Baghdad Science Journal

Volume 23 | Issue 5 Article 10

5-20-2026

# QoE Enhancement for Video Streaming Based on Attention U-Net Model

Iman Kadhum Abbood
*Department of Software, College of Information Technology, University of Babylon, Babylon, Iraq*,
iman.abbood@uobabylon.edu.iq

Doaa Ayed Mohammed
*Department of Software, College of Information Technology, University of Babylon, Babylon, Iraq*,
Doaa.aied@uobabylon.edu.iq

Israa Hadi Ali
*Department of Software, College of Information Technology, University of Babylon, Babylon, Iraq*,
israa_hadi@itnet.uobabylon.edu.iq

Sarah Abdul-Ridha Abed
*Department of Software, College of Information Technology, University of Babylon, Babylon, Iraq*,
sarahaba@uobabylon.edu.iq



## How to Cite this Article

Abbood, Iman Kadhum; Mohammed, Doaa Ayed; Ali, Israa Hadi; and Abed, Sarah Abdul-Ridha (2026) "QoE Enhancement for Video Streaming Based on Attention U-Net Model," *Baghdad Science Journal*: Vol. 23: Iss. 5, Article 10.
DOI: https://doi.org/10.21123/2411-7986.5294

BAGHDAD SCIENCE JOURNAL 2026;23(5):1659–1674



RESEARCH ARTICLE

# QoE Enhancement for Video Streaming Based on Attention U-Net Model

**Iman Kadhum Abbood** *, **Doaa Ayed Mohammed**, **Israa Hadi Ali**, **Sarah Abdul-Ridha Abed**

Department of Software, College of Information Technology, University of Babylon, Babylon, Iraq

## ABSTRACT

Video streaming services, including YouTube, Netflix, and Amazon, account for a large portion of global internet traffic. Therefore, the maintenance of high visual quality is very important to user satisfaction and overall Quality of Experience (QoE) for these users. However, fluctuating bandwidth often causes frame drops and motion discontinuities, resulting in visual artifacts and degraded QoE. Although existing video frame interpolation methods attempt to address these challenges, most remain limited in their ability to sustain performance under bandwidth-constrained environments. To overcome these limitations, IncepU-Net has been introduced. IncepU-Net is proposed as an enhanced U-Net architecture that integrates multi-scale inception modules and attention-gated skip connections for intelligent video frame interpolation in bandwidth-limited settings. IncepU-Net was trained and evaluated on the Vimeo90K and UCF101 datasets. The model at the testing stage achieved 35.19 dB PSNR and 0.958 SSIM using Vimeo90K, and 30.08 dB PSNR and 0.933 SSIM on UCF101. IncepU-Net surpasses several state-of-the-art interpolation methods. These results demonstrate that IncepU-Net effectively interpolates missing frames, mitigates playback stutter, and maintains motion continuity, thereby establishing it as a robust and practical solution for improving QoE in real-world video streaming applications, bridging the gap between restricted network delivery and perceptually meaningful frame synthesis.

**Keywords:** Attention mechanisms, Frame interpolation, Inception modules, QoE, QoS, U-Net architecture

## Introduction

Online video streaming has been completely transformed by the rapid development of Internet speeds, which have made multimedia content accessible on a variety of devices. As such, with features and resolutions that can be adjusted to suit personal preferences, users may now stream video content anywhere they go. In addition, maintaining a high standard of user experience is essential to the effective adoption and long-term expansion of digital streaming services.[1] Moreover, live video streaming, responsible for 82% of internet traffic, is based upon three main components: a client for upload, a server for the media, and a receiver with the aid of adaptive bitrate algorithms. However, a key issue is the limited uplink bandwidth, which forces the upload client to degrade video quality during encoding, thereby negatively impacting QoE despite innovations in streaming technology.[2]

Furthermore, network disruption issues such as delays, packet loss, and jitter can further degrade audiovisual quality, impacting the overall user experience,[1] and adaptive mechanisms like frame rate reduction, originally designed for energy savings in wireless networks, can introduce motion artifacts and temporal inconsistency.[3] High frame rates are essential for maintaining fluid motion, and the human visual system (HVS) is particularly sensitive to frame rate variation.[4] Service providers must guarantee users access to steady, reliable performance every

Received 12 June 2025; revised 16 August 2025; accepted 21 September 2025.
Available online 20 May 2026

* Corresponding author.
E-mail addresses: iman.abbood@uobabylon.edu.iq (I. K. Abbood), Doaa.aied@uobabylon.edu.iq (D. A. Mohammed), israa_hadi@itnet.uobabylon.edu.iq (I. H. Ali), sarahaba@uobabylon.edu.iq (S. A.-R. Abed).

https://doi.org/10.21123/2411-7986.5294

time. Then, multiple video measurement technologies have been created to develop both the quality evaluation and user experience enhancement procedures. Since video quality fundamentally depends on how the HVS interacts with the content presented to users, visual perception shapes the overall watching experience.[1] It is highly sensitive to motion smoothness and temporal coherence; it is crucial in determining the way quality of experience is evaluated. Subjective assessments are still required to evaluate user satisfaction. Therefore, QoE becomes a critical metric that measures both technical aspects of a service and user-perceived levels of satisfaction. Over recent years, the importance of QoE for users to assess their level of satisfaction with a presented item of content has increased. QoE represents both technical measurements of quality and user feedback on how well they perceive an item of content, generally rated from "poor" to "excellent". Further, the distinction between Quality of Service (QoS), and QoE should be made. Specifically, QoS involves the network performance attributes and operational features of IP-based networks and services. On the other hand, QoE relates to the subjective evaluation of an end-user's total experience and level of satisfaction when using a particular application, system or service.[5] However, traditional empirical studies evaluating the perceptual effects of adaptive streaming have consistently isolated the perceived effects of changes in video quality due to dynamic quality transitions through varying temporal fluctuations in bitrate allocations while maintaining a consistent average bitrate for all experimental conditions.

### Problem statement

Although adaptive bitrate (ABR) streaming protocols like Dynamic adaptive streaming schemes over HTTP (DASH) are ubiquitous, perceptual video quality in bandwidth-constrained scenarios is still a major issue. In order to provide continuous video streaming, many video streaming platforms use extreme compression, lower the display resolution, or drop frames. These methods maintain continuity; however, most commonly they produce temporal distortions, motion disruption, and/or visual inconsistencies that degrade the user's QoE. It has been shown by Zinner et al.,[6] that users are more sensitive to the reduction of the frame rate rather than an equivalent loss of the display resolution. This demonstrates how important it can be for adaptive strategies to have a perceptual focus. Saha et al.[7] further emphasized the complex, content-dependent interactions among low-level QoE metrics, such as buffering events, startup delay, bitrate, and switching frequency, which complicate efforts to model and optimize user experience. Therefore, this research aims to develop a new method to utilize frame interpolation at the client side utilizing the user device. More specifically, this research introduces a deep learning-based architecture called IncepU-Net. IncepU-Net is a novel neural network structure that utilizes multiple scale inception modules and attention-gated skip connections inside a standard U-net structure. The primary function of IncepU-Net will be to accurately fill in missing frames while utilizing both spatial and temporal information to restore a continuous viewing experience and therefore improve the QoS in a bandwidth-constrained network. By enabling real-time frame synthesis at the client level, IncepU-Net offers a scalable and perceptually robust solution for modern video streaming applications.

The main contributions of this work are as follows:

1 **Attention Enhanced U-Net Architecture**: Leveraged attention mechanisms within a U-Net backbone to enhance motion-aware feature refinement for video frame interpolation.
2 **Inception-Enhanced U-Net Architecture**: The incorporation of inception modules into the U-Net improves multi-scale feature extraction, improving the capability of the model to capture fine details.
3 **Charbonnier-SSIM Loss:** Unified Charbonnier-SSIM Loss for joint perceptual and pixel-wise optimization.

### Related works

There are many researchers related to the field of QoE and frame prediction, some of whom will be explained.

