Robust Strategy of Video Compression Using SRVC

M. Sita Ram Muni Reddy[1], A. Shishir[2] and Dr V Kishen Ajay Kumar[3]

[1] Undergraduate Student, Institute of Aeronautical Engineering(Autonomous), JNTUH, Telangana - 500043

[2] Undergraduate Student, Institute of Aeronautical Engineering(Autonomous), JNTUH, Telangana - 500043

[3] Associate Professor, Institute of Aeronautical Engineering(Autonomous), JNTUH, Telangana - 500043

Abstract: - Video compression is a basic part of Web video conveyance. Late work has demonstrated the way that profound learning procedures can match or beat human-planned calculations, yet these techniques are essentially less process and power efficient than existing codecs. This paper proposes a novel method for video compression using a neural super-resolution model that adapts to the content dynamically. By leveraging lightweight neural networks alongside traditional video codecs, the proposed method significantly reduces the bandwidth required for video transmission without compromising on quality. The approach encodes video content into two streams: a content stream with low-resolution frames compressed using standard codecs and a model stream that periodically updates a neural network customized to enhance short segments of the video. Experimental results demonstrate that this method outperforms both traditional and existing neural video compression schemes in terms of bit-rate efficiency and video quality, while also maintaining real-time decoding capability.

Keywords: **H.265/H.264 codec, FFmpeg, CNN(Convolution Neutral Network), Generative Adversarial Networks (GANs).**

1. INTRODUCTION

Over the years there has been a vast expansion in video traffic. Video will address over 80% of all Web traffic. Video development is so data transfer capacity raised that during flood periods, for example, the secret seemingly forever of the pandemic, Netflix and YouTube expected to smother video quality to diminish overheads [1]. Further, while PDAs support 1080p goals nowadays, cell networks are right now tormented by uninformed move cutoff and moderate changes in various regions of the planet. There needs to be a strong video strain to diminish data transmission utilization without consenting to less quality, which is more key than at any time in late memory. While the interest in video content has broadened through the long stretch, the methods used to pack and send video have normally happened as previously. Contemplations, for example, applying Discrete Cosine Changes (DCTs) to video blocks and figuring improvement vectors, which were made various years sooner, are at this point being used today [1]. Unquestionably, even the most recent H.265 codec works on these equivalent examinations by setting variable block sizes. Consistent endeavors to furthermore

cultivate video pressure have gone to critical figuring out a workable method for finding the fantastic relationship between the bits of a video pressure pipeline. These strategies have had moderate accomplishment at outsmarting current codecs, however, they are less cycle and power-valuable. We present SRVC, another framework especially critical for cell affiliations and low bitrate conditions, that hardens existing strain calculations with a lightweight [2], content-adaptable super-unbiased (SR) frontal cortex network that from an overall perspective maintains execution with low assessment cost. SRVC packs the information video into two streams a substance stream and a model stream, each with another bitrate that can be controlled wholeheartedly by the other stream. The substance stream depends upon a codec, The model stream encodes a period differentiating SR frontal cortex affiliation, which the decoder uses to help de-pressurize outlines from the substance stream. SRVC utilizes the model stream to practice the SR network for short areas of video proficiently. This makes it conceivable to utilize a little SR model, including a couple of convolutional and upsampling layers. Applying SR to work on low-bitrate stuffed video isn't new. AV1, for example, has a mode (reliably utilized in low-bitrate settings) that encodes outlines at low goals and applies an upsampler at the decoder. While AV1 depends upon standard bi-cubic [3] or bi-linear commitment for up-sampling, late ideas have shown that learned SR models can endlessly outwork the possibility of this strategy. In any case, these systems depend upon conventional SR [4] mind networks that are supposed to sum up an impressive number of information pictures. These models are monsters and can usually recreate a few edges each resulting even on first-in-class GPUs[5]. Regardless, in many use cases, speculation isn't required. Specifically, we from time to time approach the video being compacted early. We want to truly lessen the capriciousness of the SR model in such applications by practicing it (it could be said, overfitting it) to short areas of video. To make this figure work, we should guarantee that the model stream above is low. Undoubtedly, even with our little SR model, resuscitating the whole model typically would consume a high bitrate, fixing any strain benefit from chopping down the target of the substance stream. SRVC handles this test by cautiously picking a little piece of cutoff points to empower each part of the video, utilizing a "propensity facilitated" coordinate-fall system that restricts that by and large impact model quality. Our central finding is that an SR frontal cortex network changed in this manner all through a video can give such a lift to quality, that including a model trade near the compacted video is more convincing than relegating the whole piece stream to content. We propose a unique twofold exchange method for managing video online that joins a period-changing SR model with compacted low-objective video conveyed by a standard codec [6]. We encourage a heading plunge strategy to revive only an irrelevant piece of model limits for each two or three-second part of a video with low above. We propose a lightweight model with spatially-flexible parts, arranged unequivocally for content-express SR. Our model runs dynamically, taking only 11 ms (90 fps) to make a 1080p edge on a NVIDIA V100 GPU. In assessment, DVC takes 100s of milliseconds at a comparable objective. That is the very thing we show, in low bitrate frameworks, to achieve a comparative PSNR, SRVC requires only 20% of the bitrate as H.265 in its languid encoding mode 1, and 3% of DVC's pieces per pixel. SRVC's quality improvement connects across all housings in the video. Shows visual models differentiating the SRVC and these benchmark approaches at serious or higher bitrates.

