Improved predictive coding for animation-based video compression

Goluck Konuko, Giuseppe Valenzise

Improved predictive coding for animation-based video compression

Goluck Konuko[1], Giuseppe Valenzise[2]
[1]*Université Paris-Saclay*, [1,2]*Laboratoire de Signaux et systèmes,(CentraleSupelec)*,[2]*CNRS*
[1]goluck.konuko@centralesupelec.fr, [2]giuseppe.valenzise@centralesupelec.fr

***Abstract*—This paper addresses the limitations of generative face video compression (GFVC) under conditions of substantial head movement and complex facial deformations. Previous GFVC frameworks focused on perceptual compression and reconstruct videos only with the goal of perceptual quality. As a result, they often have a large disparity relative to conventional codecs when evaluated for pixel fidelity. We propose a robust framework for learned predictive coding process aiming for both perceptual quality and improved performance in terms of pixel fidelity under low bitrate conditions. Our method proposes a dual residual learning strategy. Specifically, it learns the frame residual between the animated frame and the ground truth i.e. spatial residual coding and further exploits redundancies between neighboring frame residuals i.e temporal residual coding. We specially formulate a low bitrate conditional residual coding mechanisms for both spatial and temporal residual coding. In addition, we propose a zero-cost residual alignment mechanism to refine prediction accuracy of frame residuals. Through end-to-end optimization, the proposed framework achieves a balance between perceptual quality, pixel fidelity and compression efficiency. We conduct experimental evaluations on test sequences and conditions proposed under the JVET-AH0114 standard to show significant performance gains relative to HEVC and VVC standards in terms of perceptual metrics. Compared to other GFVC frameworks, our proposed framework achieves state of the art performance on perceptual metrics and pixel fidelity metrics. It is also competitive with HDAC, HEVC and VVC in terms of pixel fidelity at low bitrates.**

***Index Terms*—generative video compression, face animation, self re-enactment**

I. Introduction

Achieving low-bitrate compression with high-quality reconstruction remains an open challenge in video coding. In recent years, a number of end-to-end learned image and video compression frameworks have been proposed [1]–[4] to solve the limitations associated with conventional video compression approaches. These have shown competitive performance relative to conventional codecs such as HEVC [5] and VVC [6]. However, learned compression frameworks are typically designed for the general purpose video compression and do not address the specific conditions of talking face compression which is typical of video conferencing *i.e.* ultra low bitrate and real-time processing. This work follows a line of generative face video compression (GFVC) frameworks proposed to use animation models for talking head video compression at ultra low bitrates [7]–[13]. GFVC methods achieve extreme high coding efficiency by encoding only a sparse motion representation and few reference frames and use a generative autoencoder network to reconstruct video sequences with high perceptual quality. Inspired by the First Order Animation model [14], GFVC frameworks assume that face and head motion can be represented through a compact set of sparse motion keypoints, which can be entropy coded and transmitted to the decoder. A decoding network trained within an adversarial learning process uses this information to reconstruct an approximation of the original talking head sequence *i.e.* face self re-enactment with minimal loss of facial information.

However, these schemes are *open-loop*, i.e., frames are predicted from a reference picture without any mechanism to correct the prediction errors. This creates a number of limitations including an error *drift* as target frames increasingly diverge from the reference frame in terms of spatial complexity and pose variations. Moreover, most GFVC frameworks are limited in rate-distortion performance since the main variable is typically the quality and number of the reference frames. On the other hand, state-of-the-art video codecs typically include a prediction and residual coding loop to maximize reconstruction accuracy but are burdened by a high computation complexity [6] that limit the deployment of their advanced optimization processes in video conferencing applications. In the case of animation-based codecs, designing and implementing such a predictive coding scheme is challenging because there is no guarantee that the coding cost of the frame residuals will be lower than the cost of the original image. However, a prior attempt in this direction *i.e.* predictive coding for animation-based video compression (RDAC) [13] explored a closed-loop coding scheme based on face animation and demonstrated a potentially viable approach to achieve predictive coding under low-bitrate conditions.