#### Frame and packet loss in video delivery

Multimedia delivery over IP networks faces significant challenges due to network bottlenecks, such as server-side congestion, Internet service provider interconnections, and last-mile connections, which cause delays and packet losses, thereby negatively impacting QoE. While Dynamic adaptive streaming, such as DASH,[8] mitigates some of these issues through adaptive bitrate selection and segment-based delivery, its reliance on the Transmission Control Protocol (TCP) introduces additional complexity due to TCP's unpredictable behavior.[9] The DASH video player dynamically adjusts its streaming parameters, such as video bitrate, spatial resolution, and frame rate, by selecting the most suitable encoded stream from the available options. This adaptive selection process is governed by multiple dynamic

parameters, including current playback speed, buffer occupancy levels, and real-time network throughput measurements.[8] The diversity of devices (smartphones, tablets, smart TVs) and connection types (cable, Wi-Fi, cellular) further complicates QoE optimization. Optimizing QoE remains a complex task due to the inherently interdependent and often nonlinear nature of its underlying metrics. Scholarly work, such as the comprehensive survey conducted by Kua et al.,[9] demonstrates an increasing interest in the measurement of QoE during dynamically changing streaming conditions. As the industry transitions from Unicast Datagram Protocol (UDP)-based transport to Internet Protocol (TCP)-based transport via Dynamic Adaptive Streaming over HTTP (DASH), DASH is becoming the standard for delivering video content over the Internet. By utilizing adaptive bitrate technology at the client level, DASH allows for near-real-time changes to video quality, which will improve continuity of play and overall QoE. Unlike continuous streaming, DASH transmits video in multi-second chunks, allowing efficient adaptation to network conditions and device capabilities. The client controls the process through real-time bandwidth assessment by choosing appropriate bitrates, which control the server transmission rate. The approach enables DASH to become a highly adaptable system that delivers exceptional streaming quality across various network conditions.[9] Authors in,[10] present "Sammy", a video streaming architecture that jointly integrates adaptive bitrate selection with application-informed pacing to reduce network traffic burstiness while maintaining high QoE. The implementation smooths data flow and minimizes network conflicts to help maintain or enhance QoE. Evaluation on a large-scale streaming platform demonstrated enhanced traffic stability and more efficient network resource utilization compared to traditional ABR algorithms. Sammy's design includes a dynamic pacing strategy that adapts pacing rates based on playback buffer occupancy, accelerating downloads when the buffer is low and slowing down when the buffer is sufficiently filled.[10] Two Machine Learning Methods are being used to improve video quality adaptation - specifically using Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) based Recurrent Neural Networks (RNN), which allow the extraction of features in video streaming that can then be predicted with an optimum bitrate per frame. This allows users to have both a personalized and most efficient way to watch their desired videos. The use of CNN and LSTM together is highly beneficial for all real-time applications like dynamic adaptive streaming over HTTP, and is able to provide the ability for users to adapt the video quality of their preferred content through video content analysis, and by creating a model of how both the user and the network behave over time.[11] Eswara et al. introduce an advanced LSTM-based model, termed LSTM-QoE, for continuous prediction of video quality. In LSTM-QoE model, there are several cascading layers of LSTMs. These layers allow for the capture of both nonlinear behavior and complex temporal relationships inherent within QoE. They show how effective their model is with respect to predicting QoE dynamics via several analyses using data from a publicly available continuous QoE database.[12]

Existing approaches in adaptive streaming have focused on reducing bursty traffic patterns produced by DASH's on-off chunk delivery. While these solutions do provide rate smoothness and congestion avoidance, they have not addressed the issue of missing frames caused by network issues. Our work will address this gap by developing a frame prediction solution to allow for content recovery during off periods, which will improve continuity of playback and user experience.

### *Video frame reconstruction through interpolation*

Video frame interpolation is a major technique in computer vision that enables the generation of intermediate frames between the already existing video frames. This ability is particularly important in areas like autonomous driving and robotics, where systems must be able to predict and determine future states of dynamic environments. Interpolation-based frame reconstruction improves the temporal resolution and leads to a stronger understanding of the scene, which is critical in decision-making in real-time, safety-critical systems. Prior to the advent of Deep Learning, Convolutional Long Short-Term Memory Models (ConvLSTM) models,[13] have greatly advanced video frame interpolation, improving both accuracy and temporal consistency. Based on what was observed in the previous frames, these models can be useful in making predictions regarding the appearance of future frames. Aravinda et al.[14] proposed an application of a hybrid architecture based on ConvLSTM and 3D CNN models for video frames prediction. They have chosen to use ConvLSTM to identify the temporal relations between successive frames and 3D CNN to extract the spatial characteristics. Their objective was to increase the precision of predictions with respect to existing techniques. They demonstrated the superiority of the hybrid model against the individual models ConvLSTM and 3D CNN by obtaining better Mean Squared Error values and higher $R^2$ values on the same benchmarks. Jatana et al.,[15] investigate frameworks for future frame prediction based on Generative Adversarial Networks, along

with combinations of CNNs, LSTMs and ConvLSTMs. A total of two models were developed. The first model was successful in predicting the 20th frame given the 19 previous frames. After performing the process three times, however, it lost accuracy because of limitations in the convolutional 3-dimensional block of the Conv3D module and also due to an increase in noise. The second model had success in predicting the 10th frame given nine input frames, capturing motion probabilistically but faltering under occlusions and fading by the tenth iteration. Both models struggled to be accurate with longer sequences, with noise sensitivity, temporal insensitivity, and resource constraints being the main problems limiting training depth. The results indicate that the architectural methods that are noise-resilient and time-sensitive, as well as increased computational capabilities, are fundamental to strong long-term predictions. Recent developments in streaming QoE prediction have progressively depended on machine learning methods to model temporal dynamics effectively.[16] Specifically, the M-3R framework uses an LSTM network-based memory-based model that captures the evolution of streaming QoE on a frame-by-frame basis. By integrating a combination of full-reference, reduced-reference, and no-reference video quality assessment metrics as input features, M-3R adeptly accounts for the combined effects of rate adaptation and rebuffering events. The results show that M-3R achieves greater accuracy and temporal consistency than traditional models and provides a more accurate and consistent prediction of QoE over time. Azadegan and Beheshti Shirazi[17] present an additive 3D CNN for improving video quality. It works fully in the decoder, allowing compatibility with conventional codecs including H.264, HEVC, and VVC. The model takes five consecutive frames of a compressed video as input and predicts the compression error, defined as the difference between the original uncompressed frame and the compressed frame. Adding this predicted distortion back into the compressed versions produces better quality outputs. This method is reference-free since no information about the original video is required, and it could easily be incorporated into current video decoding systems. Adaptive Collaboration of Flows (AdaCoF),[18] is a generalized warping module for video frame interpolation that jointly estimates kernel weights and offset vectors per pixel, enabling it to model complex motions effectively.

The method in[19] employs a U-Net architecture enhanced with residual blocks and a feature pyramid network (FPN) to improve motion representation and detail preservation. Shortcut connections are incorporated within the encoder to address the typical information loss caused by the U-Net's down-sampling and up-sampling operations. Meanwhile, the FPN in the decoder enables the model to capture features at multiple scales, strengthening the network's ability to model spatial-temporal relationships across frames. While[20] outlines scientific advances in the field of predicting future frames and suggests a generative adversarial network that uses the latest techniques in deep learning research. The model combines a U-Net-based autoencoder as the generator and a CNN as the discriminator. By training these components alternately, the system learns to generate realistic future frames, outperforming standard autoencoders in both accuracy and image clarity. An improved deformable separable convolution network has been proposed for video frame prediction in,[21] which enhances kernel-based methods by learning adaptive kernels, masks, offsets, and biases to capture non-local spatial information efficiently. Uniquely, it supports generating multiple intermediate frames between two inputs by incorporating a temporal control variable using an extended coord-conv technique. Future frame prediction using an adaptive flow network method known as Task-Oriented Flow (TOFlow)[22] integrates motion estimation and image processing within a unified end-to-end framework. By jointly training the flow estimation module alongside the video processing pipeline, TOFlow learns motion representations that are optimized for each specific task. This task-driven approach yields superior results compared to conventional methods that rely on generic optical flow.

