ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

YUV-based Deep Learning Super-Resolution for Bitrate Reduction and ROI Preservation in Modern Video Codecs

Lertluck Leela-amornsin[1], Nuttapon Vanakittistien[2], Nattee Niparnan[3], Pitchaya sitthi-amorn[4] and Attawith Sudsang[5]

ABSTRACT

High Efficiency Video Coding (HEVC) and its successors, such as Versatile Video Coding (VVC), offer substantial bitrate reductions, yet challenges remain in preserving visual fidelity under bandwidth and computational constraints. This paper proposes a deep learning-based super-resolution (SR) framework that operates natively in the YUV color space, eliminating costly RGB-YUV conversions and integrating seamlessly with modern video compression pipelines. We develop two convolutional network architectures trained on YUV-formatted video data: a full 3-channel model and a lightweight two-stream variant that separately processes luminance (Y) and chrominance (UV) channels using compact subnetworks. The proposed method enhances both full-frame and region-of-interest (ROI) quality, outperforming conventional HEVC baselines in terms of rate-distortion efficiency. Evaluations on diverse video sequences demonstrate significant bitrate savings and effective ROI preservation, with the lightweight model offering a practical solution for AI-driven applications in resource-constrained environments.

Article information:

Keywords: HEVC Encoding, VVC, Bitrate Reduction, Deep Learning, Super-resolution, YUV Color Space, ROI Preservation, Task-driven Video Coding, Light-weight, AI On-edge Device

Article history:
Received: September 25, 2025
Revised: January 29, 2026
Accepted: February 26, 2026
Published: March 7, 2026 (Online)

DOI: 10.37936/ecti-cit.2026202.263972

1. INTRODUCTION

Modern video compression systems eg. HEVC, and VVC encoding increasingly operate under strict bandwidth and computational constraints, particularly in edge-oriented and task-driven applications. In such systems, limited network bandwidth is often shared among multiple mobile and swarm agents—such as surveillance robots, aerial drones, and intelligent sensing platforms—making it infeasible to transmit all visual content at uniformly high quality.

Under these conditions, visual information does not contribute equally to downstream tasks: regions of interest (ROIs) typically contain semantically critical content, whereas non-ROI regions mainly provide contextual cues. Motivated by this observation, this work investigates a task-driven compression framework that reduces bitrate by selectively degrading non-ROI regions through spatial subsampling, while preserving and recovering visual fidelity within ROIs via post-decoding reconstruction.

Despite extensive research on task-aware video compression and deep learning-based enhancement, two important gaps remain. First, most existing super-resolution and restoration methods operate in the RGB domain, whereas modern video codecs natively encode content in the YUV color space. This mismatch introduces additional color space conversions, increases system complexity, and limits direct integration with practical video pipelines. Second, many task-driven or learning-based compression approaches require modifications to encoder internals or codec-specific bit allocation strategies, which hinders compatibility with standardized and widely deployed codecs such as HEVC and VVC.

In contrast, modern video codecs continue to evolve toward higher compression efficiency and structural complexity, making encoder-side modification increasingly impractical in real-world systems. This raises a fundamental question: can task-driven bitrate reduction and quality recovery be achieved in a codec-agnostic manner, without altering the underlying video encoder?

[1,3,4,5]The authors are with the Department of Computer Engineering, Chulalongkorn University, Bangkok, Thailand, Email: l.lertluck@gmail.com, nattee@gmail.com, pitchaya@cp.eng.chula.ac.th and attawith@cp.eng.chula.ac.th
[2]The author is with the Monk Studios, Bangkok, Thailand, Email: nuttapon.vana@gmail.com
[1]Corresponding author: l.lertluck@gmail.com

Based on this scope, the research questions addressed in this paper are: (1) how much bitrate can be reduced by selectively degrading non-ROI regions under fixed-QP or fixed-bitrate constraints, and (2) whether YUV-based super-resolution can effectively restore visual quality—particularly within ROIs—while maintaining contextual consistency in non-ROI areas.

2. OUR CONTRIBUTIONS

This paper presents a task-driven, codec-compatible framework for bitrate reduction and ROI-preserving enhancement in modern video codecs. The proposed approach integrates YUV-based deep learning with lightweight preprocessing and is designed to operate under practical encoding constraints, including fixed-QP and fixed-bitrate settings, while remaining compatible with standardized codecs such as HEVC and, conceptually, its successor VVC.

The main contributions of this work are summarized as follows:

1) A codec-compatible, task-driven framework that reduces bitrate by selectively degrading non-ROI regions while preserving visual fidelity in ROIs under fixed-QP and fixed-bitrate constraints.

2) A YUV-native super-resolution approach that operates directly on HEVC-reconstructed frames, avoiding RGB-domain processing and facilitating seamless integration with modern and next-generation video codecs.

3) A comparative study of standard and lightweight SR architectures, including a three-channel model and a Y/UV two-stream variant, demonstrating effective ROI quality recovery and contextual preservation under bandwidth-limited and edge-oriented scenarios.

It is important to note that ROI definition is inherently task-dependent. In this study, we adopt a lightweight semantic segmentation model (PIDNet) [1] as a representative example to generate ROI masks in real time. However, ROI detection itself is not the focus of this work.

Instead, our contribution lies in the design of a task-agnostic, end-to-end framework in which task-driven preprocessing enables bitrate reduction in a codec-compatible manner, while post-decoding super-resolution restores visual quality degraded by both compression and preprocessing. Together, these components jointly define a practical bitrate–quality trade-off suitable for deployment in bandwidth-constrained and edge-oriented video systems.

3. RELATED WORK

3.1 Modern Video Encoding and Rate Control Strategies

High Efficiency Video Coding (HEVC/H.265) and its successor Versatile Video Coding (VVC/H.266) significantly improve compression efficiency through techniques such as quadtree partitioning, advanced intra/inter prediction, and motion compensation. These standards use a hybrid coding structure with I-, P-, and B-frames to reduce spatial and temporal redundancy, where I-frames are key to high-fidelity reconstruction. Both codecs adopt the YUV 4:2:0 format by default, allocating most bitrate to the luminance (Y) channel to exploit perceptual sensitivity—making Y-channel enhancement particularly impactful for super-resolution.

To regulate bitrate, HEVC and VVC implement rate control by dynamically adjusting the quantization parameter (QP) at GOP, frame, and CTU levels. The widely adopted R-λ model [2] balances rate-distortion trade-offs efficiently, while VVC further introduces skipped vs. non-skipped CTU handling to optimize bit allocation [3]. Recent approaches extend beyond conventional control by integrating deep learning techniques. Reinforcement learning (RL) has shown promise for adaptive QP selection [4], [5], while task-driven RL methods [6], [7] prioritize semantic fidelity for downstream AI tasks.

In practical video coding systems, bitrate control differs fundamentally from objectives that prioritize spatial accuracy or temporal consistency under unconstrained bandwidth. When bandwidth is sufficient, encoder configurations can allocate more bits to preserve spatial detail and temporal smoothness through flexible GOP structures and reference frame usage.

