model for image animation, Advances in Neural Information Processing Systems 32 (2019).
- [59] J. Tao, B. Wang, B. Xu, T. Ge, Y. Jiang, W. Li, L. Duan, Structure-aware motion transfer with deformable anchor model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3637–3646.
- [60] H. Ni, Y. Liu, S. X. Huang, Y. Xue, Cross-identity video motion retargeting with joint transformation and synthesis, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 412–422.
- [61] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, N. Sebe, Animating arbitrary objects via deep motion transfer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2377–2386.
- [62] P.-H. Huang, F.-E. Yang, Y.-C. F. Wang, Learning identity-invariant motion representations for cross-id face reenactment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7084–7092.

- [63] F. Yin, Y. Zhang, X. Cun, M. Cao, Y. Fan, X. Wang, Q. Bai, B. Wu, J. Wang, Y. Yang, Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII, Springer, 2022, pp. 85–101.
- [64] N. Drobyshev, J. Chelishev, T. Khakhulin, A. Ivakhnenko, V. Lempitsky, E. Zakharov, Megaportraits: One-shot megapixel neural head avatars, arXiv preprint arXiv:2207.07621 (2022).
- [65] J. Wang, K. Zhao, S. Zhang, Y. Zhang, Y. Shen, D. Zhao, J. Zhou, Lipformer: High-fidelity and generalizable talking face generation with a pre-learned facial codebook, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13844–13853.
- [66] J. S. Chung, A. Zisserman, Lip reading in the wild, in: Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13, Springer, 2017, pp. 87–103.
- [67] S. Suwajanakorn, S. M. Seitz, I. Kemelmacher-Shlizerman, Synthesizing obama: learning lip sync from audio, ACM Transactions on Graphics (ToG) 36 (4) (2017) 1–13.
- [68] L. Chen, Z. Li, R. K. Maddox, Z. Duan, C. Xu, Lip movements generation at a glance, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 520–535.
- [69] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, C. Jawahar, A lip sync expert is all you need for speech to lip generation in the wild, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 484–492.
- [70] J. Wang, X. Qian, M. Zhang, R. T. Tan, H. Li, Seeing what you said: Talking face generation guided by a lip reading expert, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14653–14662.
- [71] H. Zhou, Y. Liu, Z. Liu, P. Luo, X. Wang, Talking face generation by adversarially disentangled audio-visual representation, in: Proceedings

- of the AAAI conference on artificial intelligence, Vol. 33, 2019, pp. 9299–9306.
- [72] L. Chen, R. K. Maddox, Z. Duan, C. Xu, Hierarchical cross-modal talking face generation with dynamic pixel-wise loss, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 7832–7841.
 - [73] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, D. Li, MakelTalk: speaker-aware talking-head animation, ACM Transactions On Graphics (TOG) 39 (6) (2020) 1–15.
 - [74] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, M. Nießner, Neural voice puppetry: Audio-driven facial reenactment, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16, Springer, 2020, pp. 716–731.
 - [75] Y. Sun, H. Zhou, Z. Liu, H. Koike, Speech2talking-face: Inferring and driving a face with synchronized audio-visual representation., in: IJCAI, Vol. 2, 2021, p. 4.
 - [76] Y. Lu, J. Chai, X. Cao, Live speech portraits: real-time photorealistic talking-head animation, ACM Transactions on Graphics (TOG) 40 (6) (2021) 1–17.
 - [77] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, C. Xu, Talking-head generation with rhythmic head motion, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX, Springer, 2020, pp. 35–51.
 - [78] C. Zhang, Y. Zhao, Y. Huang, M. Zeng, S. Ni, M. Budagavi, X. Guo, Facial: Synthesizing dynamic talking face with implicit attribute learning, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 3867–3876.
 - [79] X. Ji, H. Zhou, K. Wang, Q. Wu, W. Wu, F. Xu, X. Cao, Eamm: One-shot emotional talking face via audio-based emotion-aware motion model, in: ACM SIGGRAPH 2022 Conference Proceedings, 2022, pp. 1–10.