Inspired by RDAC, this paper proposes a robust predictive coding framework that learns a low bitrate representation of spatial and temporal frame residuals. In the process, we formulate conditional residual coding [15] for ultra-low bitrate compression. Further we propose a formulation of an autoencoder architecture that learns a conditional residual compression between temporally neighboring frame residuals. Finally, we include an improvement to the predictive coding frameworks by using a zero-cost alignment strategy of frame residuals to minimize the cost of differential residual coding. Specifically, by reusing the motion information conveyed by the animation keypoints, we show that in the GFVC coding process, there is a sufficient temporal correlation between

frame residuals such that effective motion compensation is necessary to reduce the coding cost associated with the frame residuals. We call our proposed approach **RDAC+** in the rest of the paper. We include a set of experiments and evaluations to characterize different elements of RDAC+ and show efficiecy gains of the proposed framework relative to conventional state-of-the-art codecs such as HEVC, VVC as well as some prior GFVC frameworks.

II. RELATED WORK

Learning-based compression of talking head videos has gained significant interest since the development of deep learning approaches for effective image animation such as [14]. These frameworks focus on the optimization of perceptual quality in reconstructed videos. Prior works in this area [7], [8], [10] demonstrated the efficiency of animation-based compression at ultra-low bitrates. Compact feature representation (CFTE) [11] optimizes the compactness of the bitstream representation of the motion keypoints with Exp-Golomb codes, leading to higher accuracy in motion prediction at lower bitrates. These frameworks rely only on image animation for frame reconstruction, which fits into the domain of perceptual compression. However, they do not have a reliable method for quality improvement when additional bit budget is allocated to the codecs. Additionaly, they have a notable drift in reconstruction performance as the temporal distance between the reference and target frame increases as well as under extreme occlusions and disocclusions. As a result, most prior research [7]–[10] has focused on talking-head videos with minimal complexity outside the facial region and are evaluated based only on the perceptual quality relative to conventional codecs at ultra low bitrates.

Previous works such as [7], [12], [13] have attempted to consider scenes with complex background and foreground deformations. In the deep animation codec (DAC) [7], an adaptive Intra refresh algorithm is introduced to limit the drift in image animation. A threshold parameter is used to measure significant deviations between the reference and the target frame, adding a new reference when needed. This approach still suffers from error drift and frequent introduction of reference frames introduces temporal artifacts in the decoded video. The spatio-temporal animation framework proposed by Chen et al. [10] addresses the jittering artifacts that are noted in a number of animation-based coding frameworks. The hybrid coding strategy (HDAC) [12], similarly addressed the problem with jittering and goes further to enable variability in reconstruction quality of animated sequences. This is achieved through a hybrid, layered coding architecture leveraging a low-quality HEVC bitstream as side information to enhance the final result of the animation codec. However, training HDAC requires an extensive data preprocessing to create multiple quality levels of the target base layer codec and a complex specialized optimization process. It is therefore evident that a fundamental limitation of previous approaches is the total or partial *open-loop* nature of the codec. The first end-to-end learned predictive coding with image animation is proposed in RDAC [13]. The results are promising and showed significant reduction in drift error propagation and allows obtaining competitive rate-distortion performance across a much larger range of bitrates than previous animation-based codecs, which often operate only at fixed rate points. This paper proposes an advanced version of the predictive coding strategy inspired by RDAC and is therefore called RDAC+. Specifically, in recognition of the entropy constraint required for GFVC frameworks, we propose a conditional residual coding approach for spatial and temporal frame residuals. This offers improved coding efficiency relative to the vanilla learned residual coding approach used in RDAC. We further improve the coding efficiency of the temporal residuals in RDAC through motion estimation and compensation. We propose a faster alternative to traditional motion estimation *i.e.* motion vector search. Specifically, we propose a spatial alignment process that reuses the motion keypoints from animation, thus avoiding the need to compute and transmit motion vectors. Our framework provides higher reconstruction performance and extends the range of bitrates beyond what is covered by [7]–[11] while achieving higher perceptual quality relative to HEVC and VVC and showing competitive performance on pixel fidelity metrics at ultra-low bitrates. We make evaluations against [7], [10], [12], [13] due to their close match in model complexity and their recent adoption in the JVET-AH0114 [1] standard.