Under strict bandwidth constraints—such as mobile or edge-based surveillance—bitrate control becomes the dominant objective. Encoder modes including fixed-QP or fixed-bitrate operation, low-delay or random-access configurations, and constrained reference structures inherently limit achievable spatial and temporal quality, often requiring trade-offs to ensure stable transmission.

This work targets such bitrate-constrained operating points. Instead of modifying internal codec rate control, we apply preprocessing strategies—such as spatial subsampling and selective non-ROI degradation—to adjust the bitrate–quality trade-off at the encoder input. A post-decoding super-resolution stage is then employed to recover spatial detail, particularly within ROIs. Consequently, spatial accuracy and temporal consistency are optimized within bitrate constraints, rather than treated as independent objectives.

3.2 RGB and YUV Color Spaces

Most deep learning models for image and video processing are trained in the RGB color space due to the availability of large-scale RGB datasets and the direct use of RGB in display devices. However, RGB processing is not natively aligned with modern video coding pipelines, which rely on YUV representations.

Conversion between RGB and YUV introduces additional computational overhead and may result in color distortion, making it less suitable for practical codec-integrated systems.

In contrast, modern video codecs such as HEVC encode content in the YUV color space—typically YUV 4:2:0—where luminance and chrominance components are separated and compressed according to human visual sensitivity. Despite its prevalence in video compression, relatively few deep learning approaches have been explicitly designed to operate directly on YUV data within a video coding context.

Recent studies have begun exploring YUV-based learning. For instance, YUVMultiNet [8] employs YUV inputs for multi-task perception networks in autonomous driving, focusing primarily on detection-related applications without addressing video reconstruction. YUVGAN [9] applies generative adversarial learning in the YUV domain for remote sensing image restoration, while LYTNet [10] investigates lightweight YUV-based transformer models for low-light image enhancement. However, these approaches are mainly limited to image-level processing or perception-oriented tasks and do not consider integration with modern video coding pipelines, bitrate-constrained encoding scenarios, or post-decoding reconstruction within standardized video codecs.

In contrast to these approaches, this study provides a deeper exploration of YUV-aware learning within a codec-compatible video compression framework. Specifically, we train super-resolution models directly on HEVC-reconstructed YUV frames and integrate them into a task-driven pipeline designed for bitrate reduction under fixed-QP and fixed-bitrate constraints. Furthermore, the proposed framework explicitly considers ROI-aware quality recovery and contextual preservation, enabling selective reconstruction without modifying codec internals. This distinguishes our work from prior YUV-based methods that are not designed for end-to-end video compression or practical deployment in bandwidth-constrained systems.

3.3 Super-Resolution for Compression Enhancement

3.3.1 Image Super-Resolution (Image SR)

Image Super-Resolution (SR) refers to the task of reconstructing high-resolution (HR) images from their low-resolution (LR) counterparts. Traditional interpolation-based methods such as nearest-neighbor, linear, bilinear, and bicubic upsampling are simple and computationally efficient, but tend to produce blurry outputs with limited detail preservation. With the advent of deep learning, SR techniques have evolved significantly and can be broadly categorized as follows [11]: **Pre-upsampling SR**: Methods such as SRCNN [12] first upscale LR images using bicubic interpolation, followed by a convolutional neural network to refine the coarse HR image. This strategy reduces the learning burden by processing images that are already close to the target resolution. **Post-upsampling SR**: Approaches such as SRGAN [13], ESRGAN [14] adopt Pix2Pix [15] to feed the original LR image into deep convolutional networks and apply learnable upsampling layers at the final stage. This end-to-end structure is capable of capturing richer textures and high-frequency details. Progressive and iterative up-sampling SR methods incrementally refine resolution using coarse-to-fine strategies or back-projection mechanisms, helping stabilize training and improve fine detail reconstruction.

These categories demonstrate the transition from conventional SR to highly adaptive learning-based frameworks that significantly boost visual quality and restoration performance.

3.3.2 Video Super-Resolution (VSR)

Video Super-Resolution (VSR) techniques aim to reconstruct high-resolution (HR) video frames from low-resolution (LR) sequences and can be broadly categorized into two types: **VSR with alignment and VSR without alignment**. The former applies explicit motion estimation and compensation, typically using optical flow to temporally align LR frames. An early example is VSRnet [16], which emphasized temporal consistency by aligning frames prior to CNN-based reconstruction.

In contrast, alignment-free approaches rely on end-to-end deep networks to implicitly perform alignment, feature fusion, and reconstruction. These include 2D convolution-based models such as VSRResFeatGAN [17], 3D convolutional methods like DUF [18] and 3DSRNet [19]. Beyond that, recurrent architectures and non-local attention-based models like Progressive Fusion Network [20] are introduced but many of these methods rely on deep and complex networks, making them less suitable for lightweight models.

Building on these foundations, Wang et al. [21] extended the Pix2Pix model to a video-to-video synthesis framework that produces temporally coherent HR outputs from semantic inputs. A comprehensive survey by Liu et al. [22] summarizes recent VSR advancements, highlighting their roles in both visual enhancement and end-to-end video coding.

While prior work primarily focuses on restoring spatial detail and temporal consistency for visual quality, our work diverges by targeting bitrate reduction under standard HEVC/VCC compression. Moreover, most VSR studies operate in the RGB domain and rarely address integration with modern video codec-compatible YUV pipelines, especially with a focus on region-of-interest (ROI) preservation.

4. PERFORMANCE EVALUATION

An overview of the complete processing pipeline corresponding to this section is illustrated in Fig. 4

4.1 Compression Pipeline

We evaluated our pipeline under a range of scenarios by combining three major experimental factors:

Input: Raw YUV video + ROI masks

Preprocessing Scenarios:

1) Original HEVC Baseline: High-resolution YUV frames are directly compressed and reconstructed using standard HEVC without super-resolution enhancement.

2) Sharp Bicubic Subsampling: Downsampling by 4× using bicubic interpolation, followed by HEVC encoding, decoding, and 4× SR reconstruction.

3) Blur Bicubic Subsampling: Input frames are first blurred before downsampling, simulating degraded low-quality inputs.

4) Non-ROI Blur Subsampling: Only non-ROI regions are blurred while ROI areas remain sharp before downsampling and compression.

Encoding Settings:

1) Fixed QP: HEVC encoding with constant quantization parameter (QP).

2) Fixed Bitrate: HEVC encoding under a constrained bitrate target.

Reconstruction: Compressed bitstream is decoded.

SR Model Decoding Enhancement:

1) 3-channels YUV SR Model: A 3-channel network trained on full YUV (Y+U+V) data.

2) Lightweight Y/uv Model: A two-stream architecture with a deeper Y-channel network and a compact UV stream to reduce complexity for edge deployment.

Final YUV video: A standard 3-channel network that jointly reconstructs luminance and chrominance components.