2. METHODOLOGY

2.1 Codecs:

Earlier work has broadly concentrated on video encoders/decoders like H.264/H.265 [4] and AV1. These codecs answer available planned calculations that exploit the worldly and spatial redundancies in video pixels, yet they can't adjust to explicit recordings. Existing codecs are especially powerful when utilized in sluggish mode for disconnected pressure. By the by, SRVC's blend of a low-goal H.265 stream with a substance-variable SR model

outflanks H.265 at high goal, even in its sluggish mode. A codec like AV1 gives the choice to encode at low goal and up-sampled utilizing bicubic insertion [3]. SRVC's learned model gives a lot bigger improvement.

2.2 Traditional Video Compression Techniques:

2.2.1. Block-Based Methods: Most traditional codecs use block-based compression, where frames are divided into blocks, and similarities within and between these blocks are exploited.
2.2.2Transform Coding: Techniques such as the Discrete Cosine Transform (DCT) and the Discrete Wavelet Transform (DWT) are commonly used to convert spatial do main data into frequency domain data, which can be more efficiently compressed.

2.3 Super-Resolution (SR) Techniques:

2.3.1 Super-resolution is the process of enhancing the resolution of an image or video. It can be used to reconstruct high-resolution images from low-resolution inputs[6]

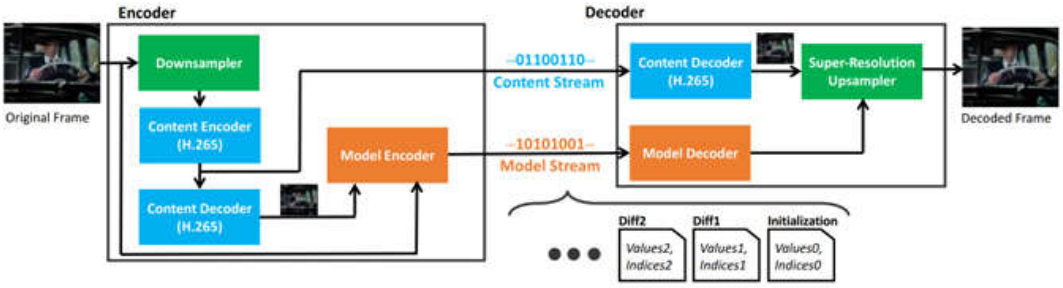

Fig. 1. Video is encoded into two bitstreams by SRVC. The content stream uses the current codec to encode low-resolution video that has been downsampled. A lightweight super-resolution neural network tailored for brief video segments receives periodic updates from the model stream.

2.3.2 Deep Learning-Based SR:

Recent advances in deep learning, particularly Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs), have significantly improved the performance of SR techniques. Notable models include SRCNN, VDSR, and ESRGAN.

2.4 Content-Adaptive Techniques:

2.4.1 Adaptive Compression: Content-adaptive methods adjust compression parameters based on the complexity of the video content. This can lead to better quality preservation for complex scenes while achieving higher compression rates for simpler scenes[7].
2.4.2 Perceptual Metrics: Techniques often utilize perceptual metrics to determine content complexity and adapt compression strategies accordingly.

2.5 Combining Super-Resolution with Compression:

2.5.1 Concept: The integration of SR with video compression involves downscaling the video during compression and applying SR techniques during decompression to restore the original resolution[7]
2.5.2 Advantages: This approach can achieve significant compression rates while maintaining high visual quality, especially with advanced SR models that can reconstruct fine details lost during downscaling[2].

2.6 Content Stream:

The encoder is the goal of the info video outlines by a variable in each aspect by utilizing region-based down testing, bringing about low-goal (LR) outlines. These LR outlines are then encoded with a standard video codec to deliver the substance bitstream. The decoder then, at that point, depressurizes this bitstream with the equivalent codec to reproduce the LR outlines. Nonetheless, because video codecs are not lossless, the LR outlines reproduced by the decoder may not impeccably match the LR outlines created by the encoder.

