MULTIDISCIPLINARY RESEARCH INSTITUTE INDIA

Open Access International Journal of Science and Engineering

ISSN: 2456-3293

Volume 9 Issue 1s (2026) | Pages: 44-49

A Hybrid Deep Learning Technique for Compressed Video Enhancement

Urvashi Tej Bhat[1], Pallavi Sagar Deshpande[2]

[1,2]Department of Electronics & Communication Engineering, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, India

[1] *urvashi.bhat@bharatividyapeeth.edu,* [2] *psdeshpande@bvucoep.edu.in*

Peer Review Information

***Type**: Article*
***Received**: 29 March 2026*
***Revised**: 26 April 2026*
***Accepted**: 30 May 2026*
***Published**: 19 June 2026*

Abstract

In recent years, the high-quality video is highly demanded. But at the same time, video quality is degraded due to factors such as compression, noise, blur and various environmental conditions, which give rise to the need for using the various video enhancement techniques These techniques aim to recover the quality of degraded videos and therefore enhancing the visual quality of the videos by addressing tasks such as denoising, super-resolution, coloration , stabilization de-flickering, and reducing compression artifacts using traditional methods as well as modern methods ,which will be discussed in the paper later. Despite using notable advancements, several issues remain unresolved, and overcoming them is vital to achieving robust, efficient, and widely applicable solutions. Traditional video enhancement methods suffer from many limitations like not balancing computational efficiency and high-quality output, which give rise to degraded performance in real-time scenarios. Over the past few years, deep learning methods have brought substantial advancements in the enhancement of compressed video, prominently surpassing the traditional techniques in terms of enhancing the visual quality by reducing the compression effects manifolds. In this paper, an extensive overview of the video enhancement techniques is presented and a Hybrid Deep Learning Enhancement Technique for Compressed Video is proposed.

Keywords: Video Enhancement; Deep Learning; Noise Reduction; Motion Artifacts; Compressed Video Enhancement; Video Restoration.

How to Cite This Article

Bhat, U. T., & Deshpande, P. S. (2026). A Hybrid Deep Learning Technique for Compressed Video Enhancement. *Open Access International Journal of Science and Engineering*, 9(1s), 44-49.

Introduction

Digital video content plays an important role in entertainment, security, modern communication, and information. The vast usage of digital cameras, smartphones, and online video platforms has led to the drastic growth in both the generation and production of video content. As, video data is now ubiquitous, that play a crucial role in various sectors like security, entertainment, social media, surveillance and healthcare. Everyday millions of videos get uploaded on large scale on various social video platforms [1], while many surveillance systems heavily rely on video data to ensure security and safety. In-spite of many recent advancements done in video technologies, the process starting from capturing a video to delivering it to end-users involves lot of obstacles that can reduce the final video quality as shown in figure 1.

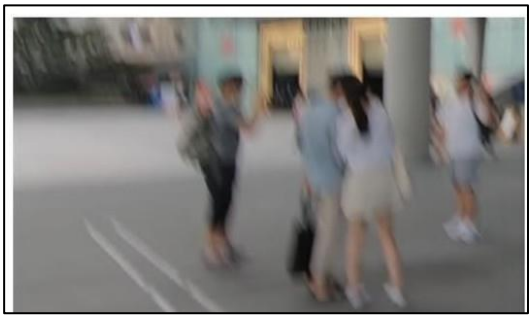

Fig. 1. *Degraded quality video frame*

Compression is one such critical stage where video quality degrades due to the introduction of artifacts [2]. Several video coding standards, such as H.264/AVC [3] and H.265/HEVC [4], rely on lossy compression techniques to minimize file size and for optimizing the usage of bandwidth efficiently. Though effective, these compression techniques introduce visible distortions, such as blocking and ringing effects, which significantly degrades the viewing experience [5]. Camera shake is another common problem, resulting in motion blur that affects scene[6] in the video under observation. Motion blur can also occur due to the movement of objects combined with a high camera exposure time. Also, Upscaling low-resolution videos to higher resolutions often leads to loss of detail and introduction of artifacts, as traditional interpolation methods are not able to reconstruct fine textures and edges, leading to blurred and unrealistic videos [7].Applying image manipulation algorithms for various tasks, such as enhancement, may introduce visual artifacts and temporal inconsistencies, further degrading the video frames [8].

In addition, environmental conditions like changing lighting, haze, and water droplets on camera lens can introduce further distortions, affecting the overall quality, clarity and sharpness [9]. Therefore, due to the above-mentioned causes, there is an immediate need for video enhancement techniques to remove all these problems and ensure high video quality.

