

Efficient Video Portrait Reenactment via Grid-based Codebook

Kaisiyuan Wang
The University of Sydney
Sydney, Australia
kaisiyuan.wang@sydney.edu.au

Hang Zhou*
Baidu Inc.
Shanghai, China
zhouhang09@baidu.com

Qianyi Wu
Monash University
Melbourne, Australia
qianyi.wu@monash.edu

Jiaxiang Tang
Peking University
Beijing, China
tjx@pku.edu.cn

Zhiliang Xu
Baidu Inc.
Shenzhen, China
xuzhiliang@baidu.com

Borong Liang
Baidu Inc.
Shenzhen, China
liangborong@baidu.com

Tianshu Hu
Baidu Inc.
Shanghai, China
hutianshu01@baidu.com

Errui Ding
Baidu Inc.
Beijing, China
dingerrui@baidu.com

Jingtuo Liu
Baidu Inc.
Beijing, China
liujingtuo@baidu.com

Ziwei Liu
Nanyang Technological University
Singapore, Singapore
zwlui.hust@gmail.com

Jingdong Wang
Baidu Inc.
Beijing, China
wangjingdong@outlook.com

ABSTRACT

While progress has been made in the field of portrait reenactment, the problem of how to efficiently produce high-fidelity and accurate videos remains. Recent studies build direct mappings between driving signals and their predictions, leading to failure cases when synthesizing background textures and detailed local motions. In this paper, we propose the Video Portrait via Grid-based Codebook (VPGC) framework, which achieves efficient and high-fidelity portrait modeling. Our key insight is to query driving signals in a position-aware textural codebook with an explicit grid structure. The grid-based codebook stores delicate textural information locally according to our observations on video portraits, which can be learned efficiently and precisely. We subsequently design a Prior-Guided Driving Module to predict reliable features from the driving signals, which can be later decoded back to high-quality video portraits by querying the codebook. Comprehensive experiments are conducted to validate the effectiveness of our approach.

CCS CONCEPTS

• **Computing methodologies** → **Animation**; *Neural networks*.

KEYWORDS

Facial Animation, Video Synthesis

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGGRAPH '23 Conference Proceedings, August 06–10, 2023, Los Angeles, CA, USA
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0159-7/23/08...\$15.00
<https://doi.org/10.1145/3588432.3591509>



Figure 1: Qualitative results of our method. We focus on high-fidelity personalized video portrait reenactment, which animates a target portrait according to the input driving videos. The synthesized results are supposed to have the same mouth shapes, facial expressions, and head poses as the driving videos. The figures are selected from the HDTF [Zhang et al. 2021] dataset ©Attribution 4.0 International (CC BY 4.0).

ACM Reference Format:

Kaisiyuan Wang, Hang Zhou, Qianyi Wu, Jiaxiang Tang, Zhiliang Xu, Borong Liang, Tianshu Hu, Errui Ding, Jingtuo Liu, Ziwei Liu, and Jingdong Wang. 2023. Efficient Video Portrait Reenactment via Grid-based Codebook. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings (SIGGRAPH '23 Conference Proceedings)*, August 06–10, 2023, Los Angeles, CA, USA. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3588432.3591509>

1 INTRODUCTION

Video portrait reenactment aims to animate a target portrait with similar movements with input driving videos. It has received considerable attention due to its increasing applications in real-world scenarios, such as filmmaking, computer game, virtual avatar creation, and multimedia entertainment. A majority of recent studies [Doukas et al. 2021; Drobyshev et al. 2022; Khakhulin et al. 2022; Siarohin et al. 2019; Sun et al. 2022b; Wang et al. 2021, 2022; Yang et al. 2022; Zakharov et al. 2020, 2019; Zhou et al. 2019, 2021] focus on animating a portrait with only one or few face images. However, they normally fail to preserve the identity of the target portrait and lead to visible artifacts.

On the other hand, researchers have also been committed to modeling realistic personalized portraits [Gafni et al. 2021; Guo et al. 2021; Kim et al. 2019, 2018; Suwajanakorn et al. 2017; Thies et al. 2016; Wang et al. 2020; Wu et al. 2018]. Their paradigm can usually be summarized as *directly* building the mapping between the driving guidance (e.g., 2D landmarks or 3D morphable parameters) and the corresponding texture. Although different techniques, including 2D generative models [Goodfellow et al. 2020; Wu et al. 2018], neural rendering [Kim et al. 2018], and neural radiance fields (NeRF) [Gafni et al. 2021; Guo et al. 2021; Mildenhall et al. 2020] are incorporated, their generated results usually suffer from perceptual degradation, unstable textures, or over-smooth issues, leading to unrealistic video portraits.

Recently, VQGAN [Esser et al. 2021a] has shown success in various generative tasks [Chang et al. 2023, 2022; Gu et al. 2022; Liu et al. 2022; Zheng et al. 2022b; Zhou et al. 2022]. It stores high-quality local textural information in a pre-learned codebook and queries tokens during inference instead of direct prediction. Such practice has been successfully adopted in face restoration studies [Gu et al. 2022; Zhou et al. 2022], which shares certain similarities with our goal of driving high-fidelity portraits.

Inspired by previous studies, we propose to integrate codebook learning into high-fidelity personalized portrait reenactment. One straightforward idea is to build one personalized codebook for each target portrait based on captured monocular videos. However, two major problems severely influence the modeling efficiency and accuracy: 1) The learning of the codebook and the auto-encoder network is extremely time-consuming. It takes more than 36 hours to model a 4-minute portrait video; 2) Local texture with similar patterns might be recognized as the same code due to the quantization error, which possibly leads to inaccurate portrait modeling.