2.7 Model Stream:

The course of video pressure in this framework is organized around two essential bitstreams: one that encodes low-goal (LR) content and one more that encodes the boundaries of the super-goal (SR) model. This second bitstream is essential as it empowers the decoder to upscale the low-goal outlines into high-goal outlines. t0,1,..., N1, the SR model is tweaked or adjusted explicitly to the edges inside that section. This transformation happens during the video encoding stage. The SR model is prepared to plan the low-goal outlines from the fragment to high-goal outlines, accordingly figuring out how to improve video quality successfully for that section of the video. To guarantee computational proficiency and smooth changes between fragments, the model variation is successive. This implies that the preparation for each section, from the past fragment. Subsequently, the model persistently advances, calibrating its capacity to upscale the video as the substance changes over the long run. The second bitstream then encodes the succession of SR model boundaries =t For all portions. The encoding system starts with the full arrangement of boundaries for the main portion Furthermore, for ensuing fragments, just encode the progressions in the model boundaries. These progressions are addressed as t=tt1, which mirrors the distinction between the boundaries of the ongoing portion and the past one. This strategy for encoding just the progressions between fragments essentially decreases how much information is expected to send the SR model, streamlining the bitstream for transmission capacity effectiveness. On the disentangling side, the decoder refreshes the SR model boundaries each. seconds. It utilizes the recently obtained boundaries t1t1 to figure out the ongoing section's boundaries t=t1+t. This steady update system guarantees that the SR model remains lightweight while keeping up with high precision in upscaling the low-goal casings to their high-goal partners. The model stream acquaints extra above with the compacted bitstream. To relieve this, we will make a smaller model that is improved for content-explicit super-goal and plan a calculation that essentially limits the above model variation. This is accomplished by preparing just a small subset of model boundaries that greatly affect super-goal quality inside each section [8].

2.8 SR Model Architecture:

Existing super-goal (SR) models for the most part de pend on huge, profound brain organizations, for example, the EDSR, which has 43 million boundaries across north of 64 layers. This intricacy allows these model tests to be carried out in a continuous video decoder. Moreover, adjusting such an enormous profound brain organization to explicit video content and communicating it to the decoder would bring about huge above. We present another lightweight design that stays minimized and shallow while being profoundly compelling for content-based variation. Our model draws motivation from traditional calculations, for example, bicubic upsampling, which commonly utilizes a solitary convolutional layer with a decent portion for picture upsampling. We hold this fundamental construction however supplant the proper portion with spatially-versatile pieces custom-made to various locales of the information outline. Our methodology

segments each casing into patches and utilizes a shallow convolutional brain organization

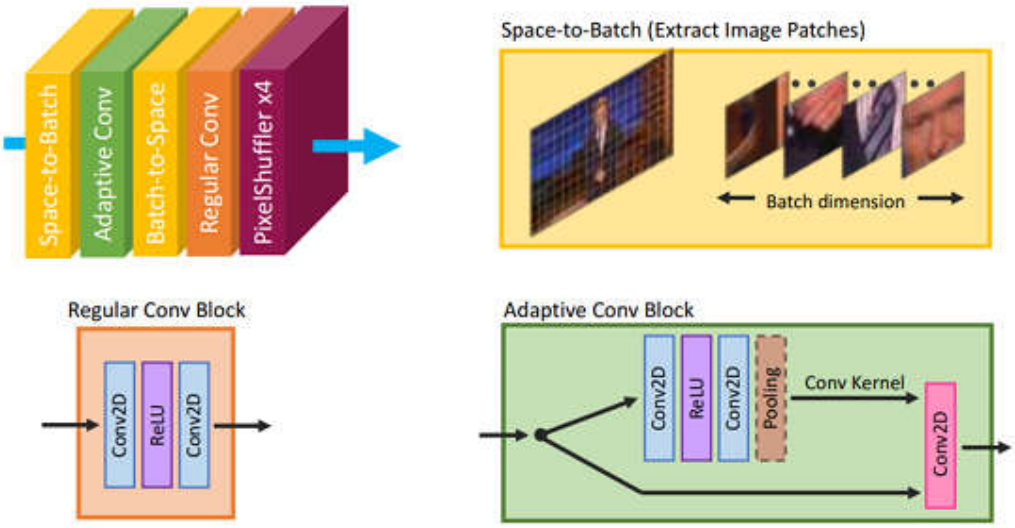

Fig. 2. SRVC model architecture

(CNN) to produce unmistakable versatile parts for each fix [9]. w = f(x), y = (w x) We utilize a two-layer convolutional brain organization (CNN) to demonstrate. In the first place, we process the component fixes and afterward reassemble them utilizing a cluster-to space activity. From that point onward, we apply another two layer CNN, trailed by a pixel shuffler to upscale the substance to a higher goal. All convolution layers utilize a 3x3 portion size, except the principal layer of the normal block, which utilizes a 5x5 part size.