- [80] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, X. Cao, F. Xu, Audio-driven emotional video portraits, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 14080–14089.
- [81] B. Liang, Y. Pan, Z. Guo, H. Zhou, Z. Hong, X. Han, J. Han, J. Liu, E. Ding, J. Wang, Expressive talking head generation with granular audio-visual control, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3387–3396.
- [82] S. Goyal, S. Uppal, S. Bhagat, Y. Yu, Y. Yin, R. R. Shah, Emotionally enhanced talking face generation, arXiv preprint arXiv:2303.11548 (2023).
- [83] Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, K. Singh, Visemenet: Audio-driven animator-centric speech animation, ACM Transactions on Graphics (TOG) 37 (4) (2018) 1–10.
- [84] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, Z. Liu, Pose-controllable talking face generation by implicitly modularized audio-visual representation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 4176–4186.
- [85] D. Min, M. Song, S. J. Hwang, Styletalker: One-shot style-based audio-driven talking head video generation, arXiv preprint arXiv:2208.10922 (2022).
- [86] L. Li, S. Wang, Z. Zhang, Y. Ding, Y. Zheng, X. Yu, C. Fan, Write-a-speaker: Text-based emotional and rhythmic talking-head generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 1911–1920.
- [87] G. Gafni, J. Thies, M. Zollhofer, M. Nießner, Dynamic neural radiance fields for monocular 4d facial avatar reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8649–8658.
- [88] P.-W. Grassal, M. Prinzler, T. Leistner, C. Rother, M. Nießner, J. Thies, Neural head avatars from monocular rgb videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18653–18664.

- [89] Y. Zheng, V. F. Abrevaya, M. C. Bühler, X. Chen, M. J. Black, O. Hilliges, Im avatar: Implicit morphable head avatars from videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13545–13555.
- [90] T. Khakhulin, V. Sklyarova, V. Lempitsky, E. Zakharov, Realistic one-shot mesh-based head avatars, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II, Springer, 2022, pp. 345–362.
- [91] W. Li, L. Zhang, D. Wang, B. Zhao, Z. Wang, M. Chen, B. Zhang, Z. Wang, L. Bo, X. Li, One-shot high-fidelity talking-head synthesis with deformable neural radiance field, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 17969–17978.
- [92] Y. Ma, S. Wang, Y. Ding, B. Ma, T. Lv, C. Fan, Z. Hu, Z. Deng, X. Yu, Talkclip: Talking head generation with text-guided expressive speaking styles, arXiv preprint arXiv:2304.00334 (2023).
- [93] A. Nagrani, J. S. Chung, A. Zisserman, Voxceleb: A large-scale speaker identification dataset, Proc. Interspeech 2017 (2017) 2616–2620.
- [94] J. S. Chung, A. Nagrani, A. Zisserman, Voxceleb2: Deep speaker recognition, Proc. Interspeech 2018 (2018) 1086–1090.
- [95] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, C. C. Loy, Mead: A large-scale audio-visual dataset for emotional talking-face generation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI, Springer, 2020, pp. 700–717.
- [96] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, R. Verma, Crema-d: Crowd-sourced emotional multimodal actors dataset, IEEE transactions on affective computing 5 (4) (2014) 377–390.
- [97] M. Cooke, J. Barker, S. Cunningham, X. Shao, An audio-visual corpus for speech perception and automatic speech recognition, The Journal of the Acoustical Society of America 120 (5) (2006) 2421–2424.

- [98] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, E. M. Provost, Msp-improv: An acted corpus of dyadic interactions to study emotion perception, *IEEE Transactions on Affective Computing* 8 (1) (2016) 67–80.
- [99] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, X. Chen, Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild, in: 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019), IEEE, 2019, pp. 1–8.
- [100] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1–11.
- [101] Z. Zhang, L. Li, Y. Ding, C. Fan, Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3661–3670.
- [102] E. Zakharov, A. Ivakhnenko, A. Shysheya, V. Lempitsky, Fast bi-layer neural synthesis of one-shot realistic head avatars, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16, Springer, 2020, pp. 524–540.
- [103] T.-C. Wang, A. Mallya, M.-Y. Liu, One-shot free-view neural talking-head synthesis for video conferencing, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 10039–10049.
- [104] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, S. Tulyakov, Motion representations for articulated animation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13653–13662.