In this paper, we propose the **Video Portrait via Grid-based Codebook (VPGC)** framework, which produces high-fidelity video portraits with high efficiency and vivid details. Our key insight is *to query driving signals in a position-aware textural codebook with an explicit grid structure*. Specifically, the grid-based codebook is designed according to the spatial distribution of portrait images, where each local region roughly corresponds to a specific semantic (e.g., the upper part for hair, the middle part for face, and the bottom part for neck or torso). We thus correlate spatial positions in the portrait space with a relatively small number of codes in the grid. In this way, the traditional codebook where textures are stored in an unordered manner is reformulated into a grid structure of three

dimensions, which shares similarities with recent efficient volume rendering approaches [Chen et al. 2022; Müller et al. 2022].

We identify several advantages of our grid-based codebook beyond the vanilla codebook [Esser et al. 2021a]. 1) The learning of our grid-based codebook can be intuitively regarded as dividing the vanilla learning procedure into sub-tasks at each spatial position. This naturally accelerates the convergence process and improves accuracy; 2) We are able to perform more fine-grained portrait modeling at each position by adopting a more flexible *Soft-Indexing* strategy, which can effectively alleviate inaccurate modeling caused by quantization [Esser et al. 2021a]. 3) The codebook itself can further facilitate reenactment training as prior information.

We then present a Prior-Guided Driving Module which maps the simplest driving signals of 2D landmarks to realistic portraits. As the grid-based codebook shares the same spatial structure with the encoded driving information, it can serve as a reliable texture prior and provide essential global contextual information to the driving stage. Specifically, the texture prior provides complementary information for the sparse 2D landmarks, which enables faster and more stable training. Furthermore, a global discriminative feature is leveraged to alleviate ambiguity in regions without specific guidance. Finally, the enhanced driving features query our learned grid-based codebook for recovering high-fidelity portraits.

Our contributions are summarized as follows:

- We propose the Video Portrait via Grid-based Codebook (VPGC) framework to achieve efficient video portrait reenactment with fine-grained details.
- We delicately design the grid-based codebook with outstanding properties that benefit both the portrait modeling process and the subsequent driving procedure.
- We propose a Prior-Guided Driving Module to mitigate uncertainty and enhance the learning procedure of the mapping between the driving guidance and portrait texture.

2 RELATED WORKS

Face Reenactment. Recent studies [Siarohin et al. 2019; Wang et al. 2021, 2022; Zakharov et al. 2019] proposed to achieve face reenactment by extracting the motion representation from the driving video and applying it to the target image. These approaches tend to generate video portraits based on only one or a few frames of the target portrait in a warping-based paradigm. However, these methods usually suffer from identity distortion and low generation quality. To achieve stable and photo-realistic video portraits, the structural representation of human-like 3D face morphable model [Blanz and Vetter 1999] is explicitly used for driving portraits. Some early attempts [Kim et al. 2019, 2018] focus on developing personalized models which rely on the 3DMM face model [Blanz and Vetter 1999] for human head rendering and a 2D generative model for the torso and background synthesis. While LSP [Lu et al. 2021] proposes to use 2D facial landmarks projected from 3D geometry as the driving guidance, its generated results are quite sensitive to the input landmarks. The latest studies [Gafni et al. 2021; Grassal et al. 2022; Guo et al. 2021; Zheng et al. 2022a] take advantage of neural radiance field to produce high-fidelity video portraits through volume rendering. Nerface [Gafni et al. 2021] builds a talking head system by combining a dynamic radiance field with a low-dimensional

morphable model. Similarly, NHA [Grassal et al. 2022] also presents a hybrid representation, including a morphable model and two feed-forward networks for vertex offset and expression texture prediction, which consumes much time on pre-processing and joint optimization.

Our VPGC avoids the time-consuming volume rendering as well as tedious pre-processing and develops an efficient framework for high-quality face reenactment by using the simplest 2d landmarks as driving signals.

Quantization-based Image Modeling. Image modeling has recently achieved significant progress after taking inspiration from Transformer networks. VQVAE [Van Den Oord et al. 2017] and VQGAN [Esser et al. 2021a], as the pioneer of quantized image modeling, have received extensive attention from multiple works on high-resolution image synthesis and editing.

Quantized image modeling is performed via an autoencoder network and a learnable codebook. The key insight of these approaches is to replace the discrete representations from the encoder with the codes queried from the learned codebook. VQGAN [Esser et al. 2021a] and ImageBART [Esser et al. 2021b] leverage a transformer to synthesize images in an auto-regressive manner, while MaskGIT [Chang et al. 2022] proposes to model an image from multiple directions instead of the sequential prediction as in VQGAN. CodeFormer [Zhou et al. 2022] achieves remarkable face restoration by learning a high-quality texture dictionary and building a reliable mapping with much less uncertainty. VQFR [Gu et al. 2022] designs a parallel decoder to replace the commonly used transformer for reconstructing high-fidelity human face details.

Despite the success achieved by these methods, all of them require large amounts of time for training due to the unordered learnable codebook. Differently, we propose a novel codebook with an explicit grid structure to store local textures for neighboring regions and enable efficient and high-quality video portrait modeling.

3 METHODOLOGY

The overview of our proposed Video Portrait via Grid-based Codebook (VPGC) framework is illustrated in Fig 2, where our framework performs high-quality video portrait modeling and reenactment training. In the following, we first introduce the formulation of our task and review the preliminaries of VQGAN in Sec. 3.1. Then grid-based codebook learning and soft indexing are demonstrated in Sec. 3.2. The Prior-Guided Driving Module is introduced in Sec. 3.3.

