Adaptive ROI Encoding and Caching for Video Surveillance Streaming

Yung-Shun Chuang, Hsu-Feng Hsiao
Department of Computer Science,
National Yang Ming Chiao Tung University, Hsinchu, Taiwan

***Abstract*—Conventional video streaming techniques in security surveillance systems often utilize uniform bit rate strategies, leading to suboptimal resource allocation. We propose a novel streaming system that dynamically assigns variable bit rates to regions of interest (ROIs) within surveillance frames. By intelligently enhancing the quality of critical areas while reducing the quality of peripheral areas, the system significantly reduces network transmission and storage costs without compromising ROI clarity. The proposed system leverages a deep learning-based ROI detection module to effectively identify regions demanding intensive monitoring. An adaptive encoding scheme then assigns lower quantization parameters to ROIs, yielding higher quality, while applying higher quantization to non-ROIs to conserve bitrate. A dynamic quality enhancement module is integrated, significantly improving the quality of both foreground and background regions, thereby enhancing recognizability for surveillance personnel or machine analysis. A key innovation is a caching mechanism that exploits the high redundancy in surveilled background scenes across frames. By reusing enhanced background blocks from preceding frames, the caching mechanism accelerates the quality enhancement module with negligible quality loss. Extensive experiments validate the framework's superior rate-distortion performance. The proposed system's improved quality, optimized resource usage, and reduced storage make it a promising solution for advancing video coding in security and surveillance domains.**

I. Introduction

Video surveillance systems are experiencing rapid growth, with the market projected to expand by 9.4% globally from 2022 to 2027, reaching an estimated $76.4 billion, driven by increasing camera deployment for security applications [1]. However, the surging volume of video data presents critical challenges in efficient streaming and storage as system scale expands. Effective surveillance requires high-quality capture of specific regions of interest (ROIs), such as entry and exit points, to enhance recognition accuracy for both human monitors and machine vision algorithms. Yet, uniformly capturing non-ROI areas at high quality leads to excessive storage consumption, necessitating adaptive bitrate allocation for optimal system performance.

Video compression techniques leveraging Region of Interest (ROI) can be broadly classified into two methodologies: pre-processing and embedded encoding. The pre-processing approach employs non-uniform blurring in non-ROI regions, effectively minimizing irrelevant information and consequently reducing the volume of data requiring compression. Itti [2] utilized this technique by tracking prominent targets within videos and blurring non-ROI areas to achieve compression. The identification of these salient targets relied on sophisticated models of visual attention, crafted through bio-inspired methods and information theory principles.

In contrast, embedded encoding focuses on selectively enhancing the visual quality of specific regions by dynamically adapting encoding parameters, such as quantization settings. Zhu et al. [3] integrated human visual attention mechanisms with a perceptually-prioritized video compression scheme. By fusing spatial and temporal saliency maps, their approach enables region-specific adjustments to compression parameters, allocating higher bitrates to visually sensitive areas while conserving resources in less critical regions. Another application was proposed in [4], where saliency was considered as an important feature for object detection and identification in security surveillance.

Choi et al. [5] enhanced the YOLO9000 object detector with deep neural networks, introducing saliency maps for bit rate allocation, prioritizing object detection over traditional PSNR enhancement. Similarly, Galteri et al. [6] introduced a self-regulating video encoding method, utilizing a Bayesian framework for saliency estimation, optimizing encoder speed.

Super-resolution imaging presents another approach for online services, offering a solution to the bandwidth constraints inherent in video streaming. By strategically transmitting low-resolution videos, bandwidth usage is minimized, while the resolution is subsequently enhanced on the client side. NAS framework [7] leverages this concept, using client resources and super-resolution networks to optimize video quality while conserving bandwidth. Similarly, LiveNAS [8] provides a real-time framework that enhances live streams by dynamically adjusting server resources for better quality. Dejavu [9] takes a different approach, focusing on video conferencing. It deliberately transmits at lower quality to ensure responsive network performance, then exploits the visual consistency in video calls to upscale frame quality. Addressing the energy demands of super-resolution on mobile devices, NEMO [10] selectively applies enhancements to specific frames, striking a balance between optimal quality and device resource management.

