# Revising Second Order Terms in Deep Animation Video Coding

Konstantin Schmidt
*Fraunhofer IIS*
Am Wolfsmantel 33, 91058 Erlangen, Germany
konstantin.schmidt@iis.fraunhofer.de

Thomas Richter
*Fraunhofer IIS*
Am Wolfsmantel 33, 91058 Erlangen, Germany
thomas.richter@iis.fraunhofer.de

***Abstract*—First Order Motion Model is a generative model that animates human heads based on very little motion information derived from keypoints. It is a promising solution for video communication because first it operates at very low bitrate and second its computational complexity is moderate compared to other learning based video codecs. However, it has strong limitations by design. Since it generates facial animations by warping source-images, it fails to recreate videos with strong head movements. This works concentrates on one specific kind of head movements, namely head rotations. We show that replacing the Jacobian transformations in FOMM by a global rotation helps the system to perform better on items with head-rotations while saving 40% to 80% of bitrate on P-frames. Moreover, we apply state-of-the-art normalization techniques to the discriminator to stabilize the adversarial training which is essential for generating visually appealing videos. We evaluate the performance by the learned metics LPIPS and DISTS to show the success our optimizations.**

***Index Terms*—video-coding, facial-animation, deep-animation, first-order-motion-model, generative-video-coding**

## I. Introduction

Recently there has been a lot of progress in the development of learning based video codecs for communication applications. These codecs are able to outperform classical engineered codec standards like e.g. H.264, H.265 or H.266 (AVC [1], HEVC [2], VVC [3]) by quite a margin [4]–[6]. Among these learning based codecs two main paths of development can be identified. First are codecs that mimic the classical design of video codecs based on motion estimation and residual coding. These codecs use deep networks to replace the hand-engineered predictors and residual coders. The main drawback of this design is the huge computational complexity caused by the decoder running in the encoder. Such codecs are able to generate videos at bitrates more than 30 % lower than H.266 [5]. Second are codecs based on First Order Motion Model (FOMM) [7]–[22] that are able to generate appealing videos at moderate computational complexity at bitrates of 3 kbps or even lower. FOMM has less than 60 million parameters and 55 G-MACS complexity per P-frame. These codecs are suited for communication applications only (coding of human faces), while the former are able to code general content. Such codecs use a single image (denoted here as I-frame) that is animated by as low as 10 keypoints that are transmitted per video-frame to be coded (denoted here as P-frame). While sometimes being able to generate almost artifact-free videos at bitrates below 3 kbps, such codecs have major limitations by design. One such limitation is strong head rotations on the roll-axis of P-frames, since these items do not occur very often in the training set. As a solution to this problem, we propose to amend or replace the Jacobian transformation of the KPs by a single rotation parameter. We show that this improves the quality of generated videos while also reducing the bitrate of P-frames.

The second optimization we propose is targeting the adversarial loss and could be applied to any generative facial animation codec. The basic principle of FOMM is to generate videos by warping I-frames based on some warping information coded in the bitstream. Often the P-frame to generate contains elements that are not present in the I-frame. In such cases the warping fails to generate meaningful content and the model relies on an adversarial loss to hallucinate these parts. The adversarial loss is known to be very unstable and as a result often only a small amount is added to the overall loss. We propose two normalization techniques to stabilize the discriminator which allows for higher amount adversarial loss and thus better image quality. First, to our best knowledge we are the first to use Gradient Normalization [23] to stabilize the discriminator in generative video coding. Second, inspired by self-supervised learning, we let the discriminator learn facial landmarks while training. In summary our contributions are:

- Optimized linear transformation of local patches that form the output frame.
- Reduced P-frame bitrate compared to the original FOMM.
- A more stable adversarial loss that results in higher image fidelity.

## II. Optimizations of the Transformations for Warping I-frames

FOMM generates animated P-frames by warping I-frame areas around so called keypoints (KP). These KPs are outputted by a DNN per video from (see Fig. 2). They are similar to facial landmarks (which identify facial elements like eyes, nose, etc.) but are learned unsupervisedly. Fig. 1a depicts the difference between KPs and landmarks on a video-frame of an

The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The hardware is funded by the German Research Foundation (DFG)

(a) Landmarks (red) and Key-points (KP).