3.1 Task Formulation and Preliminaries

Task Formulation. For each target portrait video $V = \{I_1, \dots, I_T\}$, we leverage the 3D reconstruction approach DECA [Feng et al. 2021] to produce a 3D parametric head with morphable parameters (i.e., shape, pose, and expression) and project the 3D driving landmarks as 2D heatmaps $V^l = \{I_1^l, \dots, I_T^l\}$. When animating the target portrait with a different driving portrait at the inference stage, we replace the shape parameter of the driving portrait with that of the target portrait to reproduce the driving heatmaps.

While traditional methods directly build the mapping between driving landmarks and portrait images, we propose to involve an intermediate codebook for producing results of higher quality. The whole training paradigm is divided into the training of the codebook

and the mapping between driving signals and codebook embeddings. During inference, we map the driving signals to a set of embeddings by querying the learned codebook, which are later translated back to the image domain through the pre-trained decoder.

Preliminaries of VQGAN. VQGAN [Esser et al. 2021a] is able to synthesize high-resolution images by compositing discrete quantized codes from a learnable codebook. Specifically, it is composed of two convolutional networks denoted as the Encoder E_n and the Decoder D_e . The Encoder E_n first encodes the input image $I \in \mathbb{R}^{H \times W \times 3}$ into a feature map $Z \in \mathbb{R}^{h \times w \times d}$. Then the quantization process forces each of its vectors $Z_{(u,v)} \in \mathbb{R}^d$ to search the closest code from the learnable codebook $C = \{c_i \in \mathbb{R}^d\}_{i=0}^{N-1}$ via minimizing feature distance:

$$a_{(u,v)} = \arg \min_i \|Z_{(u,v)} - c_i\|_2, \quad (1)$$

where (u, v) denotes the coordinates on the feature map. In this way, I is encoded into a set of discrete indices, which are subsequently de-quantized back to the feature map $Z_q = \{c_{a_{(1,1)}}, \dots, c_{a_{(u,v)}}\}$ via looking up the codebook.

To achieve end-to-end training, a differentiable loss function is adopted by copying the gradients from the decoder to the encoder:

$$\mathcal{L}_{VQ} = \|\mathbf{sg}[Z] - Z_q\|_2 + \|\mathbf{sg}[Z_q] - Z\|_2. \quad (2)$$

Here $\mathbf{sg}[\cdot]$ denotes the “stop gradient” operation. In terms of the image generation quality, the L_1 reconstruction loss $\mathcal{L}_{L1}$, adversarial loss $\mathcal{L}_{adv}$ [Wang et al. 2018], and perceptual loss [Johnson et al. 2016] $\mathcal{L}_{per}$ are leveraged in the self-reconstruction training. The total loss function for codebook learning is described as follows:

$$\mathcal{L}_{cb} = \mathcal{L}_{L1} + \mathcal{L}_{per} + \mathcal{L}_{VQ} + \lambda \mathcal{L}_{adv}. \quad (3)$$

3.2 Grid-based Codebook and Soft Indexing

As stated in Sec. 1, our goal is to design a more efficient portrait modeling framework. Recently, volumetric rendering [Chen et al. 2022; Müller et al. 2022; Sun et al. 2022a; Tang et al. 2022] studied to store 3D scene features in a grid structure for training acceleration. Moreover, we observe that in portrait videos, the semantics of different regions roughly keep the same in the 2D plane, which further inspires us to correlate codebook embeddings with spatial information. Thus we propose a grid-based codebook learning procedure to achieve more efficient and accurate prior learning. The meanings of the notations below are kept unchanged as in Sec. 3.1.

Grid-based Codebook. We design our codebook $G \in \mathbb{R}^{h \times w \times K_c \times d_c}$ with an explicit grid structure, which shares the same spatial dimension $h \times w$ with the encoded feature map Z . Different from the codebook $C \in \mathbb{R}^{N \times d}$ in VQGAN, which is an orderless data structure without clear semantics, our codebook stores K_c embeddings of length d_c at each grid position (u, v) . Here, the parameters used in VQGAN (N, d, h, w) are set as (1024, 256, 16, 16), while our newly proposed parameters (K_c, d_c) are set as (256, 4) by default.

Note that the mission of our grid-based codebook is just to reorganize the vanilla unordered code learning [Esser et al. 2021a] into an ordered procedure. It actually shares the *same numbers of parameters* as the vanilla codebook in VQGAN (i.e., $N \times d = h \times w \times K_c \times d_c$) without consuming extra memory. In this case, the query vector