In this paper, we introduce a novel system that dynamically allocates variable bit rates to different regions of interest (ROIs), enhancing the quality of critical areas while minimizing transmission and storage costs. The system features an ROI detection module that identifies regions requiring intensive monitoring, and employs adaptive encoding parameters to achieve discernible quality differentiation across

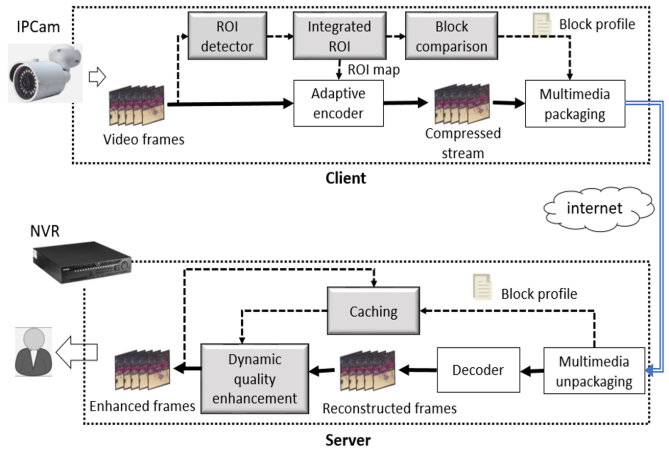

Fig. 1: Overview of the proposed ROI-based video streaming system architecture. The client-side IP camera captures video frames and performs ROI detection and adaptive encoding. The server-side NVR receives the encoded video stream along with block profile metadata, enabling dynamic quality enhancement and caching for efficient playback and storage.

various regions. A dynamic quality enhancement module further elevates the quality of both foreground and background regions, improving recognizability for surveillance personnel and machine analysis. A distinctive contribution of this system lies in its caching mechanism, which capitalizes on the high redundancy inherent in background scenes to accelerate the processing speed of the dynamic quality enhancement module. To evaluate compression efficiency, the traditional Peak Signal-to-Noise Ratio (PSNR) method, which calculates pixel differences assuming equal weight for all pixels, proves inadequate for assessing the quality compression efficiency of systems prioritizing ROI areas. Consequently, we adopt the Weighted PSNR (WPSNR) method [11], specifically tailored for evaluating the performance of such systems.

The remainder of this paper is structured as follows: Section II presents the proposed methods, followed by the simulation results in Section III, and the conclusion in Section IV.

II. Methodology

The proposed ROI-based streaming framework builds upon a foundational surveillance system architecture, divided into client and server components as illustrated in Fig. 1. The client side, comprising IP cameras, leverages the Hi3559 V100 SOC hardware chip for core system operations. Captured frames are directed to the ROI (Region of Interest) detection module and the adaptive encoder module. The ROI detection module is responsible for identifying user-defined target areas, with the resultant detections relayed to the integrated ROI module. This module's primary function is to amalgamate the dynamic regions encapsulating the target of interest, subsequently generating a comprehensive ROI map. The adaptive encoder modulates the quantization parameter (QP) in accordance with the spatial distribution of ROI positions within the ROI map, followed by hardware encoding of the frames. Concurrently, the block comparison module exploits the ROI map to assess inter-frame block similarity, yielding a block profile that encodes both ROI positions and block similarity metrics. The multimedia packaging component then transmits the encoded streams and block profiles to the remote Network Video Recorder (NVR) on the server side.

Upon reception at the server side, the multimedia unpackaging module extracts the encoded frames and block profiles from the incoming streams or recorded files. The dynamic quality enhancement module uses the information embedded in the block profiles to elevate the quality of the decoded frames. Furthermore, these enhanced frames are cached in a buffer, facilitating the reuse of blocks across frames. This innovative caching mechanism significantly reduces the execution time of the dynamic quality enhancement module, thereby improving system performance significantly.

The details of the important components are described in the following.

A. ROI Detection

The ROI detection module employs a deep learning CNN architecture specifically designed for real-time object detection. This module accurately identifies the positions and types of targets that are of critical importance to surveillance personnel. To accommodate the demanding frame rates of video streams while maintaining real-time detection performance, the YOLOv5 architecture with its efficient single-stage design is utilized.