Video Processing Techniques

A range of processes are involved in video enhancement, designed to improve the visual quality and enhancing specific features within the video sequences. As there is a direct relationship between the video quality and the effectiveness of data and consequently information extraction, whereas poor-quality visuals can lead to loss of information and hence economic losses. Therefore, the task of video enhancement plays a pivotal role in numerous applications, particularly in surveillance systems, where high-quality video is required for tasks such as criminal investigation, vehicle tracking, traffic flow analysis, and medical diagnostics [10]. By ensuring that the final video output is clear, detailed, and aligned with user expectations, video enhancement improves both the viewing experience as well as it leads to improved accuracy and more reliable performance in critical fields such as surveillance and medical imaging. Therefore, video enhancement techniques (Conventional) can be summarized as follows:

Video Super Resolution (VSR)

It is method which obtains high-resolution video frames generated from the corresponding low-resolution video frames. The methods used are:

- Interpolation Based Methods: These methods use interpolation algorithms to estimate missing pixels in the high- resolution version of the video.
- Nearest Neighbour Interpolation**:** This method replicates the value of the nearest pixel to fill in the missing pixels. It is the simplest method.
- Bilinear Interpolation: This method uses the weighted average of the four closest pixels to estimate the value of missing pixels. It is sophisticated method than the Nearest neighbour method.

- Bicubic Interpolation: This method uses the weighted average of the 16 closest pixels to estimate the value of missing pixels. It is more sophisticated method than Bilinear Interpolation method and gives the better results.

Reconstruction Based Methods

These methods use Image or video processing techniques to reconstruct the high- resolution version of video.

- Optical Flow based Method: This method calculates the motion between consecutive frames in a video or image sequence. They model how pixels move over time.
- Spatial –Temporal Method: This method is used to enhance the resolution of videos (both in space and time).

Denoising

Denoising is the technique to remove the noise from the video to improve its visual quality. The methods used are:

- Spatial filtering: This method involves applying a filter to each frame of the video to reduce noise using Median, Gaussian filters, Mean filters.
- Temporal filtering: This method involves using information from multiple frames to reduce noise using Kalman, Recursive filters.

Color Correction

It is the method to adjust the colors of the video to improve its visual quality. The techniques used are:

- White balance
- Color Grading
- Color Matching

Deblurring

It is a technique restoring a video that looks blurry so that frames appear sharper and clearer. Blurring usually happens because of camera motion, object motion, defocus, or low light during recording. There are several techniques that can be used for deblurring:

- Inverse filtering: This method uses a known blur function to reverse the blurring effect. This method is highly sensitive.
- Wiener filtering: This method uses an Image statistical model and the blur function to find the original image or frame.
- Blind deconvolution: This technique attempts to find the blur function and the original image simultaneously as the blur function is not known in this technique.

Stabilization

It is a processing technique used to reduce unwanted shake, jitter, or motion in video. The goal is to make video appear smoother, more professional, and easier to watch by compensating for unintended camera motion. The methods used are:

- Optical Flow: It measures how pixels move from one frame to the next to estimate the camera motion.
- Feature-based tracking: It detects and tracks key feature points like corners or edges across frames to estimate motion.

Deep Learning Techniques Used for Compressed Video Enhancement

Deep learning & AI algorithms have proven excellent success in video enhancement and emerged as a powerful tool for video enhancement. By dominating both spatial and temporal details, these techniques can efficiently reduce compression artifacts and restore visual detail which cannot be achieved by traditional methods [11-12]. Therefore, deep learning techniques for video enhancement are categorized into Convolutional Neural Network (CNN), Deep Neural Network (DNN), Recurrent Neural Network (RNN), and Hybrid approach.

1. Convolutional neural network (CNN): The performance of Convolutional neural network (CNN) based algorithms is good and at par by removing artifacts in compressed videos. The architecture of a Convolutional Neural Network (CNN) for video enhancement is an end-to-end network, mostly very similar to standard image enhancement models which are adapted to process the spatial (within frames) and temporal (across frames) dimensions of video data. A CNN architecture for video enhancement has following mentioned key Components:

a) Input Layer: Input layer receives the raw and degraded video, which is treated as a sequence of frames, forming a three-dimensional volume of data.

b) Convolutional Layers:

These are the essential building blocks. In the context of video, they can be 2D or 3D:

- 2D Convolutional Layer: It is applied to individual frames to extract spatial features like edges, textures, and shapes within each image.
- 3D Convolutional Layer: It is applied across multiple consecutive frames to capture temporal information (e.g., motion patterns) in addition to spatial feature.