This structured comparison allows us to investigate the impact of different preprocessing techniques on bitrate savings and reconstruction quality, particularly in maintaining fidelity for task-driven compression applications.

4.2 YUV Color Space and Data Preparation

Although RGB is common in deep learning, YUV is the standard in video compression, with YUV 4:2:0 offering efficient bitrate allocation by prioritizing the luminance (Y) channel. Recent work like YUVMultiNet [8] shows that processing directly in YUV reduces computational overhead and memory usage, benefiting edge applications such as autonomous driving. Similarly, [23] reports better PSNR and SSIM in YUV-based colorization tasks.

Motivated by these findings, we train super-resolution models directly on YUV 4:2:0 videos. The Y, U, and V planes are extracted from each frame, with U and V upscaled via bicubic interpolation to match Y resolution, forming a normalized (3, H, W) input tensor for CNNs.

We develop two SR models to balance fidelity and efficiency:

3-channel YUV model: Processes the full YUV tensor, preserving luminance and chrominance details for high-quality reconstruction.

Lightweight Y/uv model: Uses a two-stream design—one deep branch for Y and a compact one for UV—optimized for edge deployment with reduced complexity and memory usage.

This dual approach enables both high-fidelity enhancement and efficient inference under resource constraints.

4.3 ROI Extraction via Semantic Segmentation

Identifying semantically important regions is crucial for an effective ROI-based compression and reconstruction. Various methods have been proposed for region detection, including latent feature extraction from pre-trained backbones and attention-based ROI selection. However, in our work, we employed PIDNet [1], which is recognized for its balance between accuracy and real-time performance.

PIDNet achieved top ranking in the Cityscapes Panoptic Segmentation Challenge, demonstrating its ability to perform precise and efficient segmentation tasks. PIDNet is particularly well suited for edge-device applications, a priority in our study, owing to its efficient shared backbone and lightweight decoder heads.

We applied PIDNet to the original REDS dataset [24] frames to generate ROI masks. The masks are stored in the JSON format following a schema similar to Cityscapes, ensuring compatibility and ease of integration into our preprocessing pipeline.

The target objects for the ROI extraction were selected based on their relevance to surveillance and autonomous driving applications. Specifically, we used the following person, rider, car, truck, bus, caravan, trailer, train, motorcycle, and bicycle classes as the target ROIs.

These categories cover the dynamic and critical objects commonly encountered in real-world driving and surveillance scenarios. The generated binary masks guide both the selective blurring process during preprocessing and evaluation of ROI-specific metrics during model assessment.

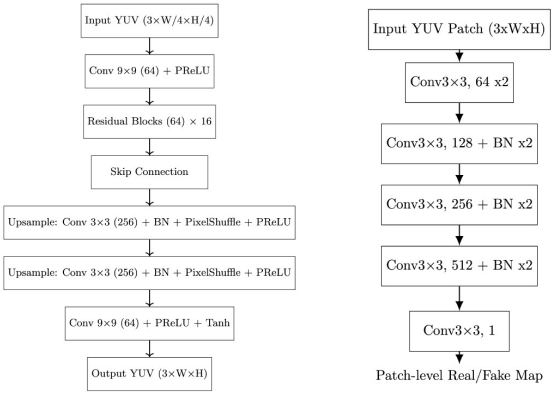
To simulate selective degradation, the non-ROI regions are blurred using a Gaussian filter. Specifically, we applied Gaussian blur with a kernel size of (15, 15) and a standard deviation of 0 to the non-ROI areas, ensuring that the ROI regions remain sharp while the surrounding context was intentionally degraded. This technique effectively emphasizes the importance

of preserving ROI fidelity during the subsequent compression and reconstruction processes.

4.4 Our Neural Network Model for YUV Super-Resolution

We adapted a Pix2Pix-style architecture [15] with a U-Net-inspired generator and a PatchGAN discriminator for YUV super-resolution. Following this framework, our generator network maps the HEVC-reconstructed low-resolution YUV frames to high-resolution outputs, whereas the discriminator enforces perceptual realism by distinguishing between real and generated HR frames.

Furthermore, unlike traditional RGB-based super-resolution methods, we operate directly on the YUV color space, avoiding unnecessary color space conversion and focusing on luminance-chrominance properties, which are critical for video compression standards such as HEVC.



(a) 3-Channel YUV Generator **(b)** PatchGAN Discriminator

Fig.1: *Network architecture diagrams: (a) Generator for 3-channel YUV super-resolution, (b) PatchGAN discriminator used for adversarial training.*

4.4.1 Full 3-Channel YUV Generator Network

The **generator network** in the full 3-channel YUV model is based on the GeneratorResNet architecture, Fig. 1(a), designed to process all three YUV channels together. This design prioritizes reconstruction quality by allowing cross-channel feature learning, which is particularly beneficial for preserving fine luminance detail and chroma consistency in high-quality reconstruction scenarios. Residual learning and long-range skip connections are employed to stabilize training and facilitate the recovery of high-frequency details lost during spatial subsampling and compression. This architecture serves as the quality-oriented baseline in our experiments.

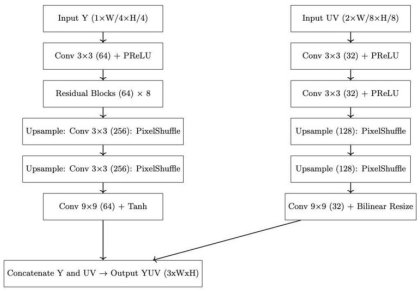


Fig.2: *Architecture of the lightweight Y/uv two-stream generator. The Y and UV branches are processed independently and concatenated after upsampling to form the final output YUV frame.*

4.4.2 Lightweight Y/uv Two-Stream Generator Network

To support deployment on resource-constrained devices, we designed a lightweight generator, Light Y/uv, with separate processing streams for luma (Y) and chroma (UV) components of standard input YUV 4:2:0 format. We introduce a lightweight two-stream generator, Fig. 2, that decouples luminance and chrominance processing. The Y channel, which carries most perceptual detail, is handled by a dedicated branch with higher representational capacity, while the subsampled U and V channels are processed jointly using a compact branch. The two streams are upsampled independently and merged to form the final YUV output. This design significantly reduces model complexity and computational cost while maintaining competitive reconstruction performance.

4.4.3 PatchGAN Discriminator Network

Both the full and lightweight models use the same PatchGAN Discriminator architecture, Fig. 1(b), which classifies the patches of the image as real or fake to provide granular feedback to the generator. Both generator variants share the same PatchGAN discriminator, which operates on local image patches rather than full frames. This choice encourages the generation of locally realistic textures and mitigates over-smoothing artifacts commonly observed in pixel-wise reconstruction losses.

5. EXPERIMENTAL SETUP

5.1 Datasets

We conducted experiments on the REDS dataset. The REDS dataset, introduced in the NTIRE 2019 Challenge on Video Deblurring and Super-Resolution [24], provides high-resolution RGB video sequences captured under real-world motion blur and degradation, making it well suited for super-resolution and video compression research.