III. PROPOSED METHOD: RDAC+

The core elements of the proposed framework in Fig. 1 is similar to RDAC [13].A deep image animation process uses a reference image and a sparse set of motion keypoints learned through self-supervised training to reconstruct talking-head video sequences. A frame residual, computed as the difference between the animated and the original frame is transmitted as a low dimensional latent representation and used to add missing features to the animated frame. We propose a conditional residual coding process in RDAC+(III-A) which offers additional coding efficiency compared to the simpler residual coding used in RDAC. In addition, we propose a zero-cost motion compensation strategy (III-B) to maximize coding efficiency by spatially aligning temporal residuals.The framework is optimized end-to-end to encode video over a larger range of bitrates than previous models and achieves a higher perceptual and pixel fidelity performance.

A. Learned Predictive Coding in Animation-Based Codecs

We interpret the output of the animation process $\hat{X}_t$ as a prediction of the actual inter frame X_t. Thus, following the classical predictive coding approach, we aim to efficiently code the prediction residual $R_t = X_t - \hat{X}_t$, achieving a reconstruction quality gain with respect to the open-loop animation-based codec. Moreover, we observe that residual frames R_t are themselves temporally correlated. Therefore, we can further leverage these correlations to perform temporal predictive

[1] https://jvet-experts.org/doc_end_user/current_document.php?id=14051

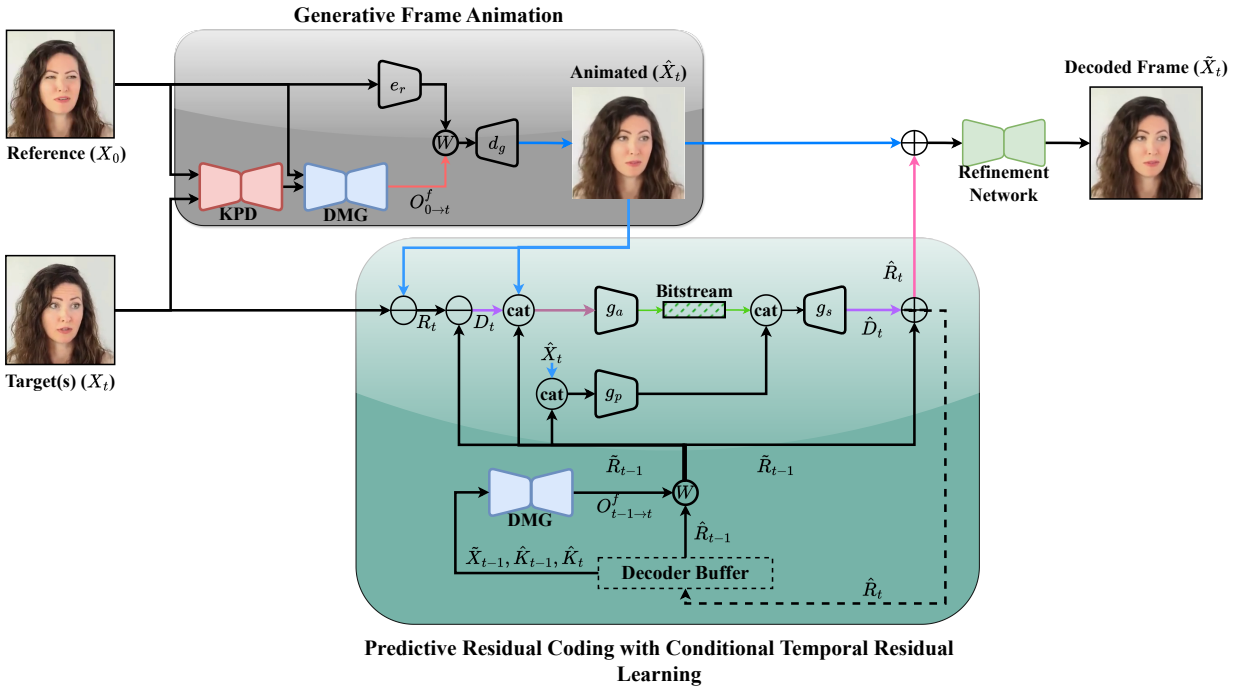


Fig. 1: **Proposed RDAC+ framework**. All the target frames are animated using information from a sparse keypoint detector (KPD) passed to a dense motion generator (DMG) to predict the optical flow ($O^f_{0\to t}$) between the target and reference frame. The optical flow is applied to the reference frame features extracted by an encoder (e_r) through a grid sampling operation (W). The deformed features are used by a decoder (d_g) to generate a prediction of the target frame. The frame residual between the original and animated frame is compressed into a bitstream using a conditional residual coder and used at the decoder side to reconstruct the final output. A refinement network is further applied to improve texture and high frequency detail reconstruction.