2.8 Model Adaptation Algorithm:

We train the SR model for each section by limiting the L2 misfortune between the model's result and the comparing high-goal outline across all casings in the fragment. Officially, the misfortune capability is characterized as:

$$L(\Theta_t) = \frac{1}{n|F_t|} \sum_{i=1}^{n} \sum_{j=1}^{|F_t|} \|Y_{ij} - X_{ij}\|^2$$

where |Ft| is the quantity of edges in the t-th portion, each with n pixels. Yij and Xij address the worth of the I-th pixel in the j-th edge of the decoded high-goal yield and the first high-goal input outline, separately. During preparation, we arbitrarily crop the examples to half of their size in each aspect. The Adam enhancer is utilized with a learning pace of 0.0001 and energy rot paces of 0.9 and 0.999 for the first and second minutes, separately. To reduce the model stream bitrate, we update only a subset of the model parameters across video segments. Our strategy involves focusing on the parameters that most significantly impact the model's accuracy. Specifically, for each new segment, we update the parameters with the largest gradient magnitudes. Toward the beginning of another section, we save a duplicate of the model. We then perform one emphasis of preparing over all edges in the new portion. From this, we recognize the fraction of boundaries with the biggest greatness of progress. We return the model boundaries to the saved duplicate and apply Adam refreshes just to the chosen boundaries, leaving the leftover boundaries unaltered. Encoding the Model Stream: To additional pack the model stream, we send just the progressions in model boundaries with each update. The model updates are encoded into a bitstream by recording the lists of the boundaries and their related changes. The encoding system in SRVC is lossless: both the encoder and

decoder update a similar subset of boundaries during each update. To refresh a small portion of the boundaries in a model with M float16 boundaries, we want a normal bitrate of at most (16 + log(M)) × ηM/τ to address the deltas and lists each τ seconds. For example, with a model size of M = 2.22 million boundaries (F=32, see Table 2), τ = 10 seconds, and η =0.01, the required bitrate is just 82 Kbits/sec to encode the model transfer for 1080p video. In correlation, Netflix suggests a transmission capacity of 5 Mbits/sec for a 1080p goal [4]. Further pressure of the model stream can be accomplished utilizing lossy pressure strategies or by progressively changing or the update recurrence in light of scene changes. Preparing the SR model for the 1080p goal and encoding the updates into the model stream at present requires around 12 minutes of the moment of video with our enhanced execution. Regardless of this, the lightweight idea of our model permits us to divide a V100 GPU between five simultaneous cycles without huge lulls effectively. Thus, the general throughput on the V100 GPU is roughly 2.5 minutes of preparing each moment of content. This length is sensible for disconnected pressure situations, where content suppliers approach recordings a long time before they are seen. We accept there is extensive potential to additionally speed up the encoding system utilizing standard strategies, like preparation on examined outlines as opposed to all edges, and through extra designing enhancements. We intend to investigate these conceivable outcomes in future work.

3. IMPLEMENTATION

3.1 Dataset:

Existing video datasets, for example, JCT-VC [9], UVG, and MCL-JCV, which contain two or three hundred edges (roughly 10 seconds of video), are inadequate for assessing SRVC's substance versatile super-goal. Thus, we prepared and tried SRVC utilizing a custom dataset involving 28 downloadable Vimeo short movies and 4 full-length recordings from the Xiph Dataset. We managed all recordings to be 10 minutes long and switched them over completely to the 1080p goal in Crude organization from their unique 4K goal and MPEG-4 configuration, utilizing region-based addition. These 1080p casings act as our high-goal source outlines. We then re-encoded every video's Crude casings at different characteristics or Consistent Rate Variables (CRFs) utilizing H.264/H.265 [4] to control the bitrate. Furthermore, we downsampled the recordings to 480p utilizing region-based interjection and encoded them at various CRFs with H.265 to accomplish changing degrees of pressure. The SR model in SRVC is prepared to plan each low-goal video at a particular CRF to its comparing high-goal 1080p video at the best. 1080p H.264/H.265 [6] We re-encode each 1080p video at different Steady Rate Variables (CRFs) utilizing the sluggish preset with the ffmpeg device and libx264/libx265 codecs