c) Activation Functions: At this stage, a Rectified Linear Unit (ReLU) activation function follows each convolutional operation to introduce nonlinearity, thereby allowing the network to model complex relationships and patterns.
d) Pooling Layers (Down sampling): This layer is periodically inserted to reduce the spatial dimensions (and sometimes temporal dimensions in 3D CNNs) of the feature maps. This reduces computational cost and makes the model more robust to small variations or shifts in the video frames. Max pooling is the most common method. Up sampling Layers (Deconvolution or Transposed Convolution): In enhancement tasks, the network usually needs to output video of the same (or higher) resolution as the input, which is achieved by Up sampling layers.
e) Output Layer: It produces the enhanced video (e.g., higher resolution, lower noise, or improved visual quality) Workflow:

- Feature Extraction (Encoder): The hierarchical features, from low-level edges to high- level information are extracted from the convolutional and pooling layers.
- Feature Integration/Processing: The network processes these features to understand motion and temporal context.
- Reconstruction (Decoder): The extracted features are used to reconstruct the final, enhanced video.
- Training: The network is trained using labelled data (pairs of low-quality and high- quality videos), using a loss function and an optimization algorithm.

2. Deep neural network (DNN):A deep neural network (DNN) for video enhancement typically integrates convolutional neural networks (CNNs) for spatial feature extraction from individual frames and recurrent neural networks (RNNs), including LSTMs, to model temporal relationships between successive frames. The goal is to produce a high-quality video from a low-quality input. A typical DNN architecture for video enhancement consists of the following mentioned key Components:
a) Input Layers: The network takes in a sequence of low-resolution (LR) or low- quality video frames and outputs a corresponding sequence of high-resolution (HR) or enhanced frames.
b) Feature Extraction (Spatial): The initial layers often consist of CNNs, which use convolutional and activation (commonly ReLU) layers to extract hierarchical features (edges, corners, textures, etc.) from each frame.
c) Up sampling /Reconstruction: The processed low-resolution feature maps are upscaled to the desired high resolution through a dedicated-up sampling layer.
d) Output Layer: The final layer reconstructs the full-colour (e.g., RGB channels) high-quality video frame.

3. Recurrent Neural Network (RNN): The Recurrent Neural Network (RNN) for video enhancement is typically composed of a hybrid architecture that is a combination of a Convolutional Neural Network (CNN) feature extractor with an RNN-based predictor. By leveraging CNNs for spatial feature extraction and RNNs for temporal modelling, the model effectively captures both spatial and temporal information. across the sequence of frames. The process can be broken down into two main components:
a) CNN Feature Extractor (Spatial Processing)
Each video frame is processed independently and relevant spatial features are extracted. The various layers are:

- Input Layer: The low-quality degraded video frames (as images) are fed into the CNN.
- Convolutional Layers: Hierarchical features, including edges, textures, and object patterns, are captured through multiple convolutional layers employing various kernel sizes (e.g., 3×3 and 5×5).
- Pooling Layers: At this stage, spatial dimensions of the feature maps are reduced which makes the model more computationally efficient and robust to minor variations and changes.
- Output Layer: A fixed-length feature vector is obtained as output for each frame, summarizing its visual content. This vector is fed as the input to the RNN component.

b) RNN Predictor/Processor (Temporal Processing) The RNN Predictor takes the feature vectors from the CNN and models the temporal relationships between frames to enhance the video.
Workflow:

- Spatial features are extracted frame wise by the CNN. Temporal features are analyzed by the RNN.
- The combined spatial and temporal features are combined to estimate the model to perform video enhancement.

4. Hybrid based Approach: A hybrid architecture for video enhancement is the combination of different processing techniques or models with deep learning models. The goal is to achieve better performance with high accuracy. The architecture can vary depending on the application but general components and configurations remain the same.

Proposed Method

In this paper, HDVE-Net which stands for Hybrid Deep Video Enhancement Network is proposed which combines spatial enhancement, temporal consistency, and optimization-based learning to reduce compression artifacts (blocking, blurring, ringing) in compressed videos while preserving motion details. Its architecture consists of:

1. Spatial Feature Enhancement Module (SFEM): In this module, Residual CNN (RCNN) is used to enhance each frame mainly focusing on blocking artifacts and texture loss. As CNNs are very effective at restoring spatial details lost during compression, it uses convolutions to capture wider context without increasing computation.

2. Temporal Motion Refinement Module (TMRM): In this module, ConvLSTM **is** used which takes multiple consecutive frames as input. It learns motion-aware features to reduce flickering and maintains temporal consistency.

3. Attention-Based Fusion Module (AFM): It employs attention mechanisms to combine features from different video frames at different scales (e.g., multi-scale, multi-modal) effectively. The primary goal is to adaptively emphasize the most relevant information and suppress noise during the fusion process. It selectively enhances important regions such as edges and moving objects. It suppresses noise from static background areas.

