The following publication Z. Huang, Y. -L. Chan, N. -W. Kwong, S. -H. Tsang, K. -M. Lam and W. -K. Ling, "Long Short-Term Fusion by Multi-Scale Distillation for Screen Content Video Quality Enhancement," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 35, no. 8, pp. 7762-7777, Aug. 2025 is available at https://doi.org/10.1109/TCSVT.2025.3544314.

# Long Short-term Fusion by Multi-scale Distillation for Screen Content Video Quality Enhancement

Ziyin Huang, Yui-Lam Chan, *Member, IEEE,* Ngai-Wing Kwong, Sik-Ho Tsang, Kin-Man Lam, *Senior Member, IEEE,* and Wing-Kuen Ling, *Senior Member, IEEE*

***Abstract*—Different from natural videos, where artifacts distributed evenly, the artifacts of compressed screen content videos mainly occur in the edge areas. Besides, these videos often exhibit abrupt scene switches, resulting in noticeable distortions in video reconstruction. Existing multiple-frame models using a fixed range of neighbor frames face challenges in effectively enhancing frames during scene switches and lack efficiency in reconstructing high-frequency details. To address these limitations, we propose a novel method that effectively handles scene switches and reconstructs high-frequency information. In the feature extraction part, we develop long-term and short-term feature extraction streams, in which the long-term feature extraction stream learns the contextual information, and the short-term feature extraction stream extracts more related information from shorter input to assist the long-term stream to handle fast motion and scene switches. To further enhance the frame quality during scene switches, we incorporate a similarity-based neighbor frame selector before feeding frames into the short-term stream. This selector identifies relevant neighbor frames, aiding in the efficient handling of scene switches. To dynamically fuse the short-term feature and long-term features, the muti-scale feature distillation focuses on adaptively recalibrating channel-wise feature responses to achieve effective feature distillation. In the reconstruction part, a high-frequency reconstruction block is proposed for guiding the model to restore the high-frequency components. Experimental results demonstrate the significant advancements achieved by our proposed Long Short-term Fusion by Multi-Scale Distillation (LSFMD) method in enhancing the quality of compressed screen content videos, surpassing the current state-of-the-art methods.**

***Index Terms*—Screen content video, quality enhancement, deep learning.**

## I. Introduction

With the development of intelligent terminal, screen content videos have received increasing attention such as the cloud gaming, video conference, online education, etc. The spread of the COVID-19 in 2020 has led to a surge in demand for online education and virtual conferences, making Screen Content Coding (SCC) [1] [2] essential for effective screen sharing. Consequently, enhancing the quality of screen content videos has become a critical challenge.

Unlike natural videos, which typically feature a dynamic range of colors, screen content videos exhibit distinct characteristics in both the spatial and temporal domains. In the spatial domain, these videos often contain large uniform, and flat areas with minimal textural complexity, as well as repeated patterns found in graphical user interfaces, spreadsheets, or web pages. Moreover, the color palette in screen content videos tends to be more limited compared to that of natural videos. This is because screen content is often sourced from digital sources that use a specific set of colors for icons, text, and simple graphics. By making use of these screen content characteristics, SCC [1] was proposed as an extension of HEVC [3] to increase coding efficiency. In addition to the conventional HEVC intra (INTRA) mode [4], the SCC standard adopts two dedicated coding modes, Intra Block Copy (IBC) and palette (PLT) [5]. IBC [6] uses the reconstructed block from the current frame as the prediction block, while PLT enumerates a color value for each coding block to generate a color table and assigns an index to each sample to indicate its corresponding color in the color table. These tools are beneficial for flat areas and repeated regions. As a result, the low-frequency region can be reconstructed well, and the artifact is mainly focused on the high-frequency region. In the temporal domain, screen content video often consists of static or rapid-moving texts and charts. Moreover, during web browsing, the video content can abruptly change in the next frame, known as "scene switch", which is frequently occurring in screen content videos. However, the abrupt changes in content typical of scene switches strain the motion compensation algorithms in SCC [1], which rely on continuity between frames. Consequently, these transitions can cause substantial drops in Peak Signal-to-Noise Ratio (PSNR), leading to noticeable visual quality degradations.