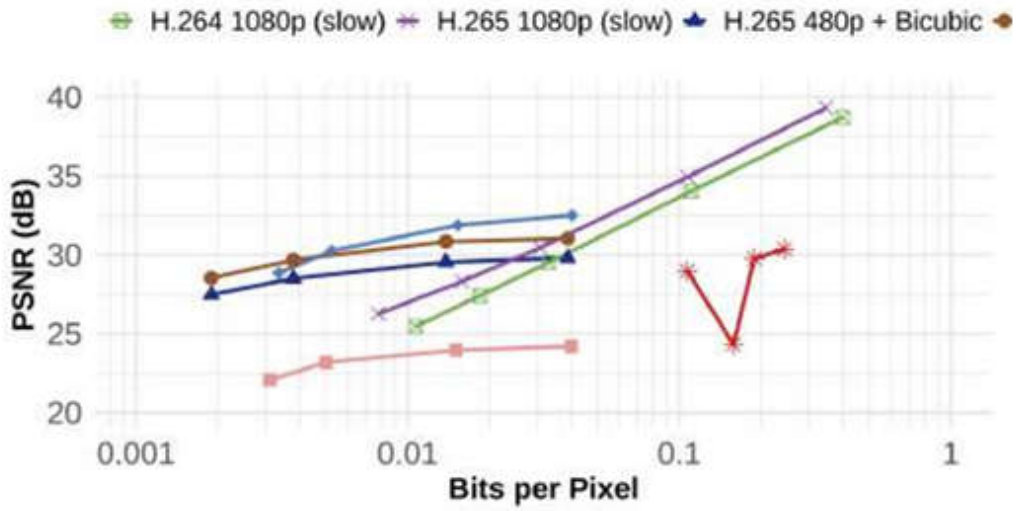

Fig. 3. A compromise between video quality and pieces per pixel for various methodologies on three long recordings from the Xiph dataset.

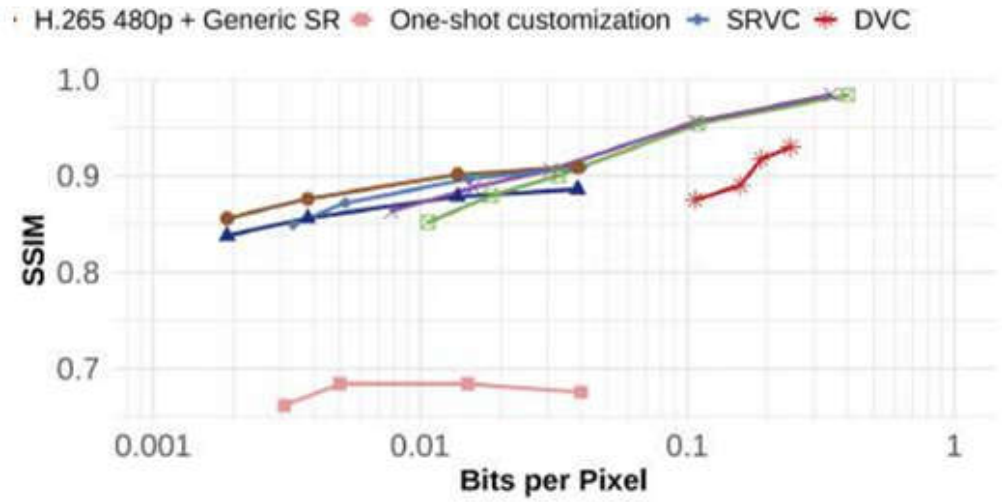

Fig. 4. Bitrate utilization of SRVC with content-versatile streaming is diminished to 16% of current codecs and around 2% of start-to-finish pressure plans, for example, DVC. Despite having video quality that is like SRVC, the nonexclusive SR approach needs constant usefulness.

3.2 Bicubic Upsampling:

We use FFmpeg and the libx265 codec to downsample the 1080p unique recordings to low-goal 480p at various CRFs utilizing region-based introduction and the sluggish preset. This technique's bitrate comes exclusively from encoding the down sampled 480p edges with H.265. We then up sample these 480p edges back to 1080p utilizing bicubic interjection, con fining the bitrate decrease from encoding at lower resolutions [10].

3.3 Generic SR:

Rather than bicubic up sampling, we utilize a DNN-based super-goal model to upscale the 480p edges to 1080p [5]. This up sampling system takes around 50 milliseconds for every casing, which is roughly multiple times slow than SRVC. We use a pre-prepared designated spot from a nonexclusive picture corpus. This approach utilizes just a

substance stream at 480p encoded with H.265 [4], bringing about a piece for every pixel esteem identical to the bicubic case.