All video sequences were converted into YUV 4:2:0 format for consistency with the HEVC compression pipeline. To enable region-of-interest (ROI)–aware

Fig.3: *Examples of preprocessing steps used in this study. Each column corresponds to a different video sequence with a randomly sampled frame. The first row shows the original subsampled frames, the second row shows the binary ROI masks generated by PIDNet, and the third row shows the preprocessed frames after applying non-ROI blurring.*

processing, we applied a pre-trained PIDNet model [1] to generate binary ROI masks for each frame. These masks primarily capture semantically important regions such as pedestrians and vehicles, which are common targets in surveillance and autonomous driving scenarios.

Figure 3 illustrates representative snapshot examples used in this study. The first row shows the original subsampled video frames after format conversion. The second row presents the corresponding binary ROI masks generated by PIDNet, where highlighted regions indicate ROIs. The third row shows the resulting preprocessed frames after applying non-ROI blurring, in which background regions outside the ROI are deliberately degraded to reduce bitrate while preserving ROI fidelity.

We split the REDS dataset into 210 sequences for training, 30 sequences for validation, and 30 held-out sequences for testing, ensuring that all splits were performed at the sequence level to avoid content leakage across sets.

5.2 HEVC Encoding Settings

Two primary encoding configurations are evaluated in this study:

- **Fixed-QP Mode:** Encoding is performed using a constant quantization parameter (QP), with QP set to 22 to balance compression efficiency and visual quality.
- **Fixed-Bitrate Mode:** Encoding targets a predefined bitrate (e.g., 1000 kbps) using rate control.

All encoding and decoding processes are carried out using FFmpeg (version 4.2.7) with the libx265 encoder, which is a widely used open-source implementation of the HEVC standard.

The proposed pipeline operates exclusively on decoded YUV frames and does not modify codec internals, rate control mechanisms, or bitstream syntax. As a result, the framework is codec-compatible and conceptually applicable to block-based hybrid video coding standards such as HEVC and VVC. However, experimental validation in this work is limited to HEVC, and no empirical performance claims are made for VVC.

5.3 Training Details

Figure 4 illustrates our training pipeline, which begins with YUV input and applies one of three preprocessing scenarios: (1) sharp bicubic subsampling, (2) blur bicubic subsampling, and (3) non-ROI blur subsampling—each downsampling by a factor of 4. The resulting low-resolution sequences are compressed with HEVC, decoded, and then reconstructed using either standard bicubic upsampling or one of our proposed super-resolution (SR) models. We evaluate two SR architectures:

3-channel YUV model (3CH): A full-capacity generator that jointly reconstructs Y, U, and V components for high-fidelity output.

Lightweight Y/uv model (LW): A two-stream design processing luminance (Y) and chrominance (UV) separately to reduce complexity while preserving ROI fidelity.

Each model was trained independently on sharp, blur, and non-ROI blurred versions of the HEVC-reconstructed YUV datasets. This setup enables the networks to learn artifact restoration under real-world compression. The resulting models are optimized for both quality and deployment on bandwidth-limited or edge environments.

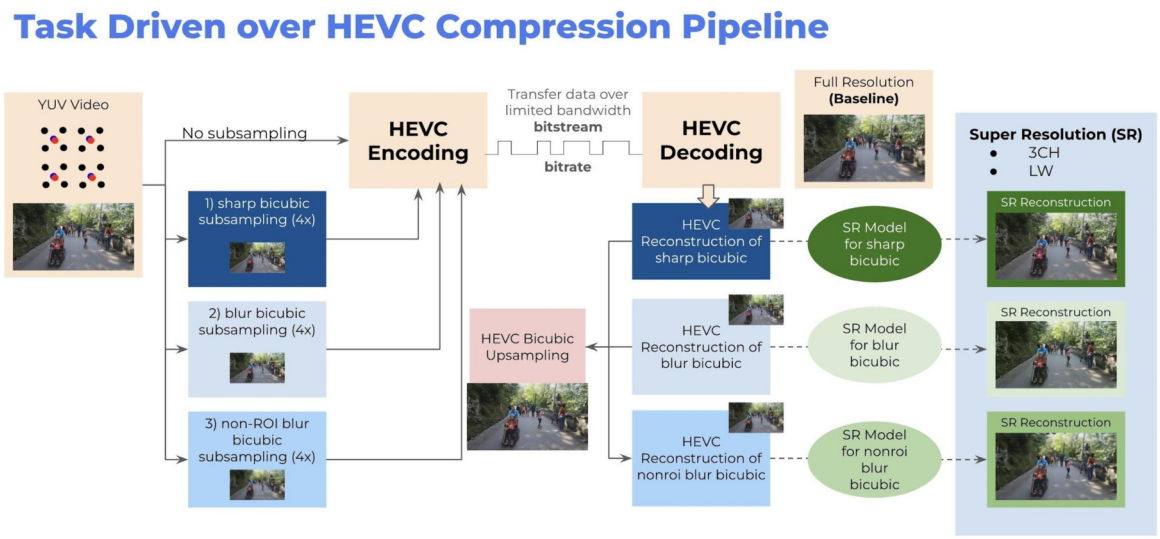


Fig.4: *Our proposed framework supporting various compression and reconstruction scenarios.* It should be noted that the framework illustrated in Fig. 4 represents multiple scenario-specific processing paths rather than a combined or fused super-resolution strategy. Each experimental setting activates only one preprocessing configuration and its corresponding SR model (e.g., sharp bicubic, blur bicubic, or non-ROI blur bicubic). The different SR models shown in this figure are therefore applied independently under different scenarios, and no model fusion or joint reconstruction is performed.

5.3.1 Training Configuration & Implementation Details

The training was performed on an Nvidia Geforce RTX 4090 GPU. Each model was trained for 2000 epochs with a batch size of 2 and an initial learning rate of 2×10^4, decayed by a factor of 0.5 every 100 epochs. The Adam optimizer is used with $\beta_1 = 0.5$ and $\beta_2 = 0.999$.

All models were implemented in PyTorch. The YUV frame extraction and color space processing were conducted using OpenCV and custom Python scripts. Gaussian blurring for non-ROI areas was performed with a kernel size of (15, 15).

• Mean Squared Error (MSE) loss is used for the adversarial loss.

• The L1 loss (mean absolute error) is used as the content loss between the generated and ground-truth high-resolution YUV frames.

• Instead of using VGG-based perceptual loss in RGB space, we directly calculate L1 loss on YUV channels, which is critical to preserve structure and color consistency in video compression tasks, as suggested in [9]. The L1 loss formulation is as follows:

$$L_1(G) = \|Y - \hat{Y}\|_1 + \|U + \hat{U}\|_1 + \|V - \hat{V}\|_1 \quad (1)$$

5.3.2 Training Scenarios

We systematically trained with the following video sequences:

• Sharp-only bicubic subsampling.