B. Integrated ROI

The integrated ROI module is a novel component that aggregates the output from the ROI detection stage. This module is engineered to leverage the capabilities of the Hi3559 V100 hardware, which sets ROI-related parameters at the granularity of Group of Pictures (GOP) units. The ROI integration process entails collecting ROI data from all frames within a GOP and performing a union operation on these regions. This union effectively captures the motion trajectory of the target throughout the GOP, providing a comprehensive representation of the ROI's spatial extent. To ensure all critical components of the target are identified, object tracking algorithms can be employed to compute the motion path of each detected entity. The system incorporates the Simple Online and Realtime Tracking (SORT) algorithm [12], renowned for its simplicity and compatibility with real-time frame rates. SORT assigns an independent linear constant velocity model to each detected object, enabling robust tracking performance.

C. Adaptive Encoding

The adaptive encoding mechanism on the IPCam side utilizes an adaptive quantization parameter (QP) strategy. The integrated ROI module categorizes ROIs as the foreground, signifying regions containing targets of interest. These ROIs are assigned lower QP values to allocate more bits and preserve finer details, resulting in enhanced visual quality. Conversely, non-target regions, classified as the background,

receive higher QP values to reduce the bitrate without compromising the overall perception of the scene.

The QP values assigned to the foreground are typically 4 to 16 smaller than those assigned to the background. However, an excessive disparity in QP values can lead to perceptible artifacts at the boundary between the foreground and background regions. These artifacts can inadvertently draw the viewer's attention away from the intended target, diminishing the effectiveness of the surveillance system. Furthermore, the reduced bitrate in the background regions may introduce compression-related distortions. To mitigate these issues, the proposed framework incorporates a dynamic quality enhancement module. This module intelligently enhances the quality of the background regions, minimizing boundary artifacts and attenuating compression-induced distortions. By maintaining a balance between the foreground and background quality, the dynamic quality enhancement module ensures that the target remains the primary focus while providing a seamless viewing experience.

D. Dynamic Quality Enhancement

To effectively enhance the quality of both foreground and background regions, the NVR employs two separate super-resolution models based on the EDSR architecture [13]. The received block table from the client is utilized to distinguish between foreground and background areas. Each model is trained independently with unique network parameters to optimize performance for its respective region type. The SR2 model is specifically designed to further boost the quality of the already higher-fidelity foreground, ensuring critical details remain sharp and recognizable. Conversely, the SR1 model focuses on alleviating compression artifacts that are more prevalent in the background due to the reduced bitrate allocation. By tailoring the super-resolution process to the distinct characteristics of foreground and background, the dynamic quality enhancement module achieves a balanced and visually pleasing output.

E. Block Comparison and Caching

To expedite the super-resolution processing and improve overall system efficiency, a block comparison and caching mechanism is integrated alongside the dynamic quality enhancement module. The primary objective of this mechanism is to construct a block table that records essential information, including the positions of regions of interest (ROIs) and the similarity metrics between corresponding blocks across consecutive frames. By leveraging this data, the system can efficiently differentiate between foreground and background regions, eliminating the need for redundant processing. Furthermore, the block table enables the identification of blocks that can be reused from previous frames, effectively reducing the number of computationally expensive super-resolution operations required. This caching strategy significantly accelerates the enhancement process without compromising the final output quality.

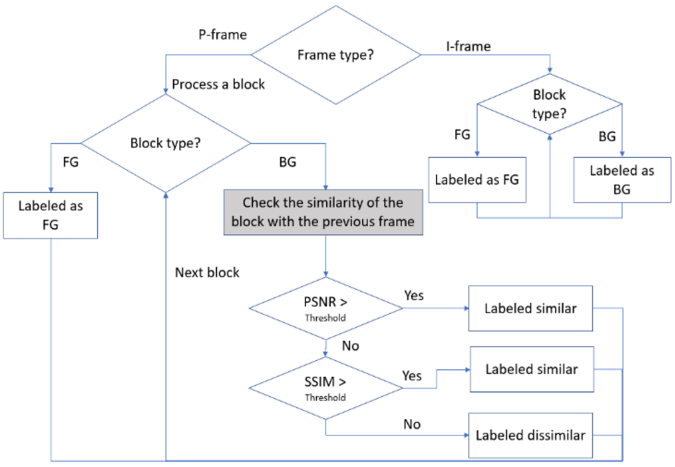

Fig. 2: Systematic process for constructing the block similarity table. This efficient two-stage approach, employing PSNR and SSIM metrics, enables the identification of static background regions for optimized caching and processing.