3.4 One-Shot Customization:

We survey a rendition of SRVC that utilizes a lightweight SR model without model transformation. Here, we train the SR model once (a single shot) utilizing the whole 1080p video

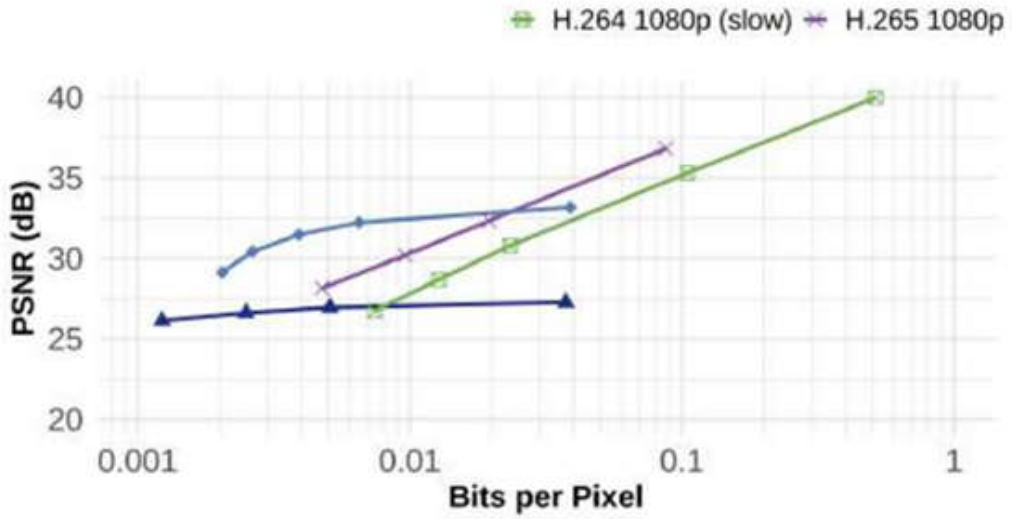

Fig. 5. A trade-off between bits-per-pixel and video quality for various strategies on 28 Vimeo videos

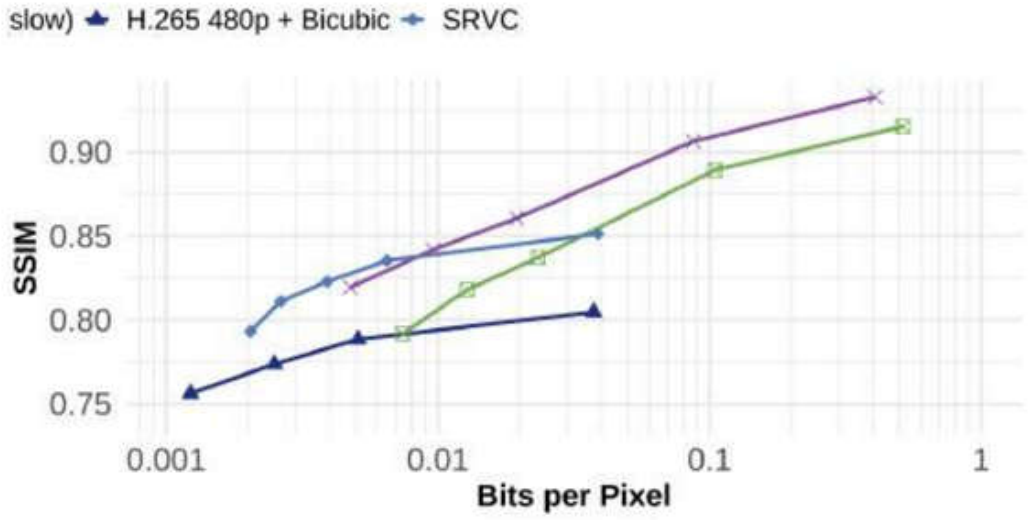

Fig. 6. To attain a PSNR of 30db, SRVC needs 10% and 25% of the bits per pixel needed in the slow modes of H.264 and H.265.

and encode it in the model stream toward the beginning, before any LR content. The substance stream comprises the 480p H.265 video, while the model transfer incorporates a solitary beginning model customized to the whole video length. The model above is disseminated across the video and added to the substance bitrate while working out the complete pieces per pixel esteem.

3.5 SRVC:

We assess SRVC, which utilizes a similar introductory SR model as A single Shot Customization however adjusts the model occasionally to the latest 5-second portion of the video. Preparing includes utilizing irregular harvests from each reference outline inside the video section. The substance stream for SRVC depends on standard H.265 encoding, while the model stream is refreshed like clockwork with our angle-directed procedure, encoding just the progressions in boundaries with the biggest slopes. The complete pieces per-pixel esteem is determined by adding the bitrate of the model stream to that of the substance stream, including the above of the underlying full model.

3.6 Model Procedure:

Our model consolidates 32 result highlight directs in the versatile convolution block, adding up to 2.22 million boundaries. Notwithstanding, the model stream refreshes just 1% of these boundaries, and this happens at regular intervals. We explore different avenues regarding various quantities of result highlight channels, shifting parts of refreshed model boundaries, and different update spans to survey their effect on SRVC's exhibition [10].