coding on residuals. Our system thus includes two prediction loops: one outer loop using image animation as predictor, and an inner loop for predictive residual coding, as shown in Figure 1. Here, we include two variational autoencoders (VAE) for conditional residual compression.These follow the common architecture proposed by [1] and adapted for low bitrate coding similar to RDAC [13]. At time $t-1$, if there are no previously decoded inter frames in the decoder buffer, then the first conditional residual coder is used to encode the frame residual R_{t-1} obtained after animation. We refer to this as a spatial conditional residual coding process which can be described as follows:

$$\hat{R}_{t-1} = g_s(round(g_a(R_{t-1}||\hat{X}_{t-1}))||g_p(\hat{X}_{t-1})) \quad (1)$$

where $||$ represents a concatenation operation, g_a and g_p are analysis networks that create low dimensional embedding for conditional reconstruction of the frame residual. A rounding operation and additional of uniform noise is used to parameterize the quantization process at training time.

Subsequently at time t, we use the previously decoded $\hat{R}_{t-1}$ to reduce the coding cost of the current frame residual R_t. We compute the temporal frame residual $D_t = R_t - \tilde{R}_{t-1}$, which is subsequently compressed into a low dimensional latent representation y_t. The condition process for conditional temporal residual coding is formulated as follows:

$$\hat{D}_t = g_s(round(g_a(D_t||\hat{R}_{t-1}||\hat{X}_t))||g_p(\hat{R}_{t-1}||\hat{X}_t)) \quad (2)$$

In Fig. 1 we show only the conditional temporal residual coding loop for clarity. The temporal frame residual is reconstructed as $\hat{y}_t$ from the bitstream and used at the decoder to generate $\hat{D}_t$. Subsequently, $\hat{D}_t$ is summed back to the decoded residual $\tilde{R}_{t-1}$ to produce a reconstructed residual $\hat{R}_t$, which together with the animated frame $\hat{X}_t$ allows decoding the frame $\tilde{X}_t$. Similar to RDAC [13], the temporal frame residuals are subsampled to 0.5 resolution relative to the spatial frame residual. Further, we experimentally determine and fix a prediction window of 8 frames for temporal residual coding. In both RDAC and RDAC+ we note that increasing the prediction window above 8 frames further reduces the bitrate but may introduce errors in some sequences with complex features in the frame residuals. We propose an investigation on how to effectively increase the temporal residual prediction window for further study due to its potential to increase the coding efficiency of the proposed framework.

B. Motion Compensated Learned Residual Coding for Animation-Based Coding

Classical video compression relies on motion compensation as an effective form of prediction, aligning the spatial content between one or multiple reference frames and the current one. In learned video compression frameworks [2]–[4], the optical flow between frames is encoded as a latent representation and transmitted as part of the bitstream. However, coding optical flows entails a significant bitrate consumption, which is unfeasible in the ultra-low-bitrate scenarios targeted in this work. An alternative used in standard video codecs is block-based motion compensation. The advantage of block matching is that it can be highly optimized in a rate-distortion sense, while producing a smaller set of motion vectors to transmit. However, in our search, we failed to identify a block-matching algorithm that can be used in our framework at training time and thus could not include it in the end-to-end optimization process.

We follow a different and novel approach for motion compensation that reuses the optical flow obtained by animation keypoints to perform spatial alignment between two consecutive residual frames R_{t-1} and R_t. Specifically, we apply the predicted optical flow between the adjacent decoded frames ($0^f_{t-1\rightarrow t}$) through a grid sampling operation. Note that the information required to predict the flow between the target frames is extracted from the decoder buffer hence the residual motion prediction introduces no additional coding cost. During training, the dense motion generator is simultaneously optimized to predict the optical flow between the reference frame and the target frames ($O^f_{0\rightarrow t}$) as well as the motion between neighboring target frames ($O^f_{t-1\rightarrow t}$). Our hypothesis is that learning temporal residual coding with a spatial alignment function reduces the effective coding cost of the frame residuals.