• Blur-only bicubic subsampling.

• ROI-preserving non-ROI blur subsampling.

For each scenario, we trained from HEVC reconstructed output to learn the artifact and impact of HEVC encoding.

5.4 Evaluation Metrics

Our evaluation protocol includes multiple dimensions to thoroughly assess the performance of the proposed super-resolution models and the compression pipeline.

- **Bitrate:** We measured the bitrate of the HEVC encoded output in kilobits per second (kbps) to evaluate the effectiveness of bitrate reduction techniques across different scenarios. The bitrate was extracted from the HEVC bitstream metadata.
- **PSNR and MS-SSIM:** Peak Signal-to-Noise Ratio (PSNR) and Multi-Scale Structural Similarity Index Measure (MS-SSIM) were used to quantify the reconstruction quality. The PSNR provides a pixel-level assessment of fidelity, whereas MS-SSIM captures structural and perceptual similarities.
- **Model Efficiency Metrics:** To assess computational feasibility, especially for edge deployment,we compared the following performance indicators between the full 3-channel YUV model and the lightweight Y/uv model:
  - **Number of Parameters:** Total trainable parameters in each model.
  - **FLOPs:** Estimated floating-point operations required per frame.

– **Processing Time:** Average elapsed time per frame (in milliseconds) measured during inference on a standardized test set using the same hardware configuration.

These metrics help quantify the trade-off between reconstruction quality and computational cost, particularly in real-time or resource-constrained environments.

This comprehensive evaluation setup ensures a detailed understanding of both compression efficiency and quality preservation, particularly emphasizing the impact on regions that are semantically significant.

6. RESULTS AND DISCUSSION

We evaluated the quality and compression trade-offs of the proposed models using the PSNR, PSNR ROI, MS-SSIM and bitrate metrics under various conditions.

6.1 Performance under fixed QP condition

Under a fixed QP setting of 22, comparisons across distinct video sequences revealed the following trends:

- **Baseline HEVC encoding** (full resolution without preprocessing or super-resolution) consistently achieved the highest PSNR and MS-SSIM across all video sequences. However, this comes at the expense of substantially higher bitrates, often exceeding 10–20× the bitrate of the subsampled pipelines (see figure 5 (c)).
- **Super-resolution (SR) models** applied after HEVC reconstruction substantially improve quality over their non-SR counterparts. Both the 3-channel YUV model and lightweight Y/uv model boost the PSNR and MS-SSIM, depending on the sequence content as shown in table 1.
- **The 3-channel model** slightly outperformed the lightweight variant in terms of the average PSNR and MS-SSIM. Nevertheless, the lightweight Y/uv model remains highly competitive, while offering clear advantages in terms of speed and computational efficiency, making it well-suited for edge deployment.

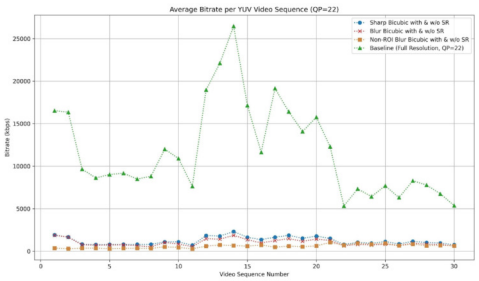


Fig.5: *Bitrate Comparison with Full Resolution at QP=22.*

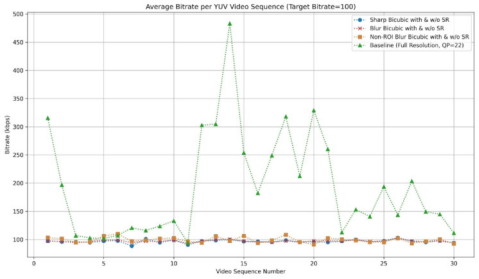


Fig.6: *Bitrate Comparison with Full Resolution at Bitrate=100Kbps.*

6.2 Performance under fixed bitrate condition

When encoding with a fixed target bitrate of 100 kbps, the analysis across sequences reveals the following key findings:

- The **baseline configuration**, HEVC applied directly to full-resolution videos, fails to meet the target bitrate in most sequences. Bitrates frequently overshoot the constraint by several hundred percent (see figure 6), rendering this configuration impractical under strict bandwidth limits.
- In contrast, the **subsampled pipelines** (4× down-scale) consistently meet the 100 kbps target. These include sharp bicubic, blurred bicubic, and non-ROI blur variants, all processed at a reduced resolution before HEVC encoding.
- **Super-resolution models** (both 3-channel and lightweight Y/uv) significantly improved the quality of the HEVC-reconstructed sequences. As shown in table 2, SR-enhanced outputs outperform their non-SR counterparts, delivering higher PSNR and MS-SSIM while maintaining compliance with the bitrate constraint.
- The **3-channel YUV model** yields the best overall reconstruction quality. However, the **lightweight Y/uv model** achieves competitive performance, particularly in sequences with moderate complexity, while requiring fewer parameters and a lower computational cost.

Under the extremely constrained fixed-bitrate condition (100 kbps), the 3CH SR model shows a slight negative cumulative ΔPSNR in the sharp bicubic scenario. At very low bitrates, the encoder's bit allocation may not sufficiently preserve high-frequency information required for effective SR reconstruction. As a result, even though the 3-channel SR model attempts to recover details, the severely quantized input limits its pixel-level reconstruction performance.

Nevertheless, the corresponding MS-SSIM increases, suggesting that structural information is still better preserved. The small magnitude of the negative ΔPSNR (-1.536) further indicates that the degradation is minor and primarily affects pixel-wise fidelity rather than perceptual structure.

Table 1: *Average reconstruction quality under fixed-QP setting. Mean PSNR and MS-SSIM are averaged over all sequences. ΣΔPSNR is computed relative to the corresponding w/o SR baseline within each scenario.*

	Mean PSNR (dB)			ΣΔPSNR (vs w/o SR)		Mean MS-SSIM		
Scenario	w/o SR	LW	3CH	LW	3CH	w/o SR	LW	3CH
Sharp Bicubic	81.986	83.105	82.689	33.581	21.076	0.976	0.982	0.982
Blur Bicubic	79.730	80.520	81.165	23.676	43.022	0.931	0.950	0.961
Non-ROI Blur	78.054	78.263	79.601	6.290	46.415	0.894	0.918	0.926

	Mean ROI PSNR (dB)			ΣΔROI PSNR (vs w/o SR)		Mean ROI MS-SSIM		
Scenario	w/o SR	LW	3CH	LW	3CH	w/o SR	LW	3CH
Sharp Bicubic	86.018	86.737	88.242	20.851	64.493	0.956	0.974	0.978
Blur Bicubic	84.157	83.278	86.223	-25.477	59.927	0.819	0.879	0.909
Non-ROI Blur	86.015	86.570	88.246	16.103	64.709	0.956	0.971	0.978