3.7 Metrics and Color Space:

We assess the typical Pinnacle Signal-to-Clamor Ratio(PSNR) and Underlying Likeness Record Measure(SSIM) across all casings after disentangling and upsampling. PSNR is determined in light of the mean square blunder of all pixels in the video, with the pixel-wise mistake estimated in the RGB variety space. SSIM is processed as the normal closeness between the decoded outlines and their related high-goal firsts. Moreover, to catch varieties in outline quality that can influence client experience, we give a combined dispersion capability (CDF) of both PSNR and SSIM across all casings in the video. We figure the substance bitrate for all approaches utilizing H.264 [4] at both 1080p and 480p goals with FFmpeg. For strategies that remember a model transfer for expansion to video outlines, we compute the model stream bitrate in light of the all-out number of model boundaries, the part refreshed in every stretch, and the update recurrence. The substance and model stream bitrates are consolidated to decide a solitary piece for every pixel metric. It's quite significant that our assessed bits-per-pixel range is a significant degree lower than results detailed in past examinations, as our methodology targets low-bitrate situations and contrasts and the sluggish method of H.264 [4], which is more effective than the" quick" and" medium" modes. We plot PSNR and SSIM measurements across various pieces per pixel values to think about different plans. Since SRVC performs derivation on decoded outlines as they are delivered to clients, its SR model should work progressively. To survey its common sense, we likewise look at SRVC's speed in outlines each second to other learning-based approaches.

3. RESULTS AND DISCUSSIONS

Compression Execution: For DVC, we present outcomes from the least bitrate model accessible, which works at 4.97 Mbps—altogether higher than the 200 Kbps bitrate of different plans in this model. To assess pressure viability across a wide scope of pieces per-pixel values, we dissect the PSNR and SSIM measurements for various techniques on three expanded Xiph recordings, as displayed in. The pieces per pixel metric records both substance and model commitments for approaches using a model stream for SR. Note that we don't report bitrate contortion measurements for A Single Shot Customization, as its PSNR results don't cover essentially H.265. As displayed in Fig. 7, SRVC conveys PSNR levels similar to the most recent H.265 standard while utilizing altogether fewer pieces per pixel. For example, accomplishing a PSNR of 30 dB with SRVC requires just 0.005

pieces per pixel, though H.265 and H.264 codecs [4], even at their slowest settings, need over 0.03 pieces per pixel. As per BD Rate and BD-PSNR measurements, SRVC gives a typical improvement of 3.41 dB over the H.265 slow preset at 1080p

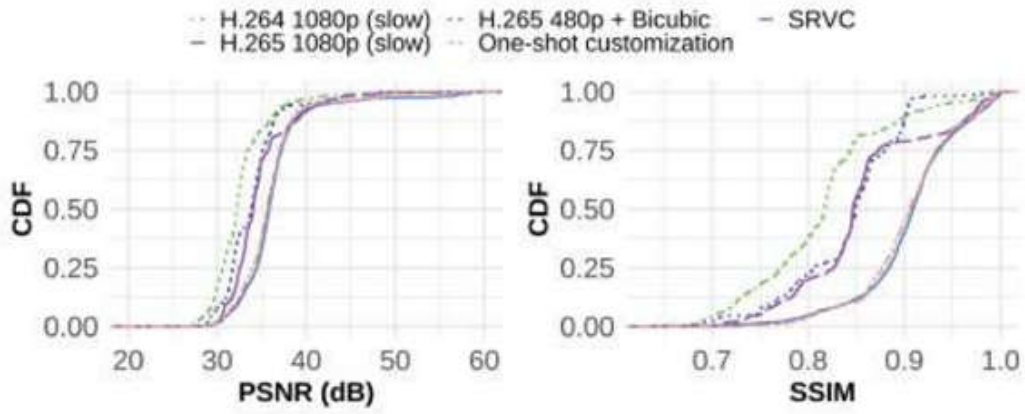

Fig. 7. CDF of PSNR and SSIM enhancements with SRVC across all video outlines at pieces for every pixel of 0.002. The quality upgrade from SRVC isn't restricted to just those approaches that follow a model update