1) Caching mechanism: In practice, the computational complexity of the super-resolution (SR) poses a significant challenge to achieving real-time performance, particularly at high frame rates. Suboptimal output frame rates can lead to visual inconsistencies that negatively impact the user experience. However, a key insight is that in the majority of security monitoring scenarios, background elements such as walls, roadsides, and the sky exhibit high temporal redundancy, remaining largely static across multiple frames. Leveraging this characteristic, we introduce an innovative caching mechanism that reuses previously processed SR blocks, effectively reducing the number of SR operations required and consequently accelerating the image output rate.

The proposed caching mechanism is comprised of two core components: the construction of the block table and the determination of SR applicability for individual blocks. As a preliminary step, frames are partitioned into fixed-size blocks. Fig. 2 illustrates the block table construction process, which is executed prior to encoding on the IP camera side. The primary function of this table is to capture the similarity between corresponding blocks in consecutive frames. The process begins by assessing the frame type. For I-frames, only the block type (foreground or background) is recorded. In the case of P-frames, the block type is determined. Blocks within the ROIs are classified as foreground (FG), while those outside are considered background (BG). For background blocks, their similarity to the corresponding block in the previous frame is evaluated using a two-stage approach. First, the peak signal-to-noise ratio (PSNR) is compared against a predefined threshold. If the PSNR exceeds this threshold, the block is labeled as similar. Otherwise, the structural similarity index (SSIM) is computed. Blocks with an SSIM above a certain threshold are classified as similar, while those below are marked as dissimilar. By employing PSNR as an initial filter, followed by the more computationally intensive SSIM

5bits	*m**16 bits	10bits	*n**8 bits	*n**5 bits
ROI number (*m* ROIs)	ROI coordinates of *m* ROIs, 16bits each	Number of dissimilar block (*n* blocks)	Coordinates of dissimilar block 8bits each	Frame index of dissimilar block 5bits each

Fig. 3: Optimized data structure of the block similarity table. The fields are strategically designed to capture essential information while minimizing storage overhead.

for refined assessment, the system strikes a balance between efficiency and accuracy.

After decoding on the network video recorder (NVR) side, the dynamic quality enhancement module utilizes the information provided in the block table to determine the appropriate processing for each block. For background blocks classified as similar, the corresponding block from the previous frame is directly used, eliminating the need for redundant SR processing. Dissimilar background blocks undergo SR1 (BG) processing, which is optimized for background regions. Foreground blocks, on the other hand, are invariably subjected to SR2 (FG) processing, ensuring the highest quality enhancement for regions of interest. It is important to note that while I-frames undergo full-frame SR, the block similarity assessment is only applied to P-frames. Consequently, background blocks in P-frames can either be sourced from the previous frame or processed using SR1 (BG), depending on their similarity classification.

2) Design of block table: The transmission of the block table alongside video data to the NVR side increases network bandwidth and storage demands. To address this challenge, we propose an innovative recording method that effectively reduces the size of the block table while preserving critical information. The key data captured in the optimized block table includes the ROI coordinates and the positions of dissimilar blocks, enabling efficient processing and caching.

Our approach leverages the Group of Pictures (GOP) structure as the fundamental unit for block table organization. As illustrated in Fig. 3, the first field of the block table utilizes 5 bits to record the number of ROIs, accommodating a maximum of 31 ROIs per GOP. The second field captures the ROI coordinates using 8 bits per coordinate, allowing for the representation of up to 240 blocks. By recording the top-left and bottom-right corners of each ROI, a total of 16 bits are allocated to precisely delineate the ROI boundaries. The third field of the block table is dedicated to indicating the number of dissimilar blocks, which primarily correspond to background blocks due to their inherent high similarity across frames. To efficiently encode this information, 10 bits are employed, supporting a maximum of 1023 dissimilar blocks per GOP. The positions of these dissimilar blocks are recorded in the fourth field, utilizing 8 bits per block. Finally, the fifth field represents the frame index of the dissimilar blocks, capitalizing on the fact that each GOP can contain a maximum of 29 P-frames. By allocating 5 bits to store the frame index, our method ensures precise temporal localization of the dissimilar blocks.