Table 2: *Average reconstruction quality under fixed-bitrate setting (100 kbps). Mean PSNR and MS-SSIM are averaged over all sequences. ΣΔPSNR is computed relative to the corresponding w/o SR baseline within each scenario.*

	Mean PSNR (dB)			ΣΔPSNR (vs w/o SR)		Mean MS-SSIM		
Scenario	w/o SR	LW	3CH	LW	3CH	w/o SR	LW	3CH
Sharp Bicubic	77.794	77.823	77.743	0.873	-1.536	0.890	0.889	0.893
Blur Bicubic	77.305	77.518	77.494	6.380	5.666	0.869	0.882	0.882
Non-ROI Blur	76.826	77.069	77.079	7.291	7.587	0.865	0.869	0.872

	Mean ROI PSNR (dB)			ΣΔROI PSNR (vs w/o SR)		Mean ROI MS-SSIM		
Scenario	w/o SR	LW	3CH	LW	3CH	w/o SR	LW	3CH
Sharp Bicubic	79.825	80.053	80.426	6.610	17.435	0.729	0.740	0.745
Blur Bicubic	79.659	79.859	80.319	5.798	19.158	0.658	0.702	0.699
Non-ROI Blur	79.825	80.536	81.584	8.240	16.836	0.729	0.787	0.805

6.3 ROI-Specific Evaluation

Although global PSNR and MS-SSIM metrics provide a broad view of reconstruction quality, many edge-centric and AI-driven applications, such as surveillance or robotics, prioritize the fidelity of specific semantic regions, often referred to as regions of interest (ROIs). We evaluated the ROI vs. full frame reconstruction quality across two encoding constraints, **fixed QP (QP=22)** and **fixed bitrate (100 Kbps)**, using both the **3-channel YUV** and **Lightweight Y/uv** models. The results are shown in the tables 1,2.

- **Non-ROI Subsampling achieves the lowest bitrate:** From the last section, among the three pre-processing strategies (sharp, blur, non-ROI blur), the *non-ROI* subsampling consistently yields the **lowest overall bitrate**, demonstrating its efficiency in suppressing irrelevant content.
- **ROI Quality Is Prioritized and Preserved:** While the *full-frame* PSNR and MS-SSIM in non-ROI setups are slightly lower, they remain competitive. Notably, the **ROI regions achieve higher PSNR and MS-SSIM** across nearly all conditions, especially under constrained bitrate, proving that this strategy effectively reallocates quality toward semantically important areas.

A negative cumulative ΔROI PSNR is observed for the lightweight (LW) SR model in the blur bicubic scenario under fixed-QP (QP=22). This can be attributed to the compact network capacity of the LW model. Since the input has already undergone blur preprocessing, the LW architecture may not possess sufficient representational capacity to fully reconstruct high-frequency details within the ROI region.

However, the corresponding MS-SSIM value increases compared to the w/o SR baseline. This indicates that although pixel-wise reconstruction fidelity (PSNR) is reduced, the LW model is still able to preserve structural consistency within the ROI. In other words, structural similarity is maintained even when absolute pixel accuracy slightly degrades.

Additionally, in the non-ROI blur scenario, the LW model performs better in ROI recovery. This suggests that the LW architecture is more suitable for spatially selective degradation patterns rather than globally blurred inputs.

This evaluation highlighted the practical benefits of combining selective degradation with SR-based

recovery. Specifically, when the bandwidth is limited, protecting the semantic quality of the ROI regions enables more intelligent and efficient compression strategies suited for edge deployment.

It should be noted that the definition of regions of interest (ROIs) is inherently task-dependent and may influence both compression behavior and evaluation outcomes. Although the ROI categories considered in this study reflect common scenarios in surveillance and autonomous systems, prioritizing specific semantic regions can introduce bias by emphasizing certain content while suppressing others. In addition, degradation of non-ROI regions may lead to partial loss of contextual information that could be relevant for downstream tasks beyond those explicitly considered. Moreover, the quality of perceptual reconstruction within ROIs is used as a proxy for semantic preservation; however, the relationship between perceptual quality and the performance of downstream tasks is task-dependent and not guaranteed.

The proposed framework itself is task-agnostic and does not depend on a specific ROI generation mechanism. Different ROI definitions or generation strategies can be incorporated depending on application requirements, allowing practitioners to balance semantic focus and contextual preservation without modifying the overall pipeline design.

6.4 3CH vs Lightweight Models Comparison

To evaluate the efficiency of our proposed models, we compared the standard 3-channel YUV generator (3CH) and the lightweight Y/uv generator (LW) in terms of parameter count, model size, FLOPs, and runtime performance across multiple video sequences. The key differences are summarized in Table 3.

Table 3: *Model Efficiency Comparison between 3CH and Lightweight Y/uv.*

Model	Input Shape	Input Size (MB)	Params (M)	FLOPs (G)
3CH SR	YUV: 3×320×180	0.69	1.55	514.74
LW SR	Y: 1×320×180 UV: 2×160×90	0.35	1.02	88.06

To evaluate computational performance, we compare the inference speed between the lightweight Y/uv (LW) model and the 3-channel YUV (3CH) model across different scenarios and target settings. The percentage speed difference is computed per video sequence, where **negative values indicate that the lightweight model is faster**. In terms of inference performance, Figure 9 illustrates the speed difference (%) between the LW and 3CH models in two representative scenarios: *blur bicubic* and *non-ROI blur*. We omit the *sharp bicubic* case as it yields similar results to the others.

For both **QP=22 and bitrate=100 Kbps** configurations, the LW model consistently demonstrates **faster processing times across most video sequences**, particularly in the *non-ROI blur* scenario. In some cases, the speed improvement exceeds **20–40%**, while only a few sequences show the 3CH model slightly outperforming. This improvement is attributed to the LW model's reduced UV input resolution and more compact architecture.

In general, the **lightweight Y/uv model achieves substantial acceleration**, offering an efficient solution for resource-constrained deployments, while maintaining competitive reconstruction quality. The practical deployment on edge devices imposes constraints on latency, memory footprint, and power consumption. Although this work does not provide hardware-specific benchmarks, these constraints are explicitly considered in model design. From a system perspective, the reduced model size and FLOPs directly translate to lower inference latency and power consumption on edge devices. This design allows practitioners to trade the fidelity of the reconstruction for computational efficiency depending on the application requirements. Hardware-specific optimization and runtime profiling are left as future work.

Although the proposed framework is motivated by task-driven applications, ROI-focused perceptual quality remains a fundamental and widely adopted evaluation criterion in video coding systems. Consequently, this study evaluates the performance of reconstruction using PSNR and MS-SSIM to assess the effectiveness of ROI preservation and overall visual fidelity under constrained encoding settings. Explicit downstream task-level evaluation, which depends on specific application objectives and task models, is beyond the scope of this work and is identified as an important direction for future research.