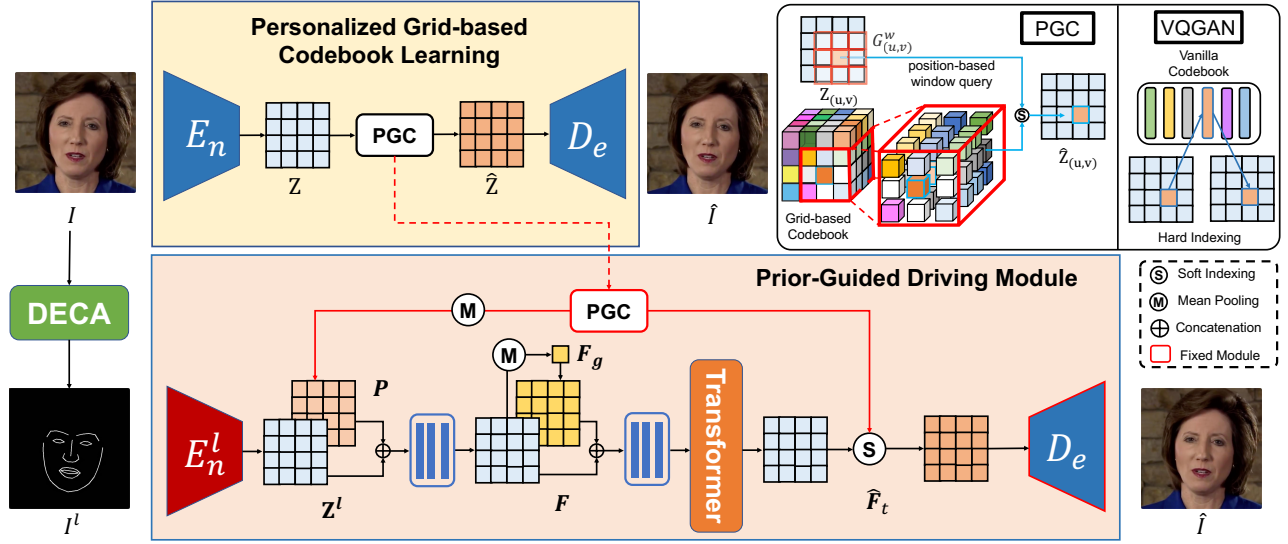


Figure 2: Overview of our VPGC framework. To store delicate textual information for the target portrait, we first build a Personalized Grid-based Codebook as shown in the upper part. The novel codebook learning strategy employs a position-aware grid structure and soft indexing to replace the vanilla unordered codebook and hard indexing used in VQGAN [Esser et al. 2021a]. By taking advantage of the textural prior from the pre-trained Personalized Grid-based Codebook, we subsequently propose a Prior-Guided Driving Module to recover high-fidelity video portraits from the input driving signal. The portrait is selected from the HDTF [Zhang et al. 2021] dataset (CC BY 4.0).

$Z_{(u,v)}$ can be represented as $K_d = d/d_c$ embeddings within $G_{(u,v)}$, where K_d is an integer. Thus $Z_{(u,v)}$ can be formulated into a matrix $Z_{(u,v)}^s \in \mathbb{R}^{K_d \times d_c}$.

Receptive Window. Considering that our grid-based codebook may overemphasize the texture synthesis at each local position while neglecting the correlation with neighboring regions, we adopt a receptive window with the size of $r \times r$ to enlarge the receptive field from a single position to a local neighborhood W around the target position.

Our intuition is that the $r \times r$ region around $G_{(u,v)}$, which consists of $r^2 \cdot K_c$ embeddings of length d_c , should be able to capture all possible textures appearing at the querying position (u,v) . All embeddings within this region can be formulated as a matrix $G_{(u,v)}^W \in \mathbb{R}^{r^2 \cdot K_c \times d_c}$, where r is set as 4 by default.

Soft Indexing Strategy. Since VQGAN quantizes feature vectors to integral tokens, textures with similar patterns may easily be recognized as the same embedding, which causes inevitable loss to the representative power of VQGAN. To handle this problem, we remove the quantization process and propose a soft indexing strategy to achieve more accurate portrait modeling.

The idea is to integrate all embeddings within the local codebook $G_{(u,v)}^W$ based on the attention mechanism according to their similarity with the query matrix $Z_{(u,v)}^s$. Particularly, we calculate the similarity matrix $M_{(u,v)}$ between $Z_{(u,v)}^s$ and $G_{(u,v)}^W$, which is defined as:

$$M_{(u,v)} = \text{softmax}(Z_{(u,v)}^s * G_{(u,v)}^{W\top}), \quad (4)$$

where the softmax operation is conducted on each row of $M_{(u,v)} \in \mathbb{R}^{K_d \times r^2 \cdot K_c}$, and $*$ denotes the matrix multiplication. Finally, we can represent the separated vector $Z_{(u,v)}^s$ with sufficient embeddings in the local dictionary $G_{(u,v)}$ via weighted summation rather than querying a fixed embedding:

$$\hat{Z}_{(u,v)} = M_{(u,v)} G_{(u,v)}. \quad (5)$$

Grid-based Codebook Learning. The training of the grid-based codebook is identical to the vanilla codebook as described in Sec. 3.1. The training objectives in Eq. 3 are all applied in training our grid-based codebook.

3.3 Prior-Guided Driving Module

After building the personalized texture dictionary, we move forward to the problem of how to learn the mapping between the driving guidance and the portrait texture. Intuitively, we start by employing an encoder E_n^l with similar architecture as E_n to encode the driving heatmaps I^l into feature map Z^l , where Z is used as supervision.

Nevertheless, we observe two interesting issues during the empirical exploration. 1) The convergence process is sometimes unstable at the early training stage, which easily leads to sub-optimal performance; 2) Unexpected jittering or artifacts occasionally occurs at the regions without landmark guidance, such as hair boundary and other clothing accessories (e.g., earrings and collars). We attribute these problems to insufficient guidance from both texture and geometry information, since the input signal is degraded from detailed portrait images to sparse facial landmarks.

To tackle these problems, we propose two feature map enhancement strategies, respectively, for alleviating the domain gap between the driving heatmaps and the portrait texture.