Fig. 4: Representative surveillance video sequences utilized in the experimental simulations, encompassing a diverse range of indoor and outdoor scenarios to comprehensively evaluate the proposed system's performance across various real-world conditions.

III. Experiments

The experimental samples encompass a diverse array of surveillance scenes, capturing both indoor and outdoor environments, as illustrated in Fig. 4. The system leverages hardware encoding utilizing the HiSilicon chip Hi3559 V100 for both traditional and ROI-based encoding paradigms. In alignment with the methodologies employed in [7] and [9], where a subset of training and testing images exhibit correlations, a portion of the video frames from sequences A, B, C, D, and the DIV2K dataset [14] are utilized to train the EDSR model. Conversely, sequences E and F are deliberately excluded from the training process to assess the system's performance on entirely unseen data. The threshold values for determining block similarity are set to PSNR = 35 dB and SSIM = 0.85.

Traditionally, the Peak Signal-to-Noise Ratio (PSNR) metric has been widely employed for evaluating video quality, treating each pixel difference with equal weight. However, in the context of ROI coding, where adaptive quantization parameters (QP) are strategically applied, the resulting quality is inherently perceptually centered. To account for this discrepancy and provide a more representative assessment, this paper adopts the Weighted PSNR (WPSNR) metric, as proposed in [15]. WPSNR assigns higher importance to the contribution of ROI regions to the overall quality, while gradually decreasing the weights of non-ROI areas based on their pixel distance from the ROI's position, following a normal distribution. This approach enables a more nuanced and perceptually meaningful evaluation of the system's performance, aligning with the primary objectives of ROI-based video compression in surveillance applications.

A. Rate-Distortion Performance

To evaluate the performance of the proposed adaptive QP encoding scheme, both with and without the dynamic quality enhancement module, we conducted a comprehensive rate-distortion analysis. This analysis compared our approach to

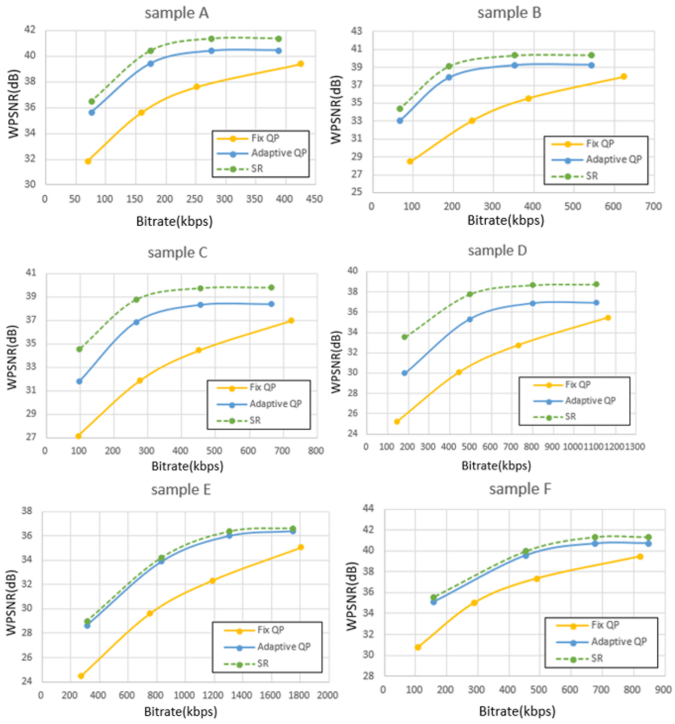

Fig. 5: Analysis of the rate-distortion performance.

the conventional HEVC encoding standard, providing valuable insights into the effectiveness of our proposed techniques.

Fig. 5 illustrates the results of this analysis, clearly demonstrating the superior performance of our adaptive QP encoding method over traditional HEVC across all tested sequences. The rate-distortion curves show a consistent improvement in quality at equivalent bitrates, highlighting the benefits of our ROI-centric bitrate allocation strategy. Furthermore, the incorporation of the super-resolution (SR) module in our dynamic quality enhancement pipeline yields even greater performance gains. As depicted by the green curves in Fig. 5, the combination of adaptive QP encoding and SR processing significantly elevates the rate-distortion performance, surpassing both conventional HEVC and standalone adaptive QP encoding.