Prior research in video streaming has predominantly advanced along two directions: network-level strategies aimed at smoothing throughput and enhancing adaptive bitrate (e.g., SAMMY[10]), and frame prediction techniques based on computationally intensive spatio-temporal models such as ConvLSTM, 3D CNNs,[14] or flow approaches (AdaCoF[18]). While the former improves network adaptability and stability, and the latter enhances frame fidelity, neither effectively addresses the joint challenge of real-time, bandwidth-constrained video streaming. High-quality interpolation methods are often computationally expensive and prone to artifacts such as blurring or occlusion, whereas ABR-centric approaches fail to improve visual quality when the bitrate is reduced. Previous interpolation systems used optical flow to create in-between frames by deforming adjacent inputs. Although effective to some degree, they were limited by motion artifacts and poor robustness in fast-moving or occluded regions. While these models have demonstrated improved performance compared to earlier kernel-based and optical-flow-based approaches, the

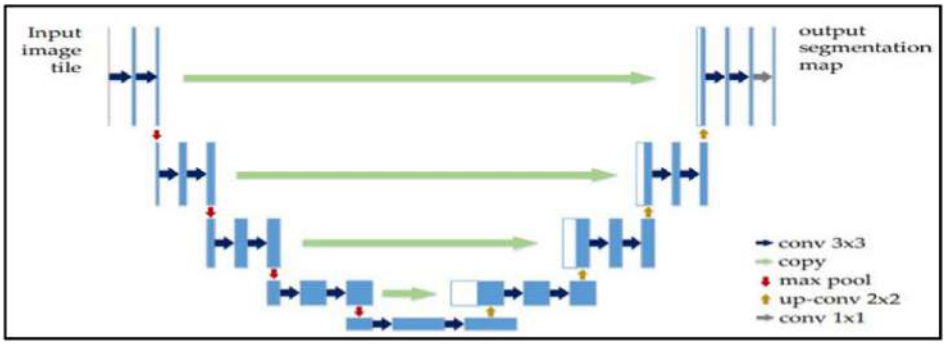


**Fig. 1.** The classic U-Net architecture.

results of current models may suffer when attempting to recreate detailed textures within an image. More recently, deep learning–based approaches, particularly CNN- and U-Net–derived architectures, have demonstrated superior performance by learning spatiotemporal features directly from data. Despite these advances, current models still face challenges in reconstructing fine-grained textures and ensuring robustness under bandwidth constraints. To mitigate these issues, we propose IncepU-Net, an enhanced U-Net variant that incorporates multi-scale inception modules, residual blocks, and attention-guided skip connections, trained with a joint Charbonnier + SSIM loss. In addition, it is well-suited for integration into bandwidth-constrained video streaming pipelines, thereby bridging the gap between network efficiency and perceptually meaningful frame synthesis.

## Materials and methods

Video streaming services dominate the Internet, generating significant revenue while competing to deliver high-quality content and superior QoE. Nonetheless, unstable network conditions tend to lead to frame loss and degradation, which adversely affect video playback. To overcome this, this work presents a neural network architecture that is based on the U-Net framework, which makes use of multi-scale feature extraction using inception blocks and attention gating to increase the relevance of features. This approach reconstructs missing frames with high precision, restoring temporal continuity and improving perceptual video quality.

Temporal frame interpolation generates temporary frames between successive frames in an attempt to increase video frame rates. Given two frames, $I_1$ and $I_3$, the goal is to estimate the missing frame $I_2$, effectively improving the frame rate and ensuring smoother playback. High-temporal resolution scenarios, such as slow-motion video and live streaming, can benefit from this method. In these cases, the model receives two sequential frames of data, which are then merged into a six-channel temporal sequence by first normalizing and augmenting the data to match the Vimeo90K dataset. The model's architecture is designed to process both encoding and decoding simultaneously. The model's architectural design is based on both an encoder and a decoder at the same time. This uses inception-based feature extraction and bottleneck stages with gated attention. Afterwards, it goes to a decoding stage with up-sampling of the decoder features along with skip connections. Finally, there is a convolutional layer for generating the final output frame. The Adamax optimizer was used during training and had a learning rate of 0.001. Two types of loss functions were used for training - one type being Charbonnier and another being SSIM. Both types were combined to generate high-quality, robust video frames.

### *U-Net*

U-Net, developed by Ronneberger et al.,[23] is among the first and most popular architectures to be used in Medical Image Segmentation. There are two primary components in the U-Net Architecture: the encoder, down-sampling of extracted features and a decoder that reconstructs them to a predetermined resolution,[24] as shown in Fig. 1.[23] A distinctive feature of U-Net is its integration of encoder-decoder layers with matching spatial dimensions via skip connections. The encoder is aimed at capturing shallow-level features, which still have detailed texture and contextual features, and the decoder processes more abstract features, though with lower spatial resolution.[25] These deep features are highly compressed, but they are useful in identifying the target accurately. U-Net improves its capability to generate high-quality and precise results by strategically combining the multi-scale characteristics of both components.[23]

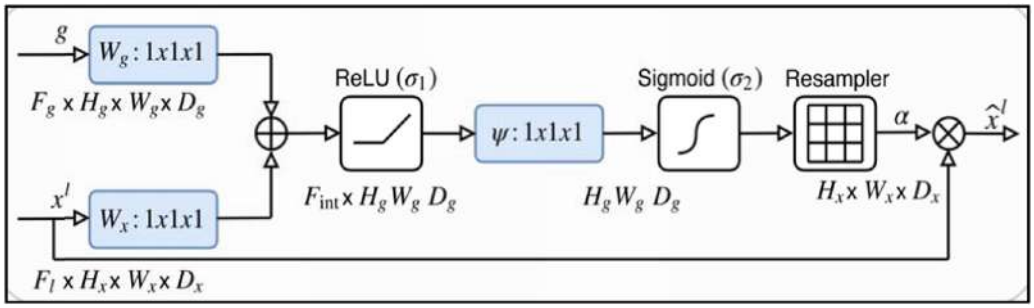


**Fig. 2.** Attention module.

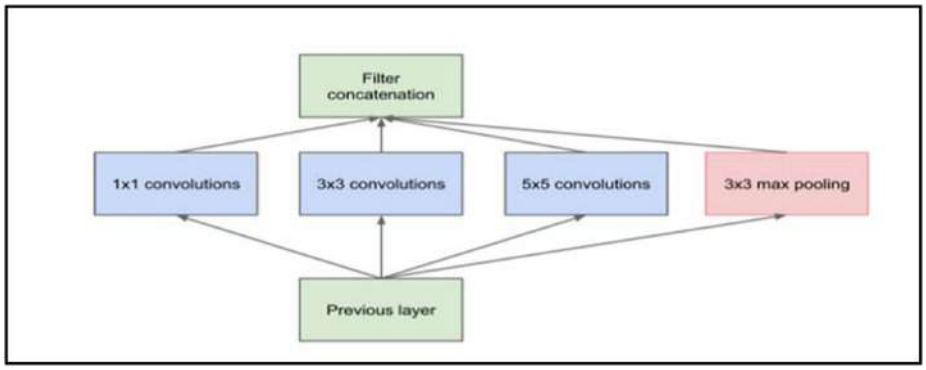


**Fig. 3.** Inception block.

### *Attention block*

Attention is an enhancement to deep learning models that helps direct focus to relevant information while eliminating the distraction of less important information. In computer vision, researchers have effectively applied attention to convolutional neural networks in order to act as a filter for selecting specific image features when performing multiple modalities, such as image captioning. Ultimately, this enhances the efficiency with which computing power is utilized, whether it be through processing sequential data or identifying spatial patterns.[26] The attention gates (AGs) merge input features ($x^l$) and a coarser gating signal (g) via bilinear up-sampling, additive fusion, and sigmoid-activated convolution to generate spatial attention coefficients ($\alpha$), which scale $x^l$ to emphasize key regions as shown in Fig. 2.[27] AGs improve medical image segmentation by filtering noise in skip connections without requiring additional localization models.[27] The elements integrate with U-Net to refine feature maps just before concatenation, thus directing the network to focus on important regions throughout both forward and backward passes. Through this implementation, U-Net achieves precision improvements and maintains computing performance levels.