6.5 Visual Reconstruction Results

To qualitatively assess the fidelity of the reconstructed video frames, we present representative visual comparisons from two different encoding configurations: fixed QP (QP=22) and fixed bitrate (100 Kbps). Each sample demonstrated a full processing pipeline from preprocessing to HEVC compression and final super-resolution reconstruction using both the 3-channel and lightweight Y/uv models.

Super-resolution models effectively restore lost quality:

- Both 3-channel YUV and lightweight Y/uv models substantially boost the PSNR compared to their non-SR counterparts.
- The 3-channel model yields the highest reconstruction fidelity, whereas the lightweight model achieved a competitive performance at a lower computational cost.

Additionally, non-ROI blur preprocessing delivers the lowest overall bitrate by sacrificing the quality in the background regions. When paired with SR, this configuration maintains strong ROI fidelity at a minimal bitrate.

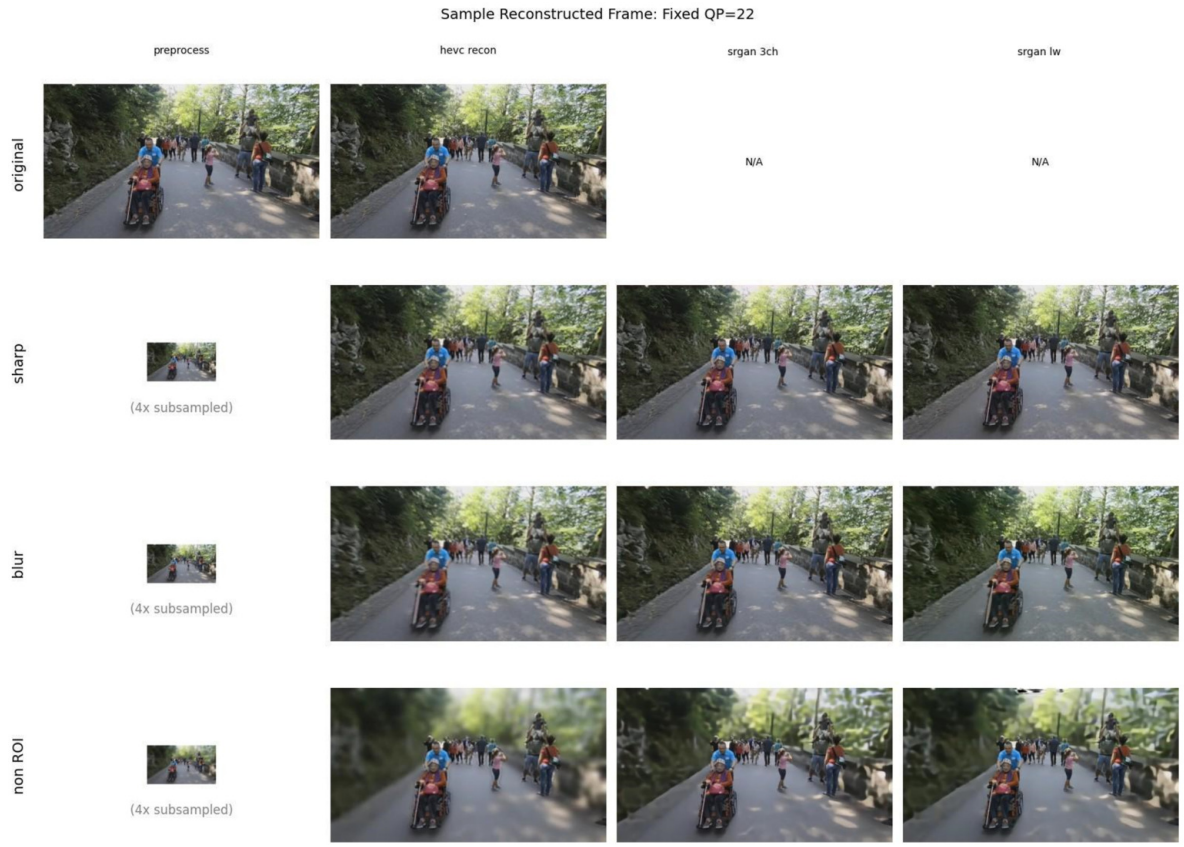


Fig.7: *Sample reconstructed frames for fixed QP=22 scenario. Each row shows preprocessing, HEVC recon, and SR results under sharp, blur, and non-ROI blur subsampling.*

7. CONCLUSION

This research proposes a YUV-based super-resolution (SR) framework for bitrate-constrained video compression, targeting both full-frame and ROI fidelity. By integrating SR into the post-HEVC pipeline, we show that aggressive downsampling (sharp, blur, non-ROI blur) can be effectively compensated with minimal quality loss.

Experiments under fixed QP and bitrate settings demonstrate that both the 3-channel and lightweight Y/uv models significantly improve PSNR and MS-SSIM over HEVC-only baselines. The lightweight model offers a strong balance of quality and efficiency, making it suitable for edge deployment.

ROI-focused analysis confirms that selectively degrading non-ROI areas allows efficient bitrate use while preserving task-relevant content. Visual results further validate the structural and perceptual gains. Overall, the framework provides a practical solution for task-driven video enhancement under real-world constraints. Future work may explore adaptive ROI selection, temporal consistency, and chroma enhancement.

ACKNOWLEDGEMENT

The author would like to thank ChatGPT for assisting in code development and paper writing. The author also acknowledges Obodroid Corporation Limited, where prior work on AI and robotics—particularly in surveillance robotics—contributed to foundational insights in this research.

AUTHOR CONTRIBUTIONS

Conceptualization, L.L., N.N., P.S. and A.S.; methodology, L.L. and P.S.; software, L.L. and N.V.; validation, L.L., N.N. and P.S.; formal analysis, L.L., P.S.; investigation, L.L., P.S.; data curation, L.L., N.V.; writing—original draft preparation, L.L.; writing—review and editing, L.L., P.S., N.N. and A.S.; visualization, L.L., N.V.; supervision, P.S., N.N. and A.S. All authors have read and agreed to the published version of the manuscript.

References

[1] J. Xu, Z. Xiong and S. P. Bhattacharyya, "PIDNet: A Real-time Semantic Segmentation Network Inspired by PID Controllers," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, pp. 19529-19539, 2023.

Fig.8: *Sample reconstructed frames for fixed bitrate (100Kbps) scenario. Rows represent different preprocessing strategies (sharp, blur, non-ROI blur); columns compare HEVC and SR outputs.*

[2] B. Li, H. Li, L. Li and J. Zhang, "λ Domain Rate Control Algorithm for High Efficiency Video Coding," in *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 3841-3854, Sept. 2014.

[3] X. Wei, M. Zhou, H. Wang, H. Yang, L. Chen and S. Kwong, "Recent Advances in Rate Control: From Optimization to Implementation and Beyond," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 1, pp. 17-33, Jan. 2024.