B. Dynamic Quality Enhancement

The proposed dynamic quality enhancement module, which incorporates super-resolution (SR) processing, effectively enhances the visual fidelity of both foreground and background regions in the adaptive QP encoded video. To illustrate the impact of this module, we present a comparative analysis of the adaptive QP approach with and without the dynamic quality enhancement, utilizing difference maps in conjunction with heat maps for visual assessment.

Fig. 6 showcases the difference maps generated using sequence A. In Fig. 6(a), the adaptive QP encoding is applied without SR processing, while Fig. 6(b) depicts the frame after undergoing SR processing. Both processed frames are compared against the original frame by computing their pixel-

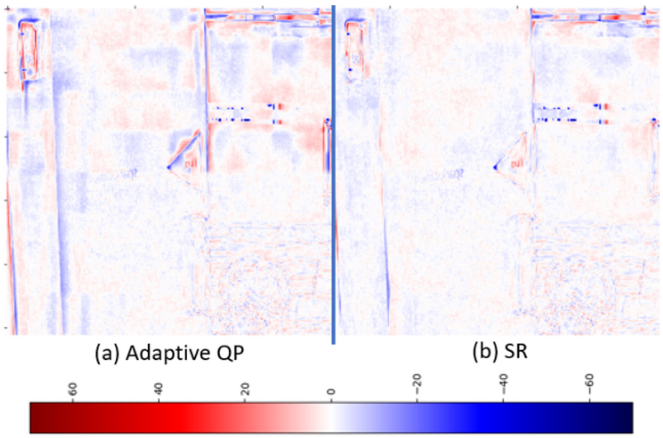

Fig. 6: Visual assessment of the impact of super-resolution processing in the dynamic quality enhancement module, utilizing difference maps and heat maps to highlight the improvement.

(a) Original image (blue region indicates the ROI)

(b) Using adaptive QP without SR

(c) Using adaptive QP with SR

Fig. 7: Subjective evaluation of the perceptual quality enhancement achieved through the application of the super-resolution technique.

wise differences. The results clearly demonstrate that the incorporation of SR processing significantly improves the quality at the boundaries between foreground and background regions. This enhancement leads to a smoother visual transition, mitigating the risk of visual distraction that may arise from stark quality disparities. Furthermore, Fig. 7 provides a subjective evaluation of the quality improvement attained through the application of super-resolution.

C. Cache Performance

To evaluate the impact of our caching mechanism on both execution time and output quality in the dynamic quality enhancement module, we conducted a comparative analysis between scenarios with caching enabled and disabled, focusing

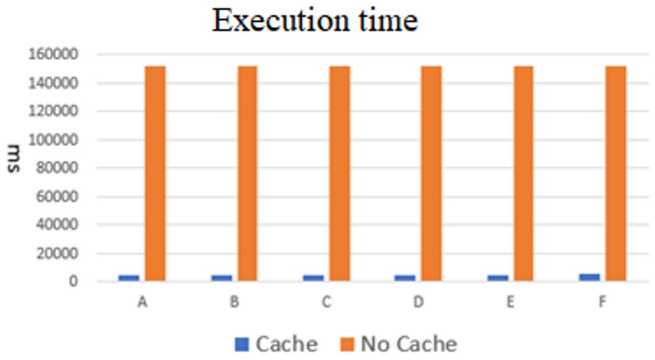

Fig. 8: Comparative analysis of the execution time through the integration of the proposed caching mechanism.

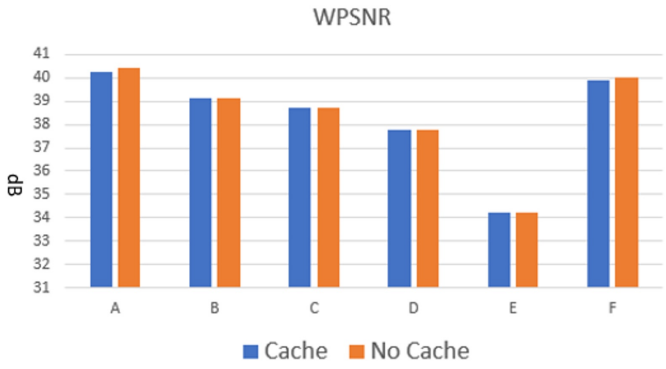

Fig. 9: The impact of the caching mechanism on output quality.

specifically on the performance of the SR1 process for background regions. Fig. 8 illustrates the stark contrast in execution times, revealing that the integration of our caching strategy yields a significant 96.65% reduction in processing time. These measurements were obtained using the Tesla V100-SXM2 GPU platform, underscoring the real-world applicability and efficiency of our approach.