C. Architecture Details

The keypoint detector (KPD), dense motion generator and frame generation networks are similar to prior works [7], [13], [14], i.e., generative autoencoder networks optimized to detect a sparse set of facial landmarks, predict a dense motion field between two frames using one as a reference and landmarks extracted from each frame and a generative autoencoder for frame reconstruction given an approximate feature representation of the target frame. The residual coding networks follow the hyperprior [1] formulation dimensioned for low bitrate compression. Specifically, we propose a latent dimension of 48x8x8 feature maps mapped to 5 bitrate levels using learned gain vectors [16]. The input dimensions of the spatial and temporal residual coding frameworks are adjusted accordingly to account for the variable number of input features for the analysis network g_a and the conditioning network g_p. The final element of our reconstruction process is a low-complexity refinement network. This is inspired by the classical UNet Architecture and is trained as a post-processing element to the proposed coding framework.

IV. RESULTS AND DISCUSSION

A. Model Training

The training dataset consists of 18k talking-head videos from the VoxCeleb2 dataset. The RDAC+ framework is trained in two steps as follows:

Face Animation. The animation-coding framework (DAC) is trained with 10 random samples per video, i.e., 180k samples per epoch for 50 epochs. This step uses the Adam optimizer with a learning rate of $2e-4$ and β parameters ($\beta_1 = 0.5$, $\beta_2 = 0.999$) using a batch size of 64. The input images are sampled in pairs, i.e., one reference frame and one target frame with random temporal distance between the frames. The data augmentations described in [7], [13], [14] are applied.

Residual Compression. We train the residual coding frameworks for 15 epochs after which only the refinement network is optimized for an additional 5 epochs. The data sampling includes 10 random samples per video in each epoch. We use the Adam optimizer with a learning rate of $1e-4$ and the MultiStepLR scheduler. Each batch sample sample consists of four frames, *i.e.*, one reference frame and three neighboring frames. The reference frame is used to animate all the target frames. The frame residual of the first target is compressed with the spatial difference coder. Subsequently, the reconstructed spatial residual is used as a temporal predictor for the residual frames in the second and third target frame. This training process emulates the temporal coding process expected at inference time. The loss value is computed as the average rate-distortion loss for the three target frames with msVGG [17] as the distortion metric.

B. Benchmark Methods and Evaluation Dataset

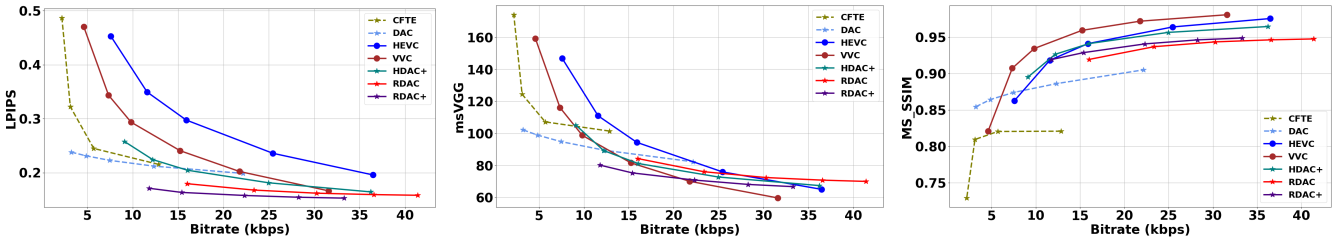
We evaluate the performance of our proposed framework against the GFVC methods, HEVC and VVC anchor codecs as proposed under the JVET-AH0114 test conditions. HDAC uses the architecture proposed in the original work [12] but uses the improved optimization conducted as part of the JVET-AH0114 standardization effort and is thus referred to as HDAC+ in this evaluation. The test videos are considered to have sufficient diversity in pose, expressions and talking head motion patterns. We use perceptual quality metrics that measure reconstruction accuracy at different levels of abstraction, from pixel level to higher-level features. We use FSIM and MS-SSIM [2] as low-level, pixel fidelity metrics. Additionally, learning-based metrics, e.g., msVGG and LPIPS [18] computed on a frame-by-frame basis are included in the evaluation process. Additionally, we include the learning-based DISTS [19] metric which measures the texture and style preservation in reconstructed videos.