(b) Warped patches with one additional patch for the background (bottom center).

item of the training set. The 68 landmarks generated with [24] are plotted in red and the KPs are plotted as yellow asterisks. These KPs, together with also unsupervisedly learned local linear transformations matrices called Jacobians (JACs) form the bitstream of the P-frames.

The warping operation used here is based on Kazemi et. al. [25] where the warped output is calculated by the per-pixel displacement information given in the matrix $WG$ (please see [25] for more details on this). This matrix is calculated as:

$$WG = (WG_{neutral} - KP_p) \cdot JAC_i \cdot JAC_p^{-1} + KP_i, \quad (1)$$

based on the unitary warping grind $WG_{neutral}$. $JAC_i$ and $JAC_p$ are local 2x2 linear transformation matrices of I-frame and P-frame that, together with the KPs allow for affine (i.e. first order) transformations of local patches. Fig. 1b (right) depicts such patches. Each patch contains a mutually exclusive part of the input image that is warped and merged by a generator DNN (Fig. 2) to form the output image. Please note that Eq. 1 is performed for each of the 10 KPs. After inspecting these local matrix-transformations, we conclude that:

- Rotations mostly follow global head rotations on the jaw-axis.
- Scalings mostly follow global head scaling.
- Transformations like shearing often appear random.
- Calculating inverse $JAC$s as in 1 is often unstable.

Especially during training an unstable matrix inversion is suboptimal for robust gradient flow. This motivates us to replace the local transformation matrices $JAC_k$ for KP $k$ by a single global rotation and a scaling:

$$JAC_k := R \cdot scf = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \cdot scf, \quad (2)$$

or a combination of a single global rotation, scaling and shearing matrices:

$$\begin{aligned} JAC_k &:= R \cdot SHR_k \cdot scf \\ &= \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \cdot \begin{bmatrix} 1+\lambda\mu_k & \lambda_k \\ \mu_k & 1 \end{bmatrix} \cdot scf. \end{aligned} \quad (3)$$

The rotation matrix $R = \left(\begin{smallmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{smallmatrix}\right)$ has a single parameter $\phi$ that is given by the KP-detection DNN (Fig. 2) and learned by a supervised loss from a head-pose estimation loss. We use the model in [26] as head-pose estimation network which gives a robust estimation of jaw, pitch and roll axis of the head. Here, we only use the roll axis which directly gives the parameter $\phi$. The inverse of such a rotation matrix is simply its transpose and is easy to calculate. As loss for learning $\phi$ we simply use an L-1 distance between $\phi$ and the head-pose estimation network output. Using a head-pose estimation loss was first done in [9] but applied to a different architecture.

Since the KPs linearly scale with the size of the head in the video, we hypothesize that the scaling operation of a general 2x2 matrix can be replaced by a single scale factor $scf$ deduced from the already trasmited KPs by linear regression at decoder side:

$$scf = \frac{\sum_k (KP_{i,k} - \overline{KP_{i,k}}) * (KP_{p,k} - \overline{KP_{p,k}})}{\sum_k (KP_{i,k} - \overline{KP_{i,k}})^2}, \quad (4)$$

where $KP_{i,k}$ and $KP_{p,k}$ are the $k$-th KPs of the I-frame and the P-fames respectively, $\overline{KP_i}$ and $\overline{KP_p}$ are the mean of the KPs.

Replacing the per-keypoint 2x2 matrix transformations with these 2 global transformations already works well and will be evaluated as a first low-bitrate system in Sec. IV. However, sometimes the performance can be improved by a shearing operation in Eq. 3 that can't be performed by the afore mentioned rotation and scaling. Such a shearing matrix contains two learnable parameters per KP. Fortunately the inverse of a shearing matrix is easy to calculate as:

$$SHR^{-1} = \begin{bmatrix} 1 & -\lambda \\ -\mu & 1+\lambda\mu \end{bmatrix} \quad (5)$$

## III. Optimizations of the Adversarial Training

As mentioned before, parts of the generated P-frames can be deduced from the I-frame and the warping information transmitted in the bitstream, while other parts need to be hallucinated by a generative model. The generative model used here is a generative adversarial network (GAN) which allows for good performance at moderate model size and computational complexity compared to e.g. diffusion models. However, the unstable training process remains a challenging problem and often the generator tends to fool the discriminator before learning to generate realistic images. This is usually caused by the sharp gradient space of the discriminator, which causes mode collapse in the training process of the generator. A promising solution to this problem is Gradient Normalization (GN) [23] which is a model-wise, non-sampling-based, and non-hard normalization of the discriminator function. A discriminator normalized with GN has increased capacity while still being Lipschitz-contrained, which ultimately results in higher fidelity of the generated content.