Texture Prior from Grid-based Codebook. As the input heatmaps can only provide limited geometry guidance from sparse landmarks, it is challenging to build a direct mapping between the input signal and the portrait texture, which leads to unstable convergence in the training process, especially at the beginning phase. Considering that our grid-based codebook shares the same spatial structure with the feature map Z^l and Z , the local texture dictionary at each position $G_{(u,v)}$ can intuitively serve as a reliable and easily accessible texture prior. In practice, we obtain the texture prior for this position $P_{(u,v)} = \text{mean}(G_{(u,v)})$ by averaging all codebook embeddings in $G_{(u,v)}$, where mean denotes the mean pooling operation. Given this reliable texture prior $P_{(u,v)} \in \mathbb{R}^{d_c}$, we concatenate it with the feature vector $Z_{(u,v)}^l$ to produce a texture-enhanced feature vector via a three-layer MLP network, which is defined as:

$$F_{(u,v)} = \text{MLP}(Z_{(u,v)}^l \oplus P_{(u,v)}), \quad (6)$$

where $\oplus$ denotes the channel-wise concatenation.

Discriminative Geometry Guidance. In terms of the temporal inconsistency in some regions, we suppose it is owing to the ambiguity caused by lacking clear guidance. Particularly, the movement information of the non-facial regions (e.g., forehead, hair, and torso) is not covered by the driving heatmaps. To cope with this issue, we propose to enlarge the difference of these uncovered regions between frames by adding additional discriminative features. Specifically, we employ mean pooling operation on the spatial dimension of F to extract global context $F_g \in \mathbb{R}^d$ for this frame, which is subsequently attached to F at each position. Up to this point, we finally obtain the feature map $\hat{F} \in \mathbb{R}^{h \times w \times d}$ enhanced by both texture and geometry priors by using another MLP network:

$$\hat{F}_{(u,v)} = \text{MLP}(F_{(u,v)} \oplus F_g). \quad (7)$$

In addition, we follow VQGAN [Esser et al. 2021a] to leverage a three-layer vision transformer module [Dosovitskiy et al. 2021] for further improving the hallucination modeling ability of our framework. The output of the transformer module $\hat{F}_t$ will be fed into our grid-based codebook and finally projected back into a high-quality portrait image via the pre-trained decoder network.

Training Objective. Since both the grid-based codebook and the decoder network aforementioned are well-trained in advance, our training objective is to reduce the difference between the final feature map of driving landmarks $\hat{F}_t$ and its corresponding ground-truth Z , where the loss function is defined based on L_1 loss:

$$\mathcal{L}_{feat} = \lambda \|\hat{F}_t - Z\|_1, \quad (8)$$

where λ is set as 10 by default. Note that Z denotes the feature map of the target portrait image encoded by the pre-trained encoder.

4 EXPERIMENTS

4.1 Experiment Settings

Dataset and Pre-processing. We use eight portrait videos captured by static cameras as our training dataset, which includes one

video from AD-Nerf [Guo et al. 2021] dataset, one video from Nerface [Gafni et al. 2021] dataset and six videos from the HDTF [Zhang et al. 2021] dataset. In terms of pre-processing, we extract frames from each video with the frame rate of 60 and crop 512×512 portrait images out of the original frames. For evaluation, we randomly select 1000 consecutive frames (i.e., around 17 seconds in length) from each test video clip as the test set.

Implementation Details. Our models are trained using Adam optimizer with an initial learning rate of $5e-5$ and batch size of 4. For a 5-minute video, it normally takes 5 hours to train the grid-based codebook and 2 hours for the driving module on a 40G NVIDIA A100 GPU. For inference, our VPGC can run at 29 fps at the cost of 5.6 GB GPU memories. Please refer to our supplementary for more details on implementation and network architectures.

Comparison Methods. We compare our VPGC with three person-agnostic methods and four person-specific methods. The person-agnostic methods consist of three latest state-of-the-art methods, LIA [Wang et al. 2022], Facev2v [Wang et al. 2021] and Style-Heat [Yin et al. 2022]. In terms of the person-specific methods, we choose a 2D generative model method LSP [Lu et al. 2021] and four 3D-based methods including neural Head Avatar (NHA) [Gras-sal et al. 2022], IMAvatar [Zheng et al. 2022a], AD-Nerf [Guo et al. 2021] and Nerface [Gafni et al. 2021] as our counterparts.

4.2 Quantitative Evaluation

Comparison Setting. We perform quantitative evaluation under the self-reenactment setting on four datasets (denoted as Testset A, B, C and Nerface dataset). Note that these comparison methods may have different requirements for cropping and alignment, we thus crop the shared facial parts detected by landmark detectors and resize them to the same size for comparison.

Evaluation Metrics. Similar to [Gafni et al. 2021], we use standard metrics PSNR, SSIM, and LPIPS to evaluate the quality of the predicted results. Furthermore, we follow [Ji et al. 2022, 2021] to adopt landmark distance on the whole face (F-LMD) for synchronization assessment, which considers both lip-sync and head pose accuracy.

Evaluation Results. The quantitative results on HDTF and Nerface datasets are summarized in Tab 1 and 2. According to the results in Tab 1, our VPGC shows superior performance in terms of all the metrics on generation quality beyond the counterparts due to the well-learned texture codebook. On the other hand, our F-LMD value is much lower than those of person-agnostic methods and comparable to the state-of-the-art personalized method LSP, which indicates that our VPGC achieves satisfactory synchronization.

In terms of the results on the Nerface dataset, neither Nerface nor DVP [Gafni et al. 2021; Kim et al. 2018] is able to synthesize satisfactory human face images, leading to much worse performance when compared with our VPGC. Please also refer to our supplementary video for visual results.