C. Qualitative Evaluation

In Figure 2 we show through visual examples the reconstruction quality of RDAC+ relative to conventional codecs such as HEVC and VVC. We report visual comparisons at

[2]https://gitlab.com/wg1/jpeg-ai/jpeg-ai-qaf

Fig. 2: **Visual Comparison of coding results.** A qualitative comparison of our proposed coding framework shows significant quality improvement of RDAC+ over VVC, HEVC and other GFVC methods. In the top example, we observe that RDAC+ has a better color reconstruction compared to RDAC. In the bottom example, RDAC+ has a higher pixel fidelity around the mouth region. These fine-scale improvements in these facial details across the reconstructed videos contributes to the notable gains in coding efficiency of our framework as shown in Fig. 3



(a) (b) (c)

Fig. 3: **Average RD Performance 15 VoxCeleb2 test sequences:**. Our framework demonstrates substantial improvement over RDAC as well as VVC and HEVC. Further, it shows a reliable approach to achieving bitrate variability within the GFVC coding framework.

between 12 kbps and 20kbps to accomodate animation-only codecs which are limitated in the range of achievable bitrates. Being a hybrid codec, RDAC+ can instead be used at this bitrate and can significantly improve visual quality compared to purely animation-based schemes. Specifically, we observe that RDAC+ has a better perceptual reconstruction quality than HEVC and VVC. We observe a notable perceptual quality difference between RDAC and RDAC+ especially when we inspect the fine-scale details on the mouth and eyes as well as the color and texture details.

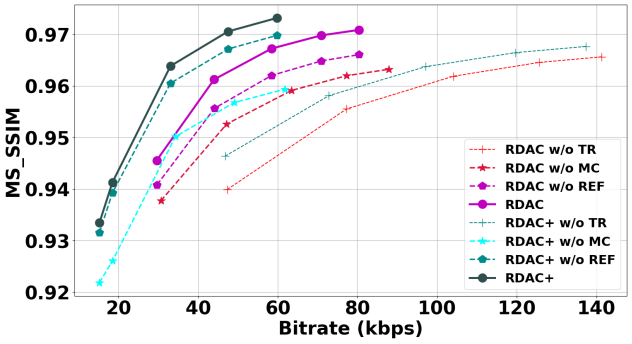
D. Rate-Distortion Performance

We present the BD-BR gains of our framework in Tab. I against anchor methods with sufficient overlap. The missing numbers when comparing against HEVC are for cases where the curves intersect leading to a failure in BD-BR computation. The highest bitrate savings are observed with respect to the perceptual metrics. This is a characteristic of all the GFVC frameworks. However, our framework achieves this in a bitrate range that is not easily achievable with the methods that use animation only such as CFTE and DAC. Further, we also observe gains in coding efficiency for previous hybrid methods such as RDAC and HDAC which similarly proposed to extend the range of achievable bitrates. Relative to RDAC our framework achieves over 20% bitrate savings on MS-

TABLE I: **Bjontengaard-Delta Bitrate (BD-BR) Performance of our proposed framework (RDAC+) relative to HDAC, RDAC, HEVC and VVC**

	FSIM	MS-SSIM	LPIPS	msVGG	DISTS
RDAC	-33.77	-22.51	-45.15	-38.05	-55.76
HEVC	-	-	-78.18	-55.23	-67.80
HDAC+	33.73	30.77	-61.43	-25.03	-25.79
VVC	17.73	23.58	-76.25	-38.01	-70.25

SSIM and FSIM and significantly narrows the gap with HEVC and VVC in this bitrate range. Note that HDAC+ builds on a HEVC base layer and thus has a robust performance on pixel fidelity metrics despite being a generative model as well. However, our method still achieves modest gains relative to HDAC+ in terms of perceptual metrics. We attribute this to the end-to-end optimization of entire RDAC+ framework, which is not the case for HDAC+ which requires offline processing of a base layer from the HEVC codec. Note that an evaluation process involving a convex hull search with a large number of configuration parameters such as previously done for RDAC [13] would produce RD curves showing a much higher coding gain. However, to remain consistent with the JVET-AH0114 evaluation protocol we do not conduct a such a convex hull search. Figure 3 shows the rate-distortion curves for a few of the evaluated metrics. Due to additional bitrate required for residual coding, the RDAC+ codec minimum bitrate is around 10 kbps. An obvious way to reduce the bitrate is deactivating the residual coding, which essentially reverts RDAC+ to DAC, i.e., an animation only codec. In practice, at low bitrates, RDAC, RDAC+ can implement a mode decision scheme where residual coding and pure DAC are put in competition. In Figure 4 we show the relative contribution of various coding tools that are applied for RDAC and RDAC+. We observe that each additional coding or reconstruction element introduces gains in coding efficiency. However, the conditional residual coding mechanism ensures that RDAC+ has a higher coding efficiency through all the optimization steps.