Several other systems that are based on FOMM propose to use a landmark loss as additional loss during training [20]. This landmark loss assures that the generated images have the same facial landmarks as the original image. Here, we propose to move the landmark loss as an additional self-supervision loss after the discriminator for two reasons: First, the predominant method of landmark estimation relies on [24], which is a non-differentiable model and can't directly be used as loss. Second, according to [27], [28] a self-supervision loss

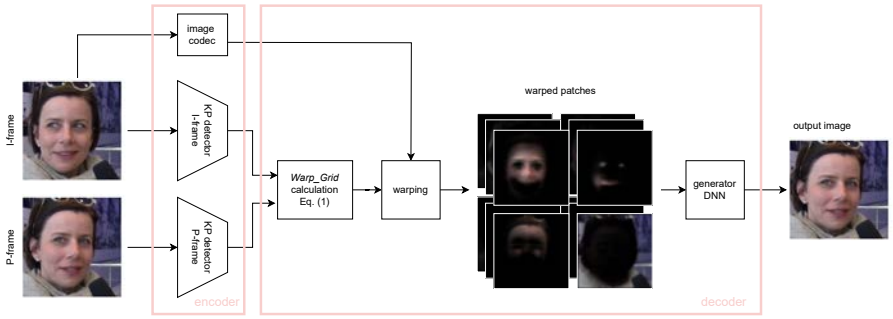


Fig. 2: Block diagram of encoder and decoder in red boxes. Input are I-frame and P-frame on the left, output is the generated P-frame on the right. DNNs are depicted with green outlines.

helps to mitigate overfitting, improves the stability and generalizability of discriminator and avoids discriminator forgetting. The loss calculation is shown by red paths in Fig. 3. To our best knowledge we are the first successfully applying GN and self-supervision by landmarks in generative models for video coding.

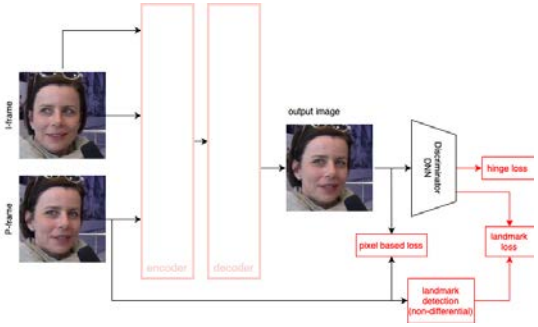


Fig. 3: Block diagram of the proposed discriminator optimization. Red blocks are needed to calculate the discriminator loss.

## IV. Evaluation

Before presenting the results, we discuss the selection of metrics we use and what data has been used for training and evaluation. It is known that metrics based on pixel distances like PSNR and SSIM are not able to predict the perceived quality of content created by generative models. Among the best performing metrics to evaluate the quality of generated videos are DNN based metrics like LPISPS [29] and DISTS [30]. These metrics estimate the quality of each video-frame and finally calculate an average over all frames. They are based on the hypothesis that features extracted from image classification networks are also able to estimate human perceived quality. According to the authors DISTS correlates better with human ratings. PSNR and SSIM values of the presented systems are given in Tab. I for completeness only. To evaluate the diversity of the generated images we use Fréchet inception distance (FID) which compares the distribution of generated images with the distribution of a set of real images. A high FID score can be used to monitor mode collapse in adversarial training.

The dataset used for training is the Vox-Celeb2 dataset [31] which contains videos at a resolution of 256×256 pixel. We used 32927 items for training and 6174 items for testing, with the test items not being part of the training set. The generated videos contain 90 frames, with only the first frame being sent as I-frame. Opposed to the reference settings in [7] we use the AdamW optimizer [32] with the parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) and a learning rate of $2e^{-4}$. The training is conducted for 480 epochs on NVIDIA H100 GPU with batch size of 68 [33].