### *Inception block*

The LeNet, AlexNet, and VGG architectures utilize different convolution sizes (e.g., 5 × 5, 3 × 3, 1 × 1), and researchers are trying to figure out which is best for a variety of data sets. On one hand, larger convolutions have the potential to have more representation; however, this comes with a higher number of parameters and greater computation. Smaller convolutions have faster execution times and require less memory, but may also result in decreased performance. The Google LeNet paper solves the trade-off problem by allowing users to combine different convolution sizes across multiple layers of blocks. This method achieves an optimal balance between fitness/accuracy, size/desirability, and time-to-completion by utilizing the Inception-style "going deeper" principle as demonstrated in inception.[28] The Inception block performs a combination of convolution and pooling operations internally to efficiently extract relevant features from images, as shown in Fig. 3.[28] The block sequentially applies 1 × 1, 3 × 3, and 5 × 5 convolutions, followed by 3 × 3 max pooling and a final 1 × 1 convolution. The final step involves a standalone 1 × 1 convolution. The approach strives to obtain features from different size ranges. The neural network

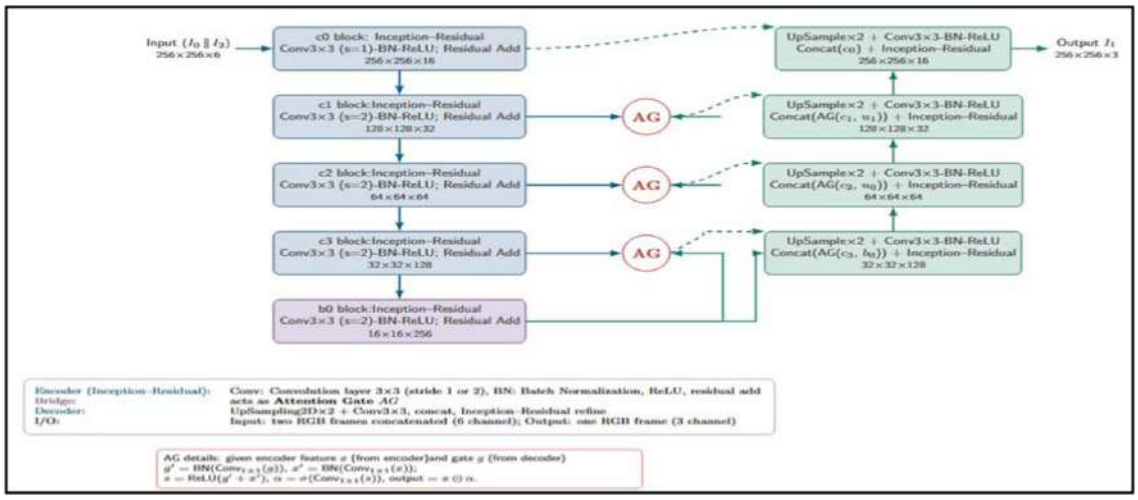


**Fig. 4.** Proposed network architecture.

ensures dimensional consistency through padding applications, where one-pixel padding is used for 3 × 3 convolution operations and two-pixel padding for 5 × 5 convolution operations. Within the block, concatenating the sequential operations preserves performance while maintaining functional diversity. The inception architecture leverages modular blocks that integrate parallel convolutional and pooling operations to optimize feature extraction and computational efficiency.[28,29] The Inception network model has undergone successive improvements that have led to better performance than the VGG and AlexNet models while maintaining improved classification accuracy.

## *Proposed network architecture*

The network synthesizes an intermediate frame from two temporally adjacent RGB inputs, making it particularly suitable for bandwidth-constrained video streaming scenarios. By integrating multi-scale feature extraction, attention-guided skip connections, and residual inception blocks, the proposed architecture improves perceptual quality and motion adaptively. The overall structure follows a classical encoder–bridge–decoder paradigm, as illustrated in Fig. 4.

### *Encoder*

The encoder is made out of a sequence of custom inception blocks that are augmented with residual connections. Every block serves as a main processing unit and is intended to facilitate multi-scale feature extraction by a combination of convolutional layers with different kernel sizes and strides. This architectural decision allows the network to absorb fine-grained and coarse patterns of space. In order to support gradient propagation and address the problem of vanishing gradients, every block has a residual connection, which is based on the ideas of Residual Networks (ResNets).[30] Inside each block, there are several sequential convolutional layers to increase the representational capacity with more complex and deeper feature interactions. Inception block structure is as follows and shown in Fig. 5:

1. The block starts with a traditional convolutional layer that has a 3 × 3 kernel size, a number of filters (which can be configured to 16), and then gradually gets larger as we descend in the network, followed by batch normalization and ReLU activation.
2. The next part of the model includes a configurable number (sequential) of convolutional layers. Each one is defined based on a user-definable filter size (i.e. 1 × 1, 3 × 3, etc.) as specified by the layer's parameter.
3. After all of these sequential convolutional layers have been processed, the original input is added to the final result as part of a residual connection. When the input and output channel dimensionality does not match, it is required to apply a 1 × 1 convolution to the input so it can be compatible with respect to the addition of the two signals. This residual mechanism facilitates more profound architectures and stabilizes training.

### *Bridge*

At the bottleneck, feature maps are further processed by an inception block that has the maximum filter capacity in order to capture global context. Attention gating is added to every skip connection

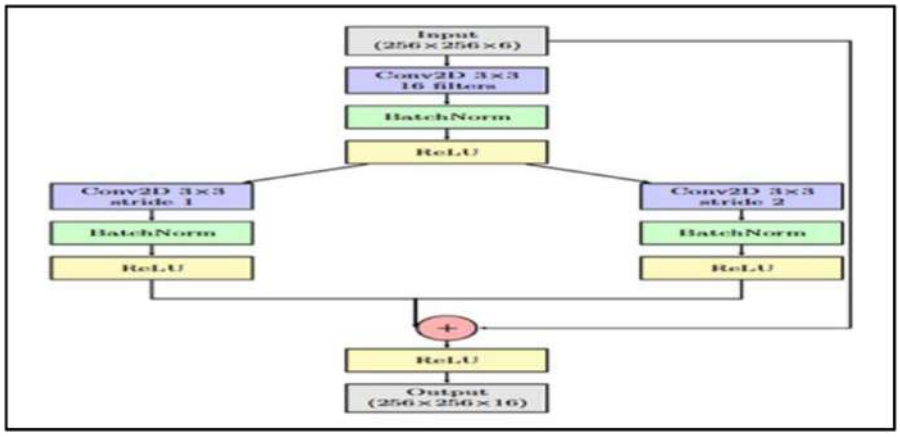


**Fig. 5.** Inception-residual block.

to enhance the combination of encoder and decoder representations. Each of the encoder feature map and the decoder gating signal is convoluted (1 × 1) in parallel to align them in the same feature space as illustrated in Fig. 4.

### *Decoder*

The decoder mirrors the architecture of the encoder. The decoder is designed to progressively increase spatial resolution while integrating Contextual features from the encoder through Attention-Guided skip connections. At each step of the decoder, there are three primary functions: (i) up-sampling, (ii) attention-based feature selection, (iii) multi-scale fusion through inception blocks. Each block of the decoder structure can be summarized as follows:

1. Every decoder stage starts with up-sampling the input feature map with UpSampling2D, then a Convolution2D to sharpen the up-sampled feature map. The use of batch normalization and ReLU activation stabilizes training and adds non-linearity.
2. The features of the modulated encoder are concatenated with the features of the up-sampled decoder. This combined tensor is subsequently fed into a tailor-made inception block. The inception block increases the richness of representation and promotes multi-scale decoding.

### *Hybrid loss formulation*

The network is trained with a mixed loss that involves Charbonnier loss as indicated in Eq. (1) with SSIM. SSIM is introduced to trade off pixel-level accuracy with perceptual quality. This loss function encourages the minimization of reconstruction error and maximum structural fidelity. This combination in the loss function guarantees that the result is not only correct with respect to pixel values, but also perceptually consistent.

- Charbonnier Robust Loss :

$$L_{\text{Charbonnier}}\ (y,\hat{y}) = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\sqrt{(y-\hat{y})^2+\varepsilon^2} \tag{1}$$

Where the variable y, $\hat{y}$ are the original and reconstructed images, respectively. While H and W are image dimensions. Finally, $\varepsilon = 0.1$, refers to preserving differentiability while mitigating outlier sensitivity in pixel-space reconstruction.

- Structural Similarity Index Measure (SSIM)

The SSIM is a metric that is employed in full-reference image quality assessment. SSIM quantifies visual similarity between an original and a processed version. It compares structural information to determine degradation by operations like compression or transmission. SSIM yields values within the interval [0,1], with scores approaching 1 reflecting a greater degree of structural correspondence between the compared images, and consequently, superior perceptual quality[31] as seen in Eq. (2).

$$L_{SSIM(y,\hat{y})} = \frac{(2\mu_y 2\mu_{\hat{y}})\,(2\sigma_{y\hat{y}}C_2)}{\left(\mu_y^2+\mu_{\hat{y}}^2+C_1\right)\left(\sigma_y^2+\sigma_{\hat{y}}^2+C_2\right)} \tag{2}$$

In the context of the SSIM, the parameters are defined as follows: $\mu_x$ and $\mu_\gamma$: Mean intensities of images $y$ and $\hat{y}$, respectively. $\sigma_y$ and $\sigma\hat{y}$: Standard deviations of images $y$ and $\hat{y}$, representing contrast measures. $\sigma_{y\hat{y}}$: Cross-covariance between images $y$ and $\hat{y}$.

The Hybrid Loss Formulation integrates these components with empirically determined weighting as shown in Eq. (3):

$$L_{Total} = L_{Charbonnier} + 0.5\, L_{SSIM} \quad (3)$$

## Experimental results and discussion

The suggested video frame interpolation approach is evaluated using the SSIM to ensure perceptual quality, Peak Signal-to-Noise Ratio (PSNR), and **Floating-Point Operations Per Second**.