(a)

Fig. 4: **Coding gains as a result of Temporal residual (TR) coding, Motion Compensation (MC) and Refinement Network (REF). The RD metrics are computed on the first 64 frames of the test videos.**

V. CONCLUSIONS

Animation-based compression offers the possibility to transmit videos with very low bitrate. However, it is often limited to reconstructing the outputs at a fixed quality level, cannot scale efficiently when higher bandwidth is available and does not compress efficiently temporal redundancies in the signal. In this paper, we propose a coding scheme that advances the predictive coding architecture RDAC. In order to efficiently exploit the both spatial and temporal dependencies to achieve a coding gain, our framework proposes a motion compensation strategy in temporal residual coding. The experimental evaluations show the potential it provides a more principled approach to residual coding within the low-bitrate range that is considered desirable for animation-based compression frameworks. The framework demonstrates significant gains in perceptual quality relative to conventional codecs such as VVC and HEVC especially at low bitrate. Simultaneously it achieves significant gains in pixel fidelity when compared to prior animation-based coding frameworks. We propose an exploration of model scalability towards higher resolution, dynamic bitrate allocation in residual coding and complexity reduction for future research.

REFERENCES

[1] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *ICLR*, 2017.

[2] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "DVC: An end-to-end deep video compression framework," in *CVPR*, 2019, pp. 11 006–11 015.

[3] J. Li, B. Li, and Y. Lu, "Deep contextual video compression," *NeurIPS*, vol. 34, pp. 18 114–18 125, 2021.

[4] T. Ladune and P. Philippe, "AIVC: Artificial intelligence based video codec," in *IEEE ICIP*. IEEE, 2022, pp. 316–320.

[5] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE TCSVT*, vol. 22, no. 12, pp. 1649–1668, 2012.

[6] B. Bross, J. Chen, J.-R. Ohm, G. J. Sullivan, and Y.-K. Wang, "Developments in international video coding standardization after AVC, with an overview of versatile video coding (VVC)," *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1463–1493, 2021.

[7] G. Konuko, G. Valenzise, and S. Lathuilière, "Ultra-low bitrate video conferencing using deep image animation," in *ICASSP 2021-2021 IEEE ICASSP*. IEEE, 2021, pp. 4210–4214.

[8] T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *CVPR)*, 2021, pp. 10 039–10 049.

[9] M. Oquab, P. Stock, O. Gafni, D. Haziza, T. Xu, P. Zhango, and O. Celebi, "Low bandwidth video-chat compression using deep generative models," in *CVPR*, 2021.

[10] B. Chen, Z. Wang, B. Li, R. Lin, S. Wang, and Y. Ye, "Beyond keypoint coding: Temporal evolution inference with compact feature representation for talking face video compression," in *Data Compression Conference (DCC)*. IEEE, 2022, pp. 13–22.

[11] Z. Chen, M. Lu, H. Chen, and Z. Ma, "Robust ultralow bitrate video conferencing with second order motion coherency," in *MMSP*. IEEE, 2022, pp. 1–6.

[12] G. Konuko, S. Lathuilière, and G. Valenzise, "A hybrid deep animation codec for low-bitrate video conferencing," in *ICIP*, 2022, pp. 1–5.

[13] G. Konuko, S. Lathuilière, and G. Valenzise, "Predictive coding for animation-based video compression," in *ICIP*. IEEE, 2023, pp. 2810–2814.

[14] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *NeurIPS*, vol. 32, 2019.

[15] F. Brand, J. Seiler, and A. Kaup, "Conditional residual coding: A remedy for bottleneck problems in conditional inter frame coding," *IEEE TCSVT*, vol. 34, no. 7, pp. 6445–6459, 2024.

[16] Interdigital.inc, "https://interdigitalinc.github.io/compressai/," 2024.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, Y. Bengio and Y. LeCun, Eds., 2015.

[18] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595.

[19] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE TPAMI*, vol. 44, no. 5, pp. 2567–2581, 2020.