We also compare our system to the one presented by Wang et. al. [9]. Since there is no official implementation available we use the implementation from [34]. Their system is labeled as "OSFV" in the plots.

The results are given in Tab. I and Fig. 4,5,6 as distribution over all test items. The proposed rotation, scaling and shearing (*rot+scale+shear*) achieving the best results in all metrics. Not sending any Jacobians at all (*no JAC*) has the strongest negative impact on the quality. Sending only a global rotation (*rot.+scal.*) already brings the quality close to the full Jacobian reference (*full JAC*). It has to be emphasized that the presented systems have much lower elements in the bitstream. In addition to the 10 KPs that are present in all systems, FOMM needs 4 additional parameters per KP. The presented system with rotation and scaling (rot. + scale) has only one global rotation while the presented system with rotation, scaling and shearing has 2 additional parameters per KP. Estimated bitrates are also given in Tab. I.

| | rot.+scale +shear | rot.+scal. | full JAC | no JAC | OSFV |
|---|---|---|---|---|---|
| LPIPS ↓ | **0.179** | 0.195 | 0.192 | 0.207 | 0.231 |
| DISTS ↓ | **0.099** | 0.105 | 0.105 | 0.119 | 0.115 |
| FID ↓ | **47.70** | 48.95 | 47.87 | 57.77 | 52.74 |
| PSNR ↑ | **24.14** | 23.6 | 23.88 | 22.81 | 20.84 |
| SSIM ↑ | **0.795** | 0.786 | 0.794 | 0.784 | 0.662 |
| bitrate [kbps] ↓ | ∼5 | ∼3.1 | ∼8 | ∼3 | 10.6 |

TABLE I: Average results of learned and classical metrics and also bitrate on test set. Arrows indicate if lower or bigger score mean better performance.

The impact of the optimizations of the discriminator is given in Tab. II. The table shows the objective metrics depending

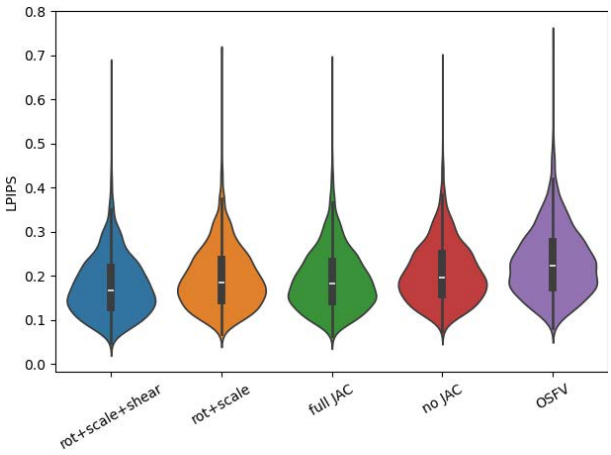


Fig. 4: LPIPS on test set. Lower means better performance.

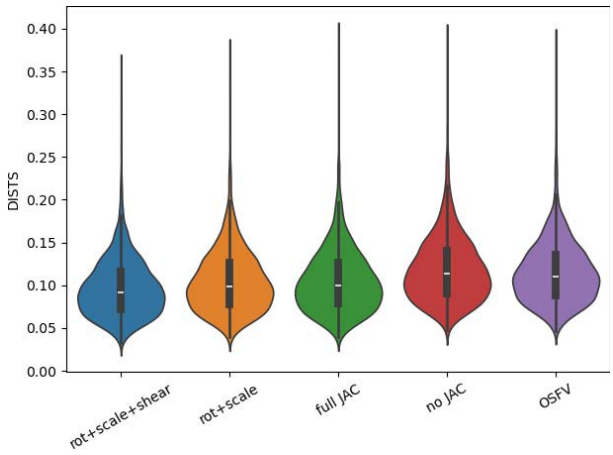


Fig. 5: DISTS on test set. Lower means better performance.

on the amount ($\lambda$) of adversarial loss used and if gradient normalization (*GN*) is present or not (*no GN*). Here it can be seen that the proposed discriminator optimizations allow for 4 times larger adversarial loss ultimately resulting in better quality. The last column shows the result from a training, where the discriminator became unstable, and the generated items were mostly noise.

| | $\lambda$ 4 GN | $\lambda$ 2 GN | $\lambda$ 1 no GN | $\lambda$ 4 no GN |
|---|---|---|---|---|
| LPIPS ↓ | **0.1793** | 0.1956 | 0.1993 | 0.721 |
| DISTS ↓ | **0.0998** | 0.109 | 0.1171 | 0.527 |

TABLE II: Impact of the optimizations of the discriminator. $\lambda$-values give the amount of adversarial loss used.