### PSNR

It is a metric used to measure the quality of a reconstructed or compressed image. It evaluates the ratio between the maximum achievable signal power and the power of background noise that degrades the accuracy of the signal's representation. PSNR is commonly used in image processing and compression applications to evaluate the quality of reconstructed images, as seen in Eq. (4). A high PSNR value indicates good image quality.[4]

$$PSNR\,(x, y) = 10log_{10}\frac{max^2}{MSE} \quad (4)$$

Where max denotes the maximum pixel value in the frame, MSE (Mean Squared Error) measures the average squared difference between actual and predicted values, making it a standard metric for evaluating model accuracy as shown in Eq. (5).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \quad (5)$$

### Floating-point operations per second (FLOPS)

It quantifies the model's computational cost, as formally defined in Eq. (6).[32]

$$\boldsymbol{FLOPS} = \sum \boldsymbol{K} \times \boldsymbol{k} \times \boldsymbol{C_{out}} \times \boldsymbol{C_{in}} \times \boldsymbol{W} \times \boldsymbol{H} \quad (6)$$

Here, K denotes the convolution kernel size, $C_{in}$ and $C_{out}$ represent the input and output channel dimensions, respectively, and H × W specifies the spatial resolution of the resulting feature map.

### Dataset

In this study, frame interpolation performance is evaluated using multiple benchmark datasets. Specifically, the UCF101 and Vimeo90K datasets are employed.[22] The training UCF101 dataset contains videos with a resolution of 320 × 240 for each frame. UCF101 dataset videos are characterized by complex backgrounds and dynamic camera motions, making it well-suited for testing interpolation methods. On the other hand, to maintain independence from the training data, 3782 triplets from the Vimeo90K dataset are also used as an evaluation set. The quality of interpolated frames was assessed using both objective and subjective metrics. For visual inspection the subjective evaluation has been used, whereas objective assessment employed PSNR and SSIM as key performance indicators.

### Training parameter

The model uses the Adamax Optimizer that is an adaptation of the Adam algorithm. The use of this type of optimizer has been shown to be effective in models which have sparse gradients as is typical when embedding layers are utilized. An initial learning rate of 0.001 was selected and we employed gradient clipping with a value of 0.5 to help ensure that gradients do not grow too large. Each model was trained for 50 iterations at a batch size of 8 to enable consistent computation among all tasks. In addition, to allow for task specific convergence characteristics while maintaining sufficient optimization ability for each task, the learning rates for each task were reduced by a value of $10^{-6}$. We used a combined Charbonnier-SSIM loss to optimize both pixel-level accuracy and perceptual quality. Input frames were randomly flipped left-right during training with a 50% chance.

### Ablation study

Ablation studies are one of the evaluation metrics used to assess how architectural changes affect model performance. In this work, four progressively modified variants of the U-Net architecture are analysed to determine their effectiveness in enhancing video frame interpolation accuracy. Each variant introduces incremental changes aimed at improving the model's capacity to capture motion dynamics and preserve fine details. Our aim is to find the architecture that produces the most accurate interpolated frames. Ablation testing was performed using a Vimeo90k dataset, which has become an industry-wide benchmark used to evaluate both the quality and robustness of frame interpolation models.

- **Attention U-Net (AttU-Net):** The AttU-Net is an extension of the classical U-Net architecture; it extends the traditional skip connections found in the U-Net with attention modules. The use of these attention modules enables the network to

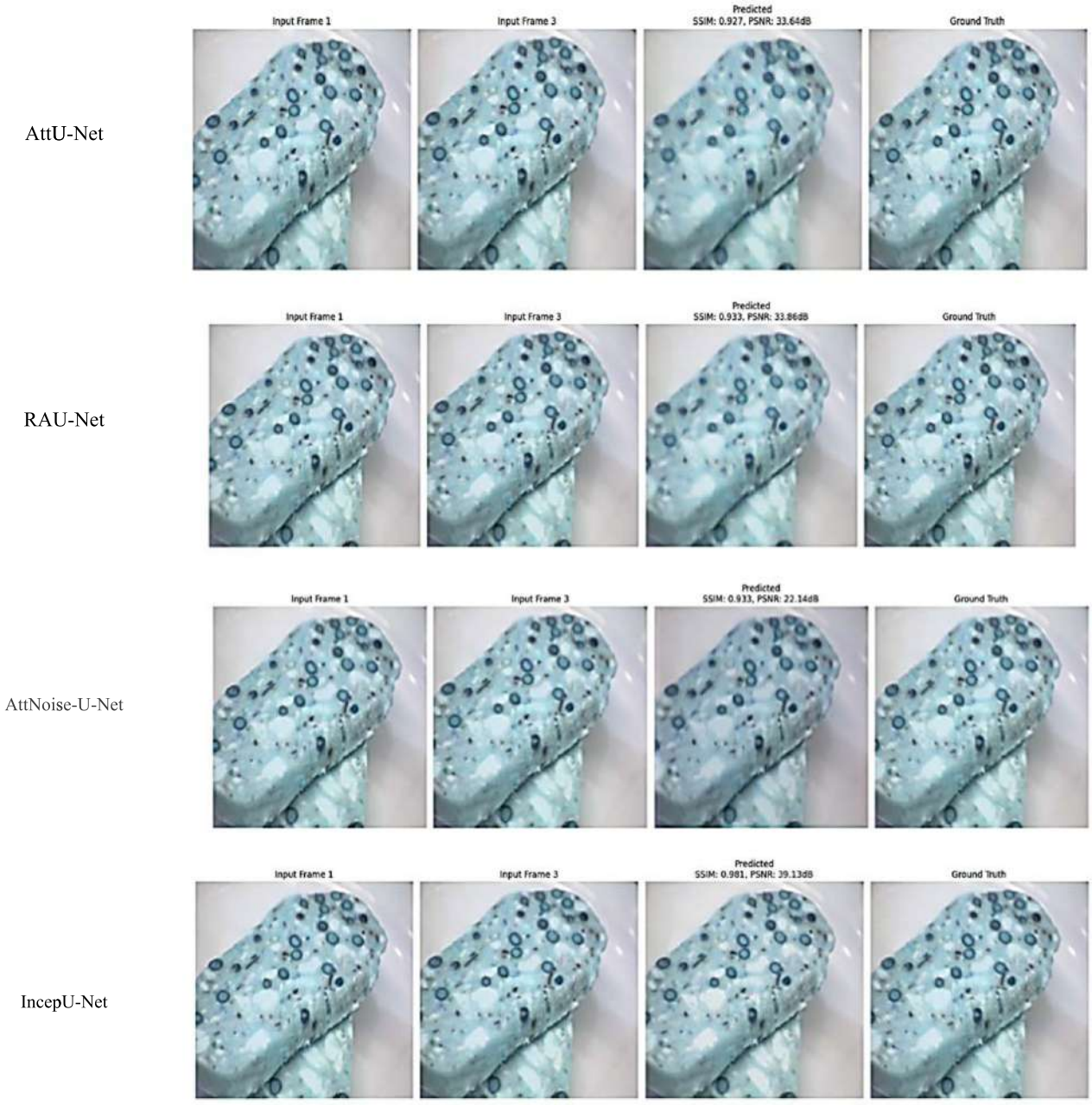


**Fig. 6.** Predicted frames using different U-Net architectures.

selectively focus on the key feature information related to motion. This occurs when combining information from the encoder and decoder. As such, the network is capable of producing higher-quality, more visually consistent results than previous architectures, as shown in Fig. 6 (first row).

- **Residual Attention U-Net (RAU-Net):** The Residual Attention U-Net enhances the traditional U-Net with both residual blocks and attention gates. The inclusion of residual blocks improves the gradient flow throughout the network and enables the deeper representation learning of features. Furthermore, the incorporation of attention gates allows for selective integration of the informative features of the encoder into the decoder portion of the network. As a result of this combination, the interpolation accuracy of the model is enhanced, while its ability to represent complex motion is also enhanced, as shown in Fig. 6 (Second row).
- **Attention U-Net architecture with Gaussian Noise (AttNoise-U-Net).** This architecture introduces an improved version of U-Net for

video frame interpolation, combining attention mechanisms with selectively noisy skip connections. In the encoder, a series of convolutional layers extract features at multiple levels, while carefully added Gaussian noise in certain skip connections helps the model become more robust. During decoding, attention gates determine how noisy encoder features are combined with the upsampled decoder outputs. During the decoding process, attention gates guide how the noisy encoder features are merged with the up sampled decoder outputs, allowing the network to better focus on important spatial details and produce more precise, more reliable frame predictions, as illustrated in the third row of Fig. 6.