4. Optimization Technique

Optimization based Hybrid Loss Function is:

$$L = \lambda_1 L_{MSE} + \lambda_2 L_{Perceptual} + \lambda_3 L_{Temporal} \qquad (1)$$

where:
L_{MSE} is MSE Loss is related to pixel accuracy $L_{Perceptual}$ is Perceptual Loss is related to visual quality.
$L_{Temporal}$ is Temporal Loss is related to frame-to-frame smoothness

Methodology

The methodology of the proposed technique is given below:

- Input: The input is Compressed video (e.g., H.264 / HEVC) which contains Compression artifacts (blocking, ringing, blurring), Quantization noise, Motion- compensation errors. These artifacts are especially strong at low bitrates and around edges and motion regions.
- Extract consecutive frames: The input video is decoded into a sequence of frames that are usually grouped into short temporal windows (e.g., 5–15 frames).
- Enhance spatial features per frame: In this step, each frame is processed independently to recover spatial details.
- Refine motion across frames: This stage models temporal relationships and motion consistency.
- Apply attention-based enhancement: Attention mechanisms selectively enhance important features. The several types of attention are Spatial attention that focuses on important regions (edges, faces, text), Temporal attention that weights reliable frames and Channel attention that emphasizes informative feature channels.in the proposed method, channel attention is employed. This step refines features into a clean, high-quality representation.
- Output high-quality reconstructed video: In this step, enhanced features are decoded back into RGB frames where frames are reassembled into a video sequence. Output matches the original resolution and frame rate.

Advantages

The advantages of the proposed method are as under:

- Reduces compression artifacts
- Maintains temporal consistency
- Computationally efficient
- Efficient for real-time applications
- Flexible

Limitations

1. High Computational Cost & Memory Requirements: The proposed method performs high quality video enhancement but it also demands significant requirement of computing resources and memory specifically for high-resolution video sequences that can make training and real-time deployment difficult.
2. Complexity vs. Processing Speed Trade-Off: As the video complexity increases, efficiency in terms of processing speed decreases, which can be viewed as the limitation, consequently making it challenging to deploy in real-time applications.
3. Training Datasets: Video enhancement tasks require diverse and high-quality video datasets for training. Lack of realistic and varied training data (especially for real-world degradations such as motion blur, compression noise, and lighting variation) limits generalization. Many existing datasets are domain-specific or synthetic.

Future Scope

Hybrid video enhancement networks promise superior performance by combining complementary learning mechanisms, yet they currently face challenges in efficiency, data requirements, temporal reliability, and real-world applicability. Future research is pushing toward lightweight, adaptive, and multimodal models that balance performance, speed, and generalization across diverse real-world video scenarios.

Acknowledgment

The author would like to express sincere gratitude to Dr. Pallavi Deshpande for her invaluable guidance, insightful suggestions, and continuous support Her expertise and encouragement significantly contributed to the successful completion of this work.

References

1. Hongming Luo et al. “Restoration of user videos shared on social media”, Proceedings of the 30th ACM International Conference on Multimedia, pp. 2749–2757, 2022
2. Kai Zeng et al. “Characterizing perceptual artifacts in compressed video streams”, Human Vision and Electronic Imaging XIX. Vol. 9014.
3. SPIE., pp. 173–182, 2014
4. Gary J Sullivan et al. “The H. 264/AVC advanced video coding standard: Overview and introduction to the fidelity range extensions”, Applications of Digital Image Processing XXVII 5558, pp. 454–474, 2004.
5. Gary J Sullivan et al. “Overview of the high efficiency video coding (HEVC) standard”, IEEE Transactions on Circuits and Systems for Video Technology, pp. 1649–1668, 2012
6. Ren Yang et al. “multi-frame quality enhancement for compressed video”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition., pp. 6664–6673, 2018
7. Shuochen Su et al. “Deep video deblurring for hand-held cameras”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1279–1288,2017
8. Hongying Liu et al. “Video super-resolution based on deep learning: a comprehensive survey”, Artificial Intelligence Review 55.8, pp. 5981–6035 ,2022
9. Chao Li et al. “Video flickering removal using temporal reconstruction optimization”. In: Multimedia Tools and Applications 79, pp. 4661–4679,2020
10. Julie Delon et al. “Stabilization of flicker-like effects in image sequences through local contrast correction”. In: SIAM Journal on Imaging Sciences 3.4, pp. 703–734, 2010
11. Jie Gui et al. “A comprehensive survey and taxonomy on single image dehazing based on deep learning”. In: ACM Computing Surveys 55.13s, pp. 1–37, 2023
12. Hong Wang et al. “Survey on rain removal from videos or a single image”, Science China Information Sciences 65.1, pp. 1–23, 2022
13. Redouane L hiadi et al. “Deep learning-based techniques for video enhancement, compression and restoration”, International Journal of Artificial Intelligence, pp 1518-1530, 2025