Finally we provide numbers of parameters and estimates of complexity of the presented models and of OSFV in Tab. III. The complexity is given in giga multiply-adds [MACs] per video frame. The proposed optimizations have only a very small impact in the complexity and memory size of FOMM.

| | Nr of parameters | complexity [giga MACs] |
|---|---|---|
| FOMM | $59.79 \times 10^6$ | 54.96 |
| ours | $59.87 \times 10^6$ | 55.22 |
| ours no JAC | $59.72 \times 10^6$ | 54.73 |
| OSFV | $173.2 \times 10^6$ | 483.85 |

TABLE III: Numbers of parameters and estimates of complexity of the presented model and of OSFV.

## V. Summary

We present optimizations of a promising video codec that can satisfy the ever-growing hunger for video-communication data rate. Besides operating at very low bitrates this codec also runs at moderate computational complexity and with low delay. Our work shows that the Jacobian in FOMM may not be needed as full 2x2 matrices. Furthermore, we show that stabilizing the discriminator can further improve the quality. This stabilization is not limited to the proposed system and can be used in all learning based communication codecs.

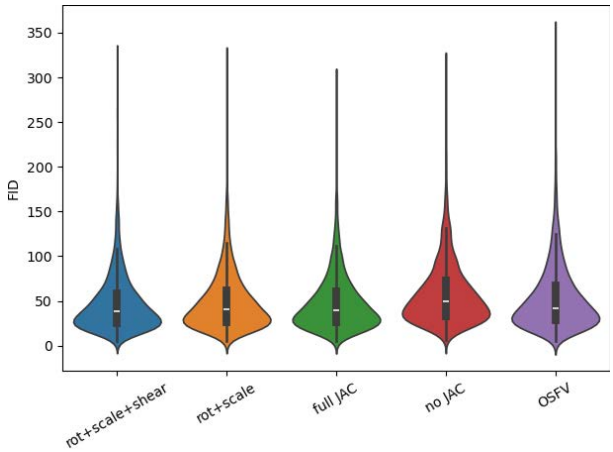


Fig. 6: Fréchet inception distance (FID) on test set. Lower means better performance.

## References

[1] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the h.264/avc video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[2] Gary J. Sullivan, Jill M. Boyce, Ying Chen, Jens-Rainer Ohm, C. Andrew Segall, and Anthony Vetro, "Standardized extensions of high efficiency video coding (hevc)," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 6, pp. 1001–1016, 2013.

[3] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, and Jens-Rainer Ohm, "Overview of the versatile video coding (vvc) standard and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, 2021.

[4] Xihua Sheng, Jiahao Li, Bin Li, Li Li, Dong Liu, and Yan Lu, "Temporal context mining for learned video compression," *IEEE Transactions on Multimedia*, vol. 25, pp. 7311–7322, 2023.

[5] Jiahao Li, Bin Li, and Yan Lu, "Neural video compression with feature modulation," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 26099–26108.

[6] Jiahao Li, Bin Li, and Yan Lu, "Neural video compression with diverse contexts," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 22616–22626.

[7] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe, "First order motion model for image animation," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.

[8] Goluck Konuko, Giuseppe Valenzise, and Stéphane Lathuilière, "Ultra-low bitrate video conferencing using deep image animation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 4210–4214.

[9] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu, " One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing ," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, June 2021, pp. 10034–10044, IEEE Computer Society.

[10] Maxime Oquab, Pierre Stock, Oran Gafni, Daniel Haziza, Tao Xu, Peizhao Zhang, Onur Celebi, Yana Hasson, Patrick Labatut, Bobo Bose-Kolanu, Thibault Peyronel, and Camille Couprie, " Low Bandwidth Video-Chat Compression using Deep Generative Models ," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Los Alamitos, CA, USA, June 2021, pp. 2388–2397, IEEE Computer Society.