- **Inception modules within the U-Net architecture (IncepU-Net):** In **IncepU-Net**, an enhanced U-Net-based architecture for video frame interpolation that incorporates attention gates and inception blocks. Using an encoder-bridge decoder architecture, the model improves motion detail and spatial accuracy by utilizing attention-guided skip connections and multi-scale feature extraction.

The efficacy of four U-Net variant architectures (AttU-Net, RAU-Net, AttNoise-U-Net, and IncepU-Net) in image interpolation tasks was assessed through a systematic evaluation of their performance. The PSNR and SSIM results from Table 1 demonstrate the outcome of the comparative ablation study. These results reflect a comparison between the methodological implications of using each architecture. The superior performance of IncepU-Net can be attributed to its integration of multi-scale inception modules. The Inception Block allows for significant improvement in hierarchical feature extraction. In contrast, architectures utilizing only attention-based mechanisms provided weaker results than those that utilized inception modules. Therefore, from what has been illustrated, it appears that attention alone will be insufficient for balancing a global context with local refinements in this particular task.

Although AttNoise-U-Net was originally designed to cope with noisy inputs, it ended up delivering the weakest results. This points to the possibility that its attention mechanism is actually enhancing noise rather than reducing it. It is apparent from the third row of Fig. 6 that while the generated frame maintains a high level of structure and spatial detail relative to the reference frame, the colors were incorrectly reproduced. These findings are supported by both the SSIM and PSNR measures, where high SSIM values indicate good correspondence; however, low PSNR values indicate decreased perceived quality. This finding makes a strong case for attention strategies that can adapt to noisy conditions, and it also shows why combining features at multiple scales can work better than relying only on local attention in image reconstruction. It appears that the attention mechanism used by AttU-Net is not able to capture the changing dynamics over time. However, since the SSIM (0.927) and PSNR (33.64 dB) of AttU-Net were competitive with those of RAU-Net, RAU-Net appears to provide a better solution than AttU-Net in some cases.

**Table 1.** The comparison of the objective quality for different U-Net architectures.

| Architecture | PSNR | SSIM |
|---|---|---|
| AttU-Net | 30.10 dB | 0.873 |
| RAU-Net | 29.86 dB | 0.869 |
| AttNoise-U-Net | 28.67 dB | 0.821 |
| IncepU-Net | 35.07 dB | 0.956 |

### *Comparison experiments*

Table 2 shows the comparison result of the proposed video frame interpolation method to other state-of-the-art architectures such as AdaCoF, ToFlow, EDSC, and U-NET + RES, across the UCF101 and Vimeo90K datasets. While existing methods reveal strong numerical accuracy, the proposed architecture demonstrates enhanced perceptual quality and structural fidelity. On UCF-101, which has many examples of motion, we tested our model's ability to keep spatial coherency when dealing with difficult motion scenarios. Although our method achieves a slightly lower PSNR equal 31.80 compared to 32.44 and 32.49 for AdaCoF and U-NET + RES respectively, this does not necessarily imply inferior visual quality, as PSNR does not fully capture structural or perceptual aspects. However, a lower PSNR does not always mean the image will be of lesser quality. In order to quantify the degree to which an image preserves the luminance, texture, and spatial structure of an original image, SSIM can be used. The SSIM is 0.933, 0.928 for proposed work and AdaCoF respectively. This enhancement reflects how we preserve more of the important visual features than AdaCoF. In addition to employing PSNR and SSIM, we incorporate the Learned Perceptual Image Patch Similarity (LPIPS) metric.[33] The LPIPS metric represents the degree of perceptual dissimilarity as measured by deep feature space embedding; thus, smaller values are indicative of greater perceptual similarity to the original image. Although our average LPIPS value was slightly greater than other comparison models, suggesting that fine texture realism can still be further improved, the results collectively show that the model emphasizes maintaining meaningful structural details over

**Table 2.** Comparison result (the best results are in bold).

| Methods | UCF101 | | | Vimeo90K | | | Parameter (million) | Training epochs | FLOP | Training Dataset |
|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR | SSIM | LPIPS | PSNR | SSIM | LPIPS | | | | |
| ToFlow [22] | 28.494† | 0.915† | 0.027* | 33.18† | 0.945† | 0.027* | **1.1*** | 15 | 44.08† | **Vimeo90K** |
| EDSC [21] | 31.940† | 0.924† | **0.023*** | 34.49† | 0.958 | **0.026*** | 8.9* | 120* | 17.96 | **Vimeo90K*** |
| Adacof [18] | 32.439† | 0.928† | 0.029* | 34.34† | 0.956† | 0.031* | 21.84* | 50 | 24.95† | **Vimeo90K*** |
| U-NET + RES [19] | **32.486**† | 0.928† | —- | 34.66† | **0.960**† | —- | 33.63† | 50 | 45.04† | **Vimeo90K**† |
| Our | 31.80 | **0.933** | 0.084 | **35.19** | 0.958 | 0.043 | 5.3 | 50 | **17.16** | **Vimeo90K** |

* the related work result copied from, [21] † the related work result copied from [19]

achieving absolute pixel-level similarity, a trade-off that is often more aligned with human perceptual preferences in dynamic motion scenarios.

The evaluation of our method was performed on Vimeo90K. Vimeo90K has more consistent and temporally stable sequences, so as to evaluate how well our method can reconstruct accurate and visually pleasing images in smoother motion scenes. Our model achieved the highest PSNR value equal to 35.19 for this dataset, representing the best possible pixel-level reconstruction, and had an SSIM score reaching 0.958, which was very similar to the best performing U-Net + RES (0.960). On the other side, LPIPS reaches 0.043, slightly higher than the best competing methods. This result highlights the architecture's ability to capture both spatial detail and temporal consistency, even in less complex motion scenarios. The methods show a clear trade-off between complexity and performance: ToFlow is lightweight (1.1M parameters, 44.08 FLOPs, 15 epochs) with moderate PSNR/SSIM; EDSC (8.9M parameters, 17.96 FLOPs, 120 epochs) improves accuracy efficiently; Adacof and U-NET + RES are heavy (21.8, 33.6M parameters, 24.95,45.04 FLOPs, 50 epochs) with higher PSNR/SSIM; our model is lightweight (5.3M, 17.16 FLOPs, 50 epochs) yet achieves competitive performance.

### *Subjective evaluation of QoE*

Subjective evaluation provides validation of the improvement in perceived video quality due to improved frame interpolation techniques by relating improved technical parameters (PSNR & SSIM) to users' perception of quality. Combining subjective evaluation with objective evaluation methods allows for an enhanced and more complete measure of QoE, which will guide the development of video streaming applications that are optimized for users' preferences.

The mean opinion score (MOS) enables a link to be made between objective human perceptions of Quality of Experience and direct human feedback that can be directly related to the existing technical measures of overall system performance. MOS is considered an important tool for measuring contextual factors that automated systems miss. In this context, MOS is measured as the mean rating provided by each evaluator of the output produced via the differing architectures of U-Nets as outlined in Eq. (7).[34] By linking the technical performance metrics used in evaluating the structure of deep learning models to the subjective experience (i.e., QoE) produced by those structures, it is possible to evaluate how specific structural decisions (i.e., attention modules, skip connections) affect both the user's perception of visual plausibility and temporal coherence.

$$MOS = \frac{\sum_{i=1}^{N} O_{ik}}{N} \tag{7}$$

Where $O_{ik}$ is participant i's opinion score for frame interpolation method k.