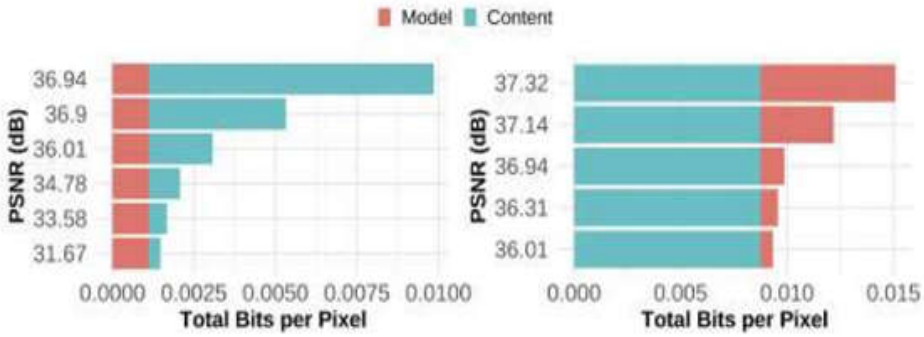

Fig. 8. Effect of shifting pieces per pixel for the substance stream for a proper model bitrate as well as the other way around. Expanding the pieces per pixel for the low-goal H.265 content stream further develops PSNR, especially at low bitrates. At higher substance bitrates, expanding the model bitrate by communicating more model boundaries further develops PSNR

for the equivalent bitrate or requires only 20% of the bitrate to accomplish identical PSNR [11]. A solitary shot customization performs more horrendously than direct bicubic inclusion. This is because SRVC's custom SR model isn't sufficiently tremendous. To summarize, the entire video, be that as it may, is good for acquiring from a little segment. Unmistakably, to achieve a comparable PSNR, SRVC requires only 3% of the pieces per pixel expected by DVC, the beginning-to-end mind pressure plot. SRVC's SSIM is identical to current codecs, but fairly better for comparable pieces per pixel, particularly at higher bitrates. Moreover, SRVC defeats a regular SR approach (EDSR) by 0.8 dB in

PSNR and by 4.8% in BD-Rate [12]. Vigor of value upgrades: To decide if SRVC's improvements are restricted to only a couple of top-notch approaches promptly following model updates, we plot the combined dissemination capability (CDF) of PSNR and SSIM values across all edges of the Meridian video in Figure 6. We look at changed approaches at a piece for every pixel worth of roughly 0.002. Since DVC works at a lot higher pieces for each pixel, and EDSR performs ineffectively in this video, we reject the two strategies from the correlation. To start with, we see that both A single shot Customization and SRVC perform better compared to different plans. Further, this improvement happens over the edges in that no casing is all more awful off with SRVC than it is with the Defacto codec. More than 50% of the casings experience a 2 3 dB improvement in PSNR and a 0.05-0.0075 improvement in SSIM with the two variants of SRVC. Effect of number of Result Element Channels. Since SRVC down examples outlines at the encoder and afterward streams a model to the getting client who settles the decoded outlines, SRVC must perform deduction quickly enough to run at the casing pace of the video on an edge gadget with restricted handling power. Watchers need something like 30 fps for good quality. Thus, the induction time on a solitary casing can't stand to be longer than 33ms. The Meridian video has a casing pace of 60 fps.

5. Conclusion

In this review, we propose SRVC, a methodology that improves current video codecs by coordinating a lightweight, content-versatile super-goal model. SRVC gives video quality similar to cutting-edge codecs while accomplishing better pressure. This plan addresses an early step towards involving super-goal as a strategy for video pressure. Future work will target advancing the harmony between model intricacy and bitrate, further developing strategies for identifying scene changes and refining update spans, as well as making more proficient lightweight super-goal brain network structures.

REFERENCES

[1] *G. J. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the High-Efficiency Video Coding (HEVC) Standard," IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1649-1668, 2012.*

[2] *O. Rippel, S. Nair, C. Lew, S. Branson, A. G. Anderson, and L. Bour dev, "Learned Video Compression," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.*

[3] *Robert Keys. Cubic convolution interpolation for digital image processing. IEEE transactions on acoustics, speech, and signal processing, 29(6):1153–1160, 1981.*

[4] *H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 9, pp. 1103-1120, Sep. 2007.*

[5] *B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 136-144, 2017.*

[6] *G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "DVC: An End-to-End Deep Video Compression Framework," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11006-11015, 2019.*

[7] *E. Agustsson, D.Minnen, N. Johnston, J. Ball´ e, S. J. Hwang, and G. Toderici, "Scale-Space Flow for End-to-End Optimized Video Compression," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8503-8512, 2020.*

[8] *Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiy ong Gao. Dvc: An end-to-end deep video compression framework. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 11006–11015, 2019.*

[9] *Alexandre Mercat, Marko Viitanen, and Jarno Vanne. Avg dataset: 50/120fps 4k sequences for video codec analysis and development. In Proceedings of the 11th ACM Multimedia Systems Conference, pages 297–302, 2020.*

[10] *P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125-1134, 2017.*

[11] *W. Shi et al., "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1874-1883,2016.*

[12] *A. Mercat, M. Viitanen, and J. Vanne, "UVG Dataset: 50/120fps 4K Sequences for Video Codec Analysis and Development," in Proceedings of the 11th ACM Multimedia Systems Conference (MMSys), pp. 297 302,2020*