[11] Anni Tang, Yan Huang, Jun Ling, Zhiyu Zhang, Yiwei Zhang, Rong Xie, and Li Song, "Generative compression for face video: A hybrid scheme," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 2022, pp. 1–6.

[12] Madhav Agarwal, Anchit Gupta, Rudrabha Mukhopadhyay, Vinay Namboodiri, and C.V. Jawahar, "Compressing video calls using synthetic talking heads," in *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. 2022, BMVA Press.

[13] Bo Chen, Zhao Wang, Binzhe Li, Shurun Wang, Shiqi Wang, and Yan Ye, "Interactive face video coding: A generative compression framework," *ArXiv*, vol. abs/2302.09919, 2023.

[14] Dahu Feng, Yan Huang, Yiwei Zhang, Jun Ling, Anni Tang, and Li Song, "A generative compression framework for low bandwidth video conference," in *2021 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2021, pp. 1–6.

[15] Bolin Chen, Zhao Wang, Binzhe Li, Rongqun Lin, Shiqi Wang, and Yan Ye, "Beyond keypoint coding: Temporal evolution inference with compact feature representation for talking face video compression," in *2022 Data Compression Conference (DCC)*, 2022, pp. 13–22.

[16] Zhao Wang, Bolin Chen, Yan Ye, and Shiqi Wang, "Dynamic multi-reference generative prediction for face video compression," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 896–900.

[17] Ruofan Wang, Qi Mao, Chuanmin Jia, Ronggang Wang, and Siwei Ma, "Extreme generative human-oriented video coding via motion representation compression," in *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2023, pp. 1–5.

[18] Anchit Gupta, Rudrabha Mukhopadhyay, Sindhu Balachandra, Faizan Farooq Khan, Vinay P. Namboodiri, and C. V. Jawahar, "Towards generating ultra-high resolution talking-face videos with lip synchronization," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 5198–5207.

[19] Weijie Yue, Jincheng Dai, Sixian Wang, Zhongwei Si, and Kai Niu, "Learned source and channel coding for talking-head semantic transmission," in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, 2023, pp. 1–6.

[20] Fa-Ting Hong and Dan Xu, " Implicit Identity Representation Conditioned Memory Compensation Network for Talking Head Video Generation ," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Los Alamitos, CA, USA, Oct. 2023, pp. 23005–23015, IEEE Computer Society.

[21] Zhehao Chen, Ming Lu, Hao Chen, and Zhan Ma, "Robust ultralow bitrate video conferencing with second order motion coherency," in *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, 2022, pp. 1–6.

[22] Lele Chen, Guofeng Cui, Ziyi Kou, Haitian Zheng, and Chenliang Xu, "What comprises a good talking-head video generation?: A survey and benchmark," 2020.

[23] Yi-Lun Wu, Hong-Han Shuai, Zhi-Rui Tam, and Hong-Yu Chiu, "Gradient normalization for generative adversarial networks," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6353–6362.

[24] Vahid Kazemi and Josephine Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.

[25] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds. 2015, vol. 28, Curran Associates, Inc.

[26] Nataniel Ruiz, Eunji Chong, and James M. Rehg, "Fine-grained head pose estimation without keypoints," 2018.

[27] Ziqiang Li, Muhammad Usman, Rentuo Tao, Pengfei Xia, Chaoyue Wang, Huanhuan Chen, and Bin Li, "A systematic survey of regularization and normalization in gans," *ACM Comput. Surv.*, vol. 55, no. 11, Feb. 2023.

[28] Spyros Gidaris, Praveer Singh, and Nikos Komodakis, "Unsupervised representation learning by predicting image rotations," 2018.

[29] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[30] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–1, 2020.

[31] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[32] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2017.

[33] "The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). the hardware is funded by the German Research Foundation (DFG)," .

[34] Zhang Long Hao, "One-shot free-view neural talking head synthesis," https://github.com/zhanglonghao1992/One-Shot_Free-View_Neural_Talking_Head_Synthesis, 2021.