4.3 Qualitative Evaluation

We also perform qualitative experiments under the cross-reenactment setting along with a user study to subjectively demonstrate the differences between our method and its counterparts. Since AD-Nerf

Table 1: The quantitative results of on Testset A, B, and C. We compare our VPGC against recent SOTA methods [Grassal et al. 2022; Guo et al. 2021; Lu et al. 2021; Wang et al. 2021, 2022; Zheng et al. 2022a] under self-reenactment setting in terms of four metrics. For LPIPS and F-LMD the lower the better, and the higher the better for other metrics.

Methods	Testset A				Testset B				Testset C			
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	F-LMD $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	F-LMD $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	F-LMD $\downarrow$
Ground Truth	N/A	1.00	0	0	N/A	1.00	0	0	N/A	1.00	0	0
Facev2v	30.16	0.70	0.16	2.52	29.58	0.67	0.14	2.47	30.99	0.78	0.09	2.45
LIA	28.79	0.71	0.22	2.51	29.41	0.66	0.18	2.82	30.22	0.74	0.13	3.03
StyleHeat	28.40	0.62	0.24	2.63	27.91	0.60	0.20	2.72	28.22	0.64	0.16	3.16
NHA	30.49	0.67	0.28	2.22	29.79	0.68	0.17	2.14	30.84	0.72	0.24	2.23
IMAvatar	29.77	0.65	0.31	2.96	29.20	0.64	0.22	3.01	29.89	0.68	0.27	2.75
AD-Nerf	30.30	0.67	0.18	3.60	29.45	0.64	0.20	3.49	30.41	0.67	0.15	2.84
LSP	31.26	0.76	0.08	2.04	30.39	0.70	0.10	2.22	31.44	0.79	0.07	2.04
VPGC	32.19	0.78	0.07	2.04	31.86	0.72	0.09	2.19	32.26	0.81	0.06	2.05

Table 2: The quantitative comparison with previous methods [Gafni et al. 2021; Kim et al. 2018] on the Nerface dataset.

Methods	Nerface Dataset			
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	F-LMD $\downarrow$
Ground Truth	N/A	1.000	0	0
DVP	28.10	0.71	0.36	4.59
Nerface	29.58	0.77	0.24	3.88
VPGC	33.17	0.85	0.07	2.16

requires a template video from the same portrait as the pose input, we do not involve it in our comparison.

Evaluation Results. The key frames from two video clips are illustrated in Fig. 3. Although Facev2v and LIA achieve satisfying movements in the animation, their results are quite dependent on the selected source images, which always cause blurry and undesirable texture. StyleHeat fails to preserve the identity information where main facial components (e.g., eyes, nose, and mouth) all suffer from severe distortion. In terms of person-specific methods, NHA and IMAvatar tend to produce over-smooth textures without essential components (e.g., teeth) and require time-consuming pre-processing. LSP fails to create reasonable textures at some local regions (e.g., face boundary and collar) due to insufficient guidance. Our VPGC can generate fine-grained details and achieve satisfactory synchronization simultaneously.

User Study. We also invite 15 users to participate in a subjective evaluation of the cross-reenactment results generated by our VPGC and other comparison methods. The participants are aged from 18 to 30 years old and come from universities or research institutes. All of them are required to rate the generated video portraits from three aspects following the Mean Opinion Scores rating protocol: 1) Generation Quality; 2) Video Realness; 3) Synchronization. The rating ranges from 1 (worst) to 5 (best).

The results are reported in Tab 3. Our VPGC outperforms all its counterparts in terms of generation quality and video realness, since

Table 3: User study results based on Mean Opinion Scores. The rating is from 1 to 5, the higher the better.

Methods	Generation Quality	Video Realness	Synchronization
LIA	3.60	4.53	4.46
Facev2v	3.86	4.66	4.53
StyleHeat	2.67	1.73	4.33
NHA	3.33	3.20	3.80
IMAvatar	3.20	2.67	3.07
LSP	3.73	3.40	4.06
VPGC (Ours)	4.93	4.80	4.40

our generated results have shown more fine-grained texture details and fewer dynamic artifacts in the animating process thanks to the personalized texture codebook and the carefully designed prior-guided driving module. In terms of synchronization, the person-agnostic methods achieve only a slight advantage over our method because of their more fluent lip movements. Overall, the users prefer our results in more aspects.

4.4 Ablation Study

We take Testset A as an example to perform ablation studies on the contributions of proposed strategies used in our VPGC, which focus on the usage of different codebooks in portrait modeling and different priors in the driving module.

Analysis on the different Codebooks. We first investigate the outstanding properties of our grid-based codebook in portrait modeling. Specifically, we train two different models by using the vanilla codebook and our grid-based codebook and evaluate the results of these two models saved at different training iterations in the early stage. The quantitative and qualitative results are shown in Tab 4 and Fig. 4, which demonstrate that our grid-based codebook enables faster and more stable portrait modeling compared to the vanilla one. For a 5-minute portrait video, it takes vanilla codebook learning 80,000 iterations (i.e., nearly 36 hours) to complete detailed texture learning (e.g., blinks in Fig. 4), while our grid-based codebook learning procedure saves *more than 80% of its time-cost*, which



Figure 3: Qualitative comparison between our method and its counterparts. We compare our VPGC with the recent state-of-the-art methods [Grassal et al. 2022; Lu et al. 2021; Wang et al. 2021, 2022; Yin et al. 2022; Zheng et al. 2022a] under cross-reenactment setting. Please zoom in for better visualization. Both the two driving videos and the target portrait video in the left column are selected from the HDTF [Zhang et al. 2021] dataset (CC BY 4.0), while the target portrait Obama in the right column is from © White House Attribution 3.0 United States (CC BY 3.0 US).

takes only 5 hours on average to finish the texture learning. Please refer to the supplementary video for better visualization.