To validate the perceptual quality of interpolated frames, a subjective evaluation was conducted with 10 participants (5 experts in video processing and five non-experts). Participants were shown 10 video sequences interpolated by each method (IncepU-Net, AttU-Net, RAU-Net, and AttNoise-U-Net) and asked to rate the visual quality on a scale of **1 (Poor)** to **5 (Excellent)** as illustrated in Table 3.[35,36]

**Table 3.** MOS evaluation.

| MOS Score | Quality Rating | Perceived Impairment Level |
|---|---|---|
| 1 | Bad | Very annoying |
| 2 | Poor | Annoying |
| 3 | Fair | Mildly annoying |
| 4 | Good | Unnoticeable but not annoying |
| 5 | Excellent | Unnoticeable |

The MOS values in Table 4 show that the proposed IncepU-Net architecture produced the best rating for the users (MOS = 4.6) compared to all other baseline architectures tested based on their ability to produce higher perceived image quality. Users continually stated that they felt the Incep U-Net generated smoother motion transitions and fewer visual artifacts than the other architectures tested, which correlates well with the superior objective performance measured by PSNR and SSIM. The IncepU-Net also had a very low standard deviation

**Table 4.** Subjective evaluation using MOS (1–5 scale).

| Architecture | MOS | SD |
|---|---|---|
| AttU-Net | 3.8 | 0.56 |
| RAU-Net | 3.5 | 0.58 |
| AttNoise-U-Net | 2.9 | 0.47 |
| IncepU-Net | **4.6** | **0.52** |

(SD = 0.52) which shows a strong agreement in how each rater rated the images generated by the IncepU-Net. Thus, it can be said that this architecture was effective at producing visually consistent interpolated frame images. Conversely, the AttNoise-U-Net was given the lowest MOS (2.9), although it did have the smallest rating spread (SD = 0.47). It appears from these results that the output of the AttNoise-U-Net was universally regarded as being of lower quality by nearly all raters with little variance among them. Although RAU-Net has a somewhat intermediate MOS (3.5), it had the largest rating spread (SD = 0.58) which probably means that users could perceive the images differently because the method did not handle either motion or texture equally consistently. When assessing the overall quality of images through MOS, it is also important to report the standard deviation of MOS to assess both the amount of variability in ratings provided by different raters and to determine whether such variability is statistically reliable. The average SD for all architectures evaluated is about 0.53, indicating moderate variability in ratings. Further, the IncepU-Net not only achieved the highest MOS but also generated the least variable high-quality MOS ratings; i.e., about one-third of raters rated it as having been within ±0.5 of the mean MOS, a meaningful range for perceiving differences in quality.

### *Limitations*

Although the IncepU-Net architecture has shown strong results in terms of objective and perceptual measures, many other constraints should be taken into account:

1. Fixed Resolution and Dataset Bias: All experiments were run with a constant resolution for both the UCF101 and Vimeo90K datasets. Although the two datasets are commonly referenced benchmark datasets, they do not cover all types of resolution variance or frame rate variance, as well as motion that is found in real-world video. As such, it remains to be determined how the model's performance will generalize across different operational environments.
2. Perceptual Texture Fidelity in High-Motion Scenes: While the model shows superior structural coherence in terms of SSIM on dynamic content, its LPIPS values are slightly worse than those of competitors (especially UCF101). This indicates that while structural coherence is preserved, fine texture realism under rapid motion remains an area for improvement.
3. Computational Trade-offs in Deployment: Even though the model has lower parameters and FLOPs than many other models, there is no comprehensive analysis of the model's real time processing performance and memory footprint for limited resources and/or edge devices. Therefore, such analyses need to be done to verify if this model can be used in applications with low-latency requirements.

## Conclusion

The paper introduced a model for video frame interpolation called IncepU-Net, which is introduced as an improved version of U-Net. In IncepU-Net, we combine inception modules and guided skip connections to maintain high levels of QoE, particularly under fluctuating network conditions that cause frame drops and visual artifacts. From a theoretical perspective, the work advances encoder–decoder design by demonstrating how multi-scale inception processing and attention mechanisms can jointly capture fine spatial structures and long-range temporal dependencies. Furthermore, using a single Charbonnier-SSIM loss function effectively provided the ability to optimize both pixel-level accuracy and perceptual quality, providing consistent gains across PSNR, SSIM, and LPIPS metrics for both UCF101 and Vimeo90K datasets. We present a method that produces a tangible, practical solution. That is, it produces realistic images with high structural coherence when restoring lost frames and thus provides smooth and stable video playback over many differing contents. At the same time, it is computationally inexpensive, having only 5.3 million parameters and 17.16 FLOPs. Although a balance between quality and efficiency is well-suited to bandwidth-limited or resource-constrained environments, there are certain limitations that must also be acknowledged. Although this method is well-suited to preserve structure within a scene; however, it also has a slightly larger LPIPS in high-motion scenes, showing that achieving realistic textures may still need additional development. Future research could therefore focus on developing hybrid motion aware architectures that combine explicit motion estimation with attention mechanisms to better align temporally and reduce

perceived distortions in fast motion. Additionally, self-supervised methods for learning temporal consistency from unlabeled data could be used to provide greater robustness across longer video sequences than would otherwise be possible using densely labeled ground truth.

## Authors' declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images that are not ours have been included with the necessary permission for republication, which is attached to the manuscript.
- No animal studies are present in the manuscript.
- No human studies are present in the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at University of Babylon.

## Authors' contributions statement

K. A. contributed to the methodology, formal analysis, and writing of the original draft. D. A. M. contributed to writing, review, and editing. I. H. A. was responsible for the conceptualization and supervision of the study. S. A. A. contributed to visualization and data creation.

## Data availability

All the data used within this research paper is publically available. The Vimeo90k dataset is accessible at http://toflow.csail.mit.edu/ and the UCF-101 dataset is located at https://www.crcv.ucf.edu/data/UCF101.php.

## References


1. Ghani RF, Ajrash AS. Quality of experience measurement for video streaming based on adaptive neural fuzzy inference system. Iraqi J Sci. 2019 Jul 19;60(7):1609–17. https://doi.org/10.24996/ijs.2019.60.7.21.
2. Liborio Filho J da M, Oliveira J, Melo CAV. Super-resolution with perceptual quality for improved live streaming delivery on edge computing. Comput Networks. 2024 Jun;248:110463. https://doi.org/10.1016/j.comnet.2024.110463.
3. Abbood IK, Idrees AK. Distributed video transmission reduction approach for energy saving in WMSNs. Int J Commun Syst [Internet]. 2024 Oct 24;37(15):e5880. https://doi.org/10.1002/dac.5880.
4. Taha M, Ali A. Smart algorithm in wireless networks for video streaming based on adaptive quantization. Concurrency and Computation: Practice and Experience. 2023 Jan 19;35(9). https://doi.org/10.1002/cpe.7633.
5. YAMAZAKI T. Quality of experience (QoE) studies: Present state and future prospect. IEICE Trans Commun. 2021 Jul 1;104(7):716–724. https://doi.org/10.1587/transcom.2020cqi0003.
6. Zinner T, Hohlfeld O, Abboud O, Hossfeld T. Impact of frame rate and resolution on objective QoE metrics. 2010 2nd International Workshop on Quality of Multimedia Experience, QoMEX 2010 - Proceedings. 2010:29–34. https://doi.org/10.1109/qomex.2010.5518277.
7. Saha A, Pentapati SK, Shang Z, Pahwa R, Chen B, Gedik HE, *et al.* Perceptual video quality assessment: The journey continues!. Frontiers in Signal Processing. 2023 Jun 27;3:1193523. https://doi.org/10.3389/frsip.2023.1193523.
8. Zhou W, Min X, Li H, Jiang Q. A brief survey on adaptive video streaming quality assessment. J Vis Commun Image Represent. 2022 Jul;86:103526. https://doi.org/10.1016/j.jvcir.2022.103526.
9. Kua J, Armitage G, Branch P. A survey of rate adaptation techniques for dynamic adaptive streaming over HTTP. IEEE Commun Surv Tutorials. 2017;19(3):1842–66. https://doi.org/10.1109/comst.2017.2685630.
10. Spang B, Kunamalla S, Teixeira R, Huang T-Y, Armitage G, Johari R, *et al.* Sammy: Smoothing video traffic to be a friendly internet neighbor. In: Proceedings of the ACM SIGCOMM 2023 Conference. New York, NY, USA: ACM; 2023:754–68. https://doi.org/10.1145/3603269.3604839.
11. Darwich M, Bayoumi M. Video quality adaptation using CNN and RNN models for cost-effective and scalable video streaming Services. Cluster Comput. 2024 Aug 28;27(5):6355–75. https://doi.org/10.1007/s10586-024-04315-8.
12. Eswara N, Ashique S, Panchbhai A, Chakraborty S, Sethuram HP, Kuchi K, *et al.* Streaming video QoE modeling and prediction: A long short-term memory approach. IEEE Trans Circuits Syst Video Technol. 2020 Mar;30(3):661–73. https://doi.org/10.1109/tcsvt.2019.2895223.
13. Desai P, Sujatha C, Chakraborty S, Ansuman S, Bhandari S, Kardiguddi S. Next frame prediction using ConvLSTM. J Phys Conf Ser. 2022 Jan 1;2161(1):012024. https://doi.org/10.1088/1742-6596/2161/1/012024.
14. Aravinda C V., Al-Shehari T, Alsadhan NA, Shetty S, Padmajadevi G, Reddy KRUK. A novel hybrid architecture for video frame prediction: combining convolutional LSTM and 3D CNN. J Real-Time Image Process. 2025 Feb 23; 22(1):50. https://doi.org/10.1007/s11554-025-01626-w.
15. Jatana N, Wadhwa D, Singh NK, Hassen OA, Gupta C, Darwish SM, *et al.* Future frame prediction using generative adversarial networks. Karbala Int J Mod Sci. 2024 Jan 12;10(1):19–30. https://doi.org/10.33640/2405-609X.3338.
16. Ghosh M, Singhal C. M-3R: A memory based approach for streaming QoE prediction under 3R settings. In: 2021 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS). IEEE; 2021:432–7. https://doi.org/10.1109/ants52808.2021.9936944.
17. Azadegan H, Shirazi AAB. Improving video quality by predicting inter-frame residuals based on an additive 3D-CNN model. Journal of Visual Communication and Image Representation. 2023 Feb 1;90:103734. https://doi.org/10.2139/ssrn.4088133.
18. Lee H, Kim T, Chung T, Pak D, Ban Y, Lee S. AdaCoF: Adaptive collaboration of flows for video frame interpolation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2020:5315–24. https://doi.org/10.1109/cvpr42600.2020.00536.