[4] L. -C. Chen, J. -H. Hu and W. -H. Peng, "Reinforcement Learning for HEVC/H.265 Frame-level Bit Allocation," *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*, Shanghai, China, pp. 1-5, 2018.

[5] M. Zhou, X. Wei, S. Kwong, W. Jia and B. Fang, "Rate Control Method Based on Deep Reinforcement Learning for Dynamic Video Sequences in HEVC," in *IEEE Transactions on Multimedia*, vol. 23, pp. 1106-1121, 2021.

[6] J. Shi and Z. Chen, "Reinforced Bit Allocation under Task-Driven Semantic Distortion Metrics," *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, Seville, Spain, pp. 1-5, 2020.

[7] X. Li, J. Shi and Z. Chen, "Task-Driven Semantic Coding via Reinforcement Learning," in *IEEE Transactions on Image Processing*, vol. 30, pp. 6307-6320, 2021.

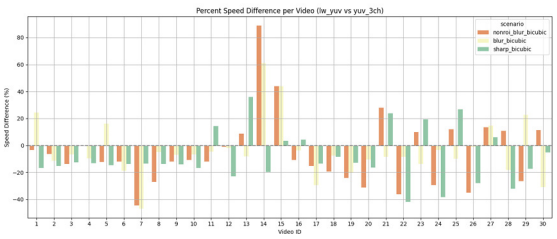
[8] T. Boulay, S. El-Hachimi, M. K. Surisetti, P. Maddu and S. Kandan, "Yuvmultinet: Real-time yuv multi-task cnn for autonomous driving," *arXiv preprint* arXiv:1904.05673, 2019.

[9] X. Wen, Z. Pan, Y. Hu and J. Liu, "Generative adversarial learning in yuv color space for thin cloud removal on satellite imagery," *Remote Sensing*, vol. 13, no. 6, p. 1079, 2021.

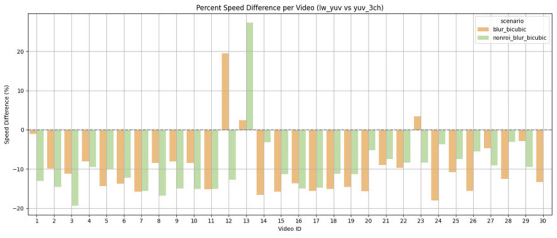
[10] Z.F.E. Mohammed Y. Abbass and H. Kasban, "Low-light image enhancement via improved lightweight yuv transformer-based models," *Journal of Visual Communication and Image Representation*, 2025.

[11] Z. Wang, J. Chen and S. C. H. Hoi, "Deep Learning for Image Super-Resolution: A Survey," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365-3387, 1 Oct. 2021.

[12] C. Dong, C. C. Loy, K. He and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295-307, 1 Feb. 2016.



(a) Speed Performance at QP=22



(b) Speed Performance at Target Bitrate = 100Kbps

Fig.9: *Percent speed difference between the Lightweight Y/UV and 3CH YUV models across video sequences.*

[13] C. Ledig *et al.*, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 105-114, 2017.

[14] X. Wang, K. C. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *European Conference on Computer Vision (ECCV) Workshops*, 2018.

[15] P. Isola, J. -Y. Zhu, T. Zhou and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 5967-5976, 2017.

[16] A. Kappeler, S. Yoo, Q. Dai and A. K. Katsaggelos, "Video Super-Resolution With Convolutional Neural Networks," in *IEEE Transactions on Computational Imaging*, vol. 2, no. 2, pp. 109-122, June 2016.

[17] R. Yang *et al.*, "Vsrresfeatgan: Video super-resolution with residual feature and adversarial networks," in *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[18] Y. Jo, S. W. Oh, J. Kang and S. J. Kim, "Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 3224-3232, 2018.

[19] J. Caballero *et al.*, "Real-Time Video Super-Resolution with Spatio-Temporal Networks and Motion Compensation," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 2848-2857, 2017.

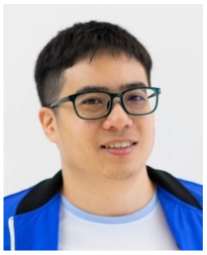
[20] Y. Tian, Y. Zhang, Y. Fu and C. Xu, "TDAN: Temporally-Deformable Alignment Network for Video Super-Resolution," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 3357-3366, 2020.

[21] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz and B. Catanzaro, "Video-to-video synthesis," *arXiv preprint* arXiv:1808.06601, 2018.

[22] Q. Liu *et al.*, "Video super-resolution based on deep learning: A comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[23] A. Bugeau, R. Giraud and L. Raad , "Influence of color spaces for deep learning image colorization," *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging*, pp. 847-878, 2023.

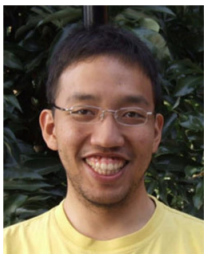
[24] S. Nah *et al.*, "NTIRE 2019 Challenge on Video Deblurring and Super-Resolution: Dataset and Study," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, pp. 1996-2005, 2019.



Lertluck Leela-amornsin received the B.Eng. degree in Electrical Engineering from Chulalongkorn University, Bangkok, Thailand, and the M.Sc. degree from the Graduate School of Information Science and Technology, University of Tokyo, Japan. During his M.Sc. studies, he conducted research on congestion control in delay-tolerant networks. His current research interests include video compression, super-resolution, task-driven video coding and AI deployment on edge devices.

Nuttapon Vanakittistien received the B.Eng. and M.Eng. in Computer Engineering from Chulalongkorn University, Thailand, in 2014 and 2017, respectively. He is currently a Pipeline Engineer at The Monk Studios in Bangkok, where the tools he developed support animated features. His research interests include procedural grooming, real-time simulation, task-driven video compression and super-resolution.

Nattee Niparnan received B.Eng and Ph.D. degrees in Computer Science from University of Virginia in 2007 and 2011. He then joined Computer Graphics Group at MIT as a Post-Doc in 2011 to 2014. He is currently an Assistant Professor at the Department of Computer Engineering, Chulalongkorn University. His research interests include real-time rendering, input/output devices and computational fabrication.



Pitchaya sitthi-amorn received the B.Eng., M.Eng., and Ph.D. degrees in computer engineering from Chulalongkorn University, Bangkok, Thailand, in 2001, 2003, and 2008, respectively. He is currently an Associate Professor with the Department of Computer Engineering, Chulalongkorn University, Thailand. His research interests include robotic grasping, localization, and mapping.

Attawith Sudsang was born in Bangkok, Thailand, in 1971. He received the B.Eng. degree in computer engineering from Chulalongkorn University, Thailand, in 1991, and the M.S. and Ph.D. degrees in computer science from the University of Illinois at Urbana–Champaign, USA, in 1994 and 1999, respectively. He is currently an Assistant Professor with the Department of Computer Engineering, Chulalongkorn University. His research interests include robotic grasping, mobile robot navigation, and virtual reality.