Analysis on the different priors in Driving Module. We continue to evaluate the contributions of the texture prior P and the discriminative geometry prior F_g used in our driving module. Specifically, we build a baseline without priors (denoted as “Baseline”). In order to evaluate these two priors, we first construct another variant

“Baseline w P ” by adding only P into the baseline network, and the baseline with both priors is denoted as “Full model”.

The quantitative results under the self-reenactment setting are reported in Tab 5. Since these two priors are proposed to handle the unstable texture issues, there is no obvious gap between their scores on the metric F-LMD. While for image quality metrics, these two variants without using the priors all suffer from certain degradation. The employment of P tends to provide texture prior even when

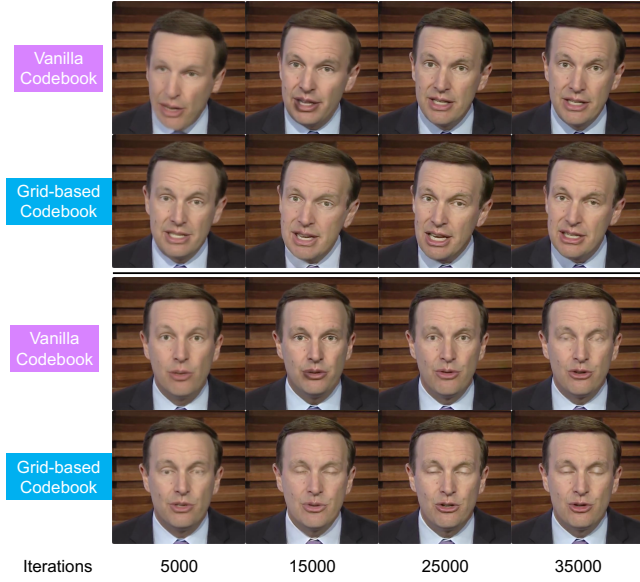


Figure 4: Qualitative comparison of different codebooks. It takes much fewer iterations for our grid-based codebook than the vanilla codebook to learn accurate and fine-grained local textures on the mouth and eyes areas. The portrait is from the HDTF dataset [Zhang et al. 2021] (CC BY 4.0)

Table 4: Visual comparison between vanilla codebook and our grid-based codebook at different iterations.

Codebook	PSNR			
	5000	15000	25000	35000
Vanilla	28.15	29.37	30.86	31.29
Grid-based	30.82	31.69	32.01	32.27

Table 5: Ablation study results when using different driving priors.

Methods	Testset A			
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	F-LMD $\downarrow$
Baseline	30.25	0.73	0.08	2.59
Baseline w P	31.69	0.74	0.07	2.73
Full model	32.36	0.79	0.06	2.27

processing unseen pose, while the usage of F_g can dramatically reduce the uncertainty of the prediction and alleviate unexpected jittering around the hair area. Please refer to our demo video for better understanding.

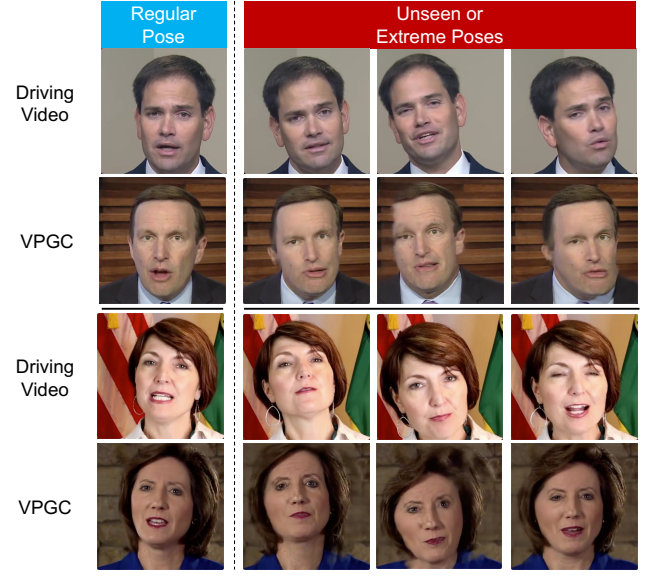


Figure 5: Visualization results of our method driven by unseen or extreme poses. The synthesized portrait suffers from obvious artifacts. All the portraits are selected from the HDTF dataset [Zhang et al. 2021] (CC BY 4.0)

5 CONCLUSION AND DISCUSSIONS

This paper presents the Video Portrait via Grid-based Codebook (VPGC) framework, which synthesizes robust and high-fidelity face reenactment results for specific portraits. We identify several advantages of our design: 1) The grid-based codebook facilitates the training of both the dictionary and the mapping between driving signals and codebook embeddings. 2) Detailed expressions can be accurately captured with our local learning paradigm. 3) Our method produces high-fidelity and more realistic results that show superiority over previous methods.

Limitations. Although our VPGC achieves superior performance over the previous approaches, we still notice some challenging cases that our VPGC cannot handle well, especially when there are rarely seen or extreme movements in the driving videos. We suppose this sensitivity is derived from the local-dependent nature of the grid-based codebook, in which the preserved personalized textual information is strongly entangled with the position (or explicit grid structure). Here we provide several examples of the failure cases in Fig. 5, from which we observe the generated portraits suffer from severe facial texture loss and distortion. These artifacts usually occur in certain regions that cannot be covered in the training data.