It is important to acknowledge that the caching mechanism's retrieval of blocks from preceding frames may result in a slight degradation of output quality. As depicted in Fig. 9, the maximum observed decrease in WPSNR (Weighted Peak Signal-to-Noise Ratio) after applying caching is a mere 0.169 dB. This marginal compromise in quality is more than compensated for by the substantial gains in processing efficiency. By sacrificing an almost imperceptible amount of visual fidelity, our caching system dramatically reduces the computational burden of the SR1 process for background regions, ultimately enabling a significant improvement in the overall output frame rate. This strategic trade-off positions our dynamic quality enhancement module as a compelling solution for real-time video streaming applications, where the balance between quality and efficiency is of utmost importance.

IV. Conclusion

We have developed an adaptive video streaming system that optimizes bitrate allocation by prioritizing Regions of Interest (ROIs). By strategically assigning higher bitrates to critical regions while conserving resources in less pivotal areas, our system achieves a remarkable synergy between enhanced target clarity and reduced bandwidth and storage requirements. To further elevate visual quality, we integrate super-resolution techniques that refine both foreground and background elements. Moreover, our system capitalizes on the inherent redundancy in security footage backgrounds by incorporating an innovative caching mechanism, which significantly bolsters processing efficiency. Extensive experimental results validate the efficacy of our dynamic ROI video compression framework, demonstrating substantial improvements in overall video quality.

References

[1] S. W. Market. (2022) Video surveillance market set to register $76 billion by 2027. Accessed: 2024-12-09. [Online]. Available: https://www.securityworldmarket.com/int/News/Business-News/video-surveillance-market-set-to-register-76-billion-by-20271

[2] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE transactions on image processing*, vol. 13, no. 10, pp. 1304–1318, 2004.

[3] S. Zhu, C. Liu, and Z. Xu, "High-definition video compression system based on perception guidance of salient information of a convolutional neural network and hevc compression domain," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 1946–1959, 2019.

[4] J. Xiao, Z. Wang, Y. Chen, L. Liao, J. Xiao, G. Zhan, and R. Hu, "A sensitive object-oriented approach to big surveillance data compression for social security applications in smart cities," *Software: Practice and Experience*, vol. 47, no. 8, pp. 1061–1080, 2017.

[5] H. Choi and I. V. Bajic, "High efficiency compression for object detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1792–1796.

[6] L. Galteri, M. Bertini, L. Seidenari, and A. Del Bimbo, "Video compression for object detection algorithms," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3007–3012.

[7] H. Yeo, Y. Jung, J. Kim, J. Shin, and D. Han, "Neural adaptive content-aware internet video delivery," in *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, 2018, pp. 645–661.

[8] J. Kim, Y. Jung, H. Yeo, J. Ye, and D. Han, "Neural-enhanced live streaming: Improving live video ingest via online learning," in *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*, 2020, pp. 107–125.

[9] P. Hu, R. Misra, and S. Katti, "Dejavu: Enhancing videoconferencing with prior knowledge," in *Proceedings of the 20th International Workshop on Mobile Computing Systems and Applications*, 2019, pp. 63–68.

[10] H. Yeo, C. J. Chong, Y. Jung, J. Ye, and D. Han, "Nemo: enabling neural-enhanced video streaming on commodity mobile devices," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 1–14.

[11] J. Erfurt, C. R. Helmrich, S. Bosse, H. Schwarz, D. Marpe, and T. Wiegand, "A study of the perceptually weighted peak signal-to-noise ratio (wpsnr) for image compression," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2339–2343.

[12] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3464–3468.

[13] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, 2017, pp. 136–144.

[14] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[15] Z. Wang and A. C. Bovik, "Embedded foveation image coding," *IEEE Transactions on image processing*, vol. 10, no. 10, pp. 1397–1410, 2001.