19. Yang X, Zhang H, Qu Z, Feng Z, Tian J. Video frame interpolation via residual blocks and feature pyramid networks. IET Image Process. 2023 Mar 23;17(4):1060–70. https://doi.org/10.1049/ipr2.12695.
20. Itagi S, Gowda S, Udupa T, Shylaja SS. Future frame prediction using deep learning. In: Lecture Notes in Electrical Engineering. 2022:187–99. https://doi.org/10.1007/978-981-16-6448-9_21.
21. Cheng X, Chen Z. Multiple video frame interpolation via enhanced deformable separable convolution. IEEE Trans Pattern Anal Mach Intell. 2022 Oct 1;44(10):7029–45. https://doi.org/10.1109/tpami.2021.3100714.
22. Xue T, Chen B, Wu J, Wei D, Freeman WT. Video enhancement with task-oriented flow. Int J Comput Vis. 2019 Aug 12;127(8):1106–25. https://doi.org/10.1007/s11263-018-01144-2.
23. Yang D, Liu G, Ren M, Xu B, Wang J. A multi-scale feature fusion method based on U-net for retinal vessel segmentation. Entropy. 2020 Jul 24;22(8):811. https://doi.org/10.3390/e22080811.
24. Xu Q, Ma Z, HE N, Duan W. DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation. Comput Biol Med. 2023 Mar;154:106626. https://doi.org/10.1016/j.compbiomed.2023.106626.
25. Ibtehaz N, Kihara D. ACC-UNet: A completely convolutional UNet model for the 2020s. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2023:692–702. https://doi.org/10.1007/978-3-031-43898-1_66.
26. Cha J, Jeong J. Improved U-net with residual attention block for mixed-defect wafer maps. Appl Sci. 2022 Feb 20;12(4):2209. https://doi.org/10.3390/app12042209.
27. Chen Z, Peng Y, Jiao J, Wang A, Wang L, Lin W, *et al.* MD-Unet for tobacco leaf disease spot segmentation based on multi-scale residual dilated convolutions. Sci Rep. 2025 Jan 22;15(1):2759. https://doi.org/10.3390/app12042209.
28. Chai J, Zeng H, Li A, Ngai EWT. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. Mach Learn with Appl. 2021 Dec;6:100134. https://doi.org/10.1016/j.mlwa.2021.100134.
29. Zhu Z, Xu Z, Liu J. Retracted: Application of lightweight deep learning model in vocal music education in higher institutions. Comput Intell Neurosci. 2023 Jan 11;2023(1). https://doi.org/10.1155/2023/9781291.
30. Shafiq M, Gu Z. Deep residual learning for image recognition: A survey. Applied sciences. 2022 Sep 7;12(18):8972. https://doi.org/10.3390/app12188972.
31. Bakurov I, Buzzelli M, Schettini R, Castelli M, Vanneschi L. Structural similarity index (SSIM) revisited: A data-driven approach. Expert Syst Appl. 2022 Mar 1;189:116087. https://doi.org/10.1016/j.eswa.2021.116087.
32. Pan P, Shao M, He P, Hu L, Zhao S, Huang L, *et al.* Lightweight cotton diseases real-time detection model for resource-constrained devices in natural environments. Front Plant Sci. 2024;15(June):1–15. https://doi.org/10.3389/fpls.2024.1383863.
33. Hernández-Cámara P, Vila-Tomás J, Laparra V, Malo J. Dissecting the effectiveness of deep features as a perceptual metric. 2023. https://doi.org/10.2139/ssrn.4609207.
34. Yu H, Benois-Pineau J, Bourqui R, Giot R, Zhukov A. Mean opinion score as a new metric for user-evaluation of XAI methods. In 2025:443–57. https://doi.org/10.1007/978-3-031-88223-4_31.
35. Laghari AA, Zhang X, Shaikh ZA, Khan A, Estrela V V., Izadi S. A review on quality of experience (QoE) in cloud computing. J Reliab Intell Environ. 2024 Jun 25;10(2):107–21. https://doi.org/10.1007/978-981-10-7200-0_15.
36. Abdul Hamid MF, Jofri MH, Meerangani KA, Sharipp MTM, Ariffin MFBM, Ayub MS, *et al.* Integrated management model of malaysian private tahfiz centers for user acceptance satisfaction by using linear regression. Baghdad Sci J. 2024 Aug 19;22(1). https://doi.org/10.21123/bsj.2024.9560.

# تحسين جودة تجربة المستخدم (QOE) في بث الفيديو بالاعتماد على نموذج Attention U-Net

**ايمان كاظم عبود، دعاء عايد محمد ، اسراء هادي علي ، ساره عبد الرضا عبد**

قسم البرمجيات، كلية تكنولوجيا المعلومات، جامعة بابل، بابل، العراق.

**الخلاصة**

تُشكّل خدمات بث الفيديو، بما فيها يوتيوب ونتفليكس وأمازون، نسبةً كبيرةً من حركة الإنترنت العالمية. لذا، يُعدّ الحفاظ على جودة صورة عالية أمرًا بالغ الأهمية لرضا المستخدمين وجودة تجربتهم الشاملة. مع ذلك، غالبًا ما يتسبب تذبذب عرض النطاق الترددي في فقدان الإطارات وانقطاع الحركة، ما يؤدي إلى تشوهات بصرية وتدهور جودة التجربة. ورغم محاولات طرق استكمال إطارات الفيديو الحالية معالجة هذه التحديات، إلا أن معظمها لا يزال محدودًا في قدرته على الحفاظ على الأداء في بيئات ذات عرض نطاق ترددي محدود. وللتغلب على هذه القيود، طُرح نموذج IncepU-Net تم اقتراح IncepU-Net كبنية U-Net محسنة تدمج وحدات inception متعددة المقاييس ووصلات تخطي ببوابة الانتباه من أجل استيفاء إطارات الفيديو الذكي في البيئات ذات النطاق الترددي المحدود. وقد تم تدريب IncepU-Net وتقييمه على مجموعتي بيانات Vimeo90K و UCF101. حقق النموذج في مرحلة الاختبار نسبة إشارة إلى ضوضاء (PSNR) بلغت 35.19 ديسيبل ومؤشر تشابه هيكلي (SSIM) بلغ 0.958 باستخدام Vimeo90K، ونسبة إشارة إلى ضوضاء بلغت 30.08 ديسيبل ومؤشر تشابه هيكلي بلغ 0.933 باستخدام UCF101. يتفوق نموذج IncepU-Net على العديد من أحدث طرق الاستيفاء. تُظهر هذه النتائج أن IncepU-Net يستوفي الإطارات المفقودة بكفاءة، ويُخفف من تقطع التشغيل، ويحافظ على استمرارية الحركة. مما يجعله حلاً قوياً وعملياً لتحسين جودة تجربة المستخدم في تطبيقات بث الفيديو الواقعية، ويسد الفجوة بين محدودية نقل البيانات عبر الشبكة وتوليف الإطارات ذات الدلالة الإدراكية.

**الكلمات المفتاحية:** آليات الانتباه، استيفاء الإطارات، وحدات Inception، جودة تجربة المستخدم(QoE)، جودة الخدمة(QoS)، بنية U-Net.