Ethical Considerations. As our method creates high-fidelity video portraits, it might lead to negative and harmful effects on society when used by the wrong hands. We would strictly limit the usage of our model for research purposes only and share our results with the deep fake detection community, which can benefit the development of advanced detection algorithms. We believe that proper usage of our technique will enhance development in both machine learning research and multimedia entertainment in our daily life.

REFERENCES

- Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 187–194.
- Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. 2023. Muse: Text-To-Image Generation via Masked Generative Transformers. *arXiv preprint arXiv:2301.00704* (2023).
- Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11315–11325.
- Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. 2022. Tensorf: Tensorial radiance fields. In *Proceedings of the European Conference on Computer Vision*. 333–350.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
- Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. 2021. HeadGAN: One-shot Neural Head Synthesis and Editing. In *IEEE/CVF International Conference on Computer Vision*.
- Nikita Drobyshev, Jenya Chelisev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. 2022. MegaPortraits: One-shot Megapixel Neural Head Avatars. *Proceedings of the 30th ACM International Conference on Multimedia*.
- Patrick Esser, Robin Rombach, Andreas Blattmann, and Bjorn Ommer. 2021b. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in Neural Information Processing Systems* 34 (2021).
- Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021a. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12873–12883.
- Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. 2021. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images. *ACM Transactions on Graphics* 40, 8. <https://doi.org/10.1145/3450626.3459936>
- Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. 2021. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8649–8658.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.
- Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022. Neural head avatars from monocular RGB videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18653–18664.
- Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. 2022. VQFR: Blind Face Restoration with Vector-Quantized Dictionary and Parallel Decoder. In *Proceedings of the European Conference on Computer Vision*.
- Yudong Guo, Keyu Chen, Sen Liang, Yongjin Liu, Hujun Bao, and Juyong Zhang. 2021. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. In *IEEE/CVF International Conference on Computer Vision*.
- Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. 2022. EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model. In *ACM SIGGRAPH 2022 Conference Proceedings (SIGGRAPH '22)*. <https://doi.org/10.1145/3528233.3530745>
- Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. 2021. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14080–14089.
- Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*. Springer, 694–711.
- Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. 2022. Realistic one-shot mesh-based head avatars. In *Proceedings of the European Conference on Computer Vision*. Springer, 345–362.
- Hyeonwoo Kim, Mohamed Elgharib, Michael Zollhöfer, Hans-Peter Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. 2019. Neural style-preserving visual dubbing. *ACM Transactions on Graphics* 38, 6 (2019), 1–13.
- Hyeonwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. 2018. Deep video portraits. *ACM Transactions on Graphics* 37, 4 (2018), 1–14.
- Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. 2022. Audio-Driven Co-Speech Gesture Video Generation. *Advances in Neural Information Processing Systems* (2022).
- Yuanxun Lu, Jinxiang Chai, and Xun Cao. 2021. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics* 40, 6 (2021), 1–17.
- Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision*. Springer, 405–421.
- Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics* 41, 4 (2022), 1–15.
- Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First order motion model for image animation. *Advances in Neural Information Processing Systems* 32 (2019), 7137–7147.
- Cheng Sun, Min Sun, and Hwann-Tzong Chen. 2022a. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5459–5469.
- Yasheng Sun, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Zhibin Hong, Jingtuo Liu, Errui Ding, Jingdong Wang, Ziwei Liu, and Koike Hideki. 2022b. Masked Lip-Sync Prediction by Audio-Visual Contextual Exploitation in Transformers. In *SIGGRAPH Asia 2022 Conference Papers*. 1–9.
- Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics* 36, 4 (2017), 1–13.
- Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Gang Zeng, and Jingdong Wang. 2022. Real-time Neural Radiance Talking Portrait Synthesis via Audio-spatial Decomposition. *arXiv preprint arXiv:2211.12368* (2022).
- Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2387–2395.
- Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in Neural Information Processing Systems* 30 (2017).
- Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. MEAD: A Large-scale Audio-visual Dataset for Emotional Talking-face Generation. In *ECCV*.
- Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10039–10049.
- Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. 2022. Latent Image Animator: Learning to Animate Images via Latent Space Navigation. In *International Conference on Learning Representations*.
- Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. 2018. Reenactgan: Learning to reenact faces via boundary transfer. In *Proceedings of the European Conference on Computer Vision*. 603–619.
- Kewei Yang, Kang Chen, Daoliang Guo, Song-Hai Zhang, Yuan-Chen Guo, and Weidong Zhang. 2022. Face2Face p: Real-Time High-Resolution One-Shot Face Reenactment. In *Proceedings of the European Conference on Computer Vision*. Springer, 55–71.
- Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. 2022. StyleHEAT: One-Shot High-Resolution Editable Talking Face Generation via Pre-trained StyleGAN. In *Proceedings of the European Conference on Computer Vision*.
- Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. 2020. Fast Bi-layer Neural Synthesis of One-Shot Realistic Head Avatars. In *Proceedings of the European Conference on Computer Vision*. Springer, 524–540.
- Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9459–9468.
- Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 2021. 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*. 3661–3670.
- Chuanxia Zheng, Tat-Jen Cham, Jianfei Cai, and Dinh Phung. 2022b. Bridging Global Context Interactions for High-Fidelity Image Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11512–11522.
- Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. 2022a. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13545–13555.
- Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9299–9306.
- Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4176–4186.
- Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. 2022. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems* 35 (2022), 30599–30611.