In recent years, various neural network architectures have been proposed for video quality enhancement in natural videos. Compared with single-frame methods [7]–[11], multiple-frame methods have shown even better enhancement results by utilizing the information from the neighbor frames. However, these multiple-frame methods, such as flow-based alignment [12], [13] and deformable-based alignment [14]–[16], based on motion consistency between frames in natural videos. They may always struggle to compensate for substantial content variations between frames [17], particularly

This work is supported by the Hong Kong Research Grants Council (RGC) under Research Grant PolyU 152069/18E, and Innovation and Technology Fund - Partnership Research Programme (ITF-PRP) under PRP/036/21FX. *(Corresponding author: Yui-Lam Chan)*

Ziyin Huang, Yui-Lam Chan, Ngai-Wing Kwong, and Kin-Man Lam are with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail: ziyin.huang@connect.polyu.hk; enylchan@polyu.edu.hk; ngai-wing.kwong@connect.polyu.hk; enkmlam@polyu.edu.hk)

Wing-Kuen Ling is with the Center for Integrated Circuits and Artificial Intelligence, Tsientang Institute for Advanced Study, Zhejiang, China (email: wkling@tias.ac.cn)

Sik-Ho Tsang is with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong, and the Department of Computer Science, Hong Kong Chu Hai College, Hong Kong (email: harristsang@chuhai.edu.hk)

during scene switches. This is due to the irrelevant information provided by neighbor frames during these switches, which can decrease the effectiveness of the quality enhancement model. As aforementioned, the particularity of screen content videos, with rapid scene changes and artifacts in high-frequency regions, makes it challenging for both optical flow and deformable convolution methods to accurately determine positions from preceding and following frames, limiting their effectiveness in enhancing screen content videos. Therefore, there is a critical need for the development of a new multi-frame video quality enhancement method that effectively addresses the unique challenges posed by scene switches in screen content videos.

Based on the unique characteristics of screen content, we propose a Long Short-term Fusion by Multi-scale Distillation (LSFMD) method to effectively restore high-frequency details and improve quality during scene switches in compressed screen content videos. This method consists of a long short-term feature extraction module and a high-frequency reconstruction module. The long short-term feature extraction module is designed to retain useful information from neighbor frames while minimizing their impact during scene transitions, allowing the model to enhance video quality despite scene switches and rapid motion. The high-frequency reconstruction module focuses on reconstructing the sharp edges, as the artifacts in screen content (SC) videos predominantly occur around these regions. In the short-term feature extraction stream, we introduce a Similarity-based Neighbor Frame Selector (SNFS) that identifies and selects relevant frames among neighbor frames to minimize disturbances from unrelated frames. This selector ensures that short-term information is extracted from frames with similar content, enhancing the accuracy of the reconstruction. The selected frames pass through the Multi-scale Residual Block (MSRB) to capture short-term features for flat areas and text regions using different kernel sizes, while a 3D Residual Block extracts long-term features for contextual information. To effectively fuse short-term and long-term information, we design a Multi-scale Hierarchical Feature Distillation (MHFD) mechanism. This mechanism transforms features from different scales to refine the hierarchical features at various network depths using local-global attention to distill significant features to the target frame which can capture more information for uneven noise distribution in screen content videos. This allows for better handling of scene switches and consistency in scenes. The fused features are then used as input for our proposed High-Frequency Reconstruction Block (HFRB), which utilizes the scale-space theory [18], [19] to factorize the feature map tensors and extract the high-frequency information to guide the model in restoring fine details of the target frame. This approach ensures the preservation and enhancement of critical high-frequency details, resulting in better video quality.

The main contributions of this work are summarized below:

- To the best of our knowledge, our proposed LSFMD is the first approach in screen content video quality enhancement to extract and fuse the long short-term features in the corresponding frames to improve frame quality during scene switches and restore the high-frequency detail.
- Instead of using a fixed set of neighbor frames to enhance the target frame, an SNFS is proposed to dynamically identify and select the most relevant frames based on content similarity. This adaptive frame selection mechanism minimizes the disturbance from unrelated frames, enhancing the accuracy of the reconstruction.
- To avoid the loss of features with the depth of the network, we propose the MHFD to capture the correlation of hierarchical features between short-term and long-term feature extraction streams to distillate the useful information related to the target frame, making the reconstructed frame more high-quality.
- Different from the conventional reconstruction part using vanilla convolution, the HFRB is proposed to parallelly reuse the high-frequency information of the target frame to adaptively restore the high-frequency details of the reconstructed frame.

The rest of this paper is organized as follows. Section II briefly introduces related works. In Section III, the proposed LSFMD model is presented in detail. In Section IV, we describe the experimental setting and analyze the performance of the experimental results. Finally, Section V concludes the paper.

## II. Related Works

### A. Single-frame Quality Enhancement

Numerous studies have been developed to enhance the quality of compressed videos using spatial information from a single frame. For instance, the In-loop Filtering CNN (IFCNN) [7] replaces the conventional Sample Adaptive Offset (SAO) filter with a three-layer CNN module to improve video quality within the codec. Similarly, the Variable-filter-size Residue-learning CNN (VRCNN) [8] aims to reduce distortion in videos by modifying internal codec modules. Other approaches focus on post-processing techniques to enhance video quality after decoding. For example, the Deep CNN-based Auto Decoder (DCAD) [9] employs ten convolutional layers to utilize spatial information and improve videos on the decoder side. The Quality Enhancement CNN (QE-CNN) [20] was designed to enhance both I and P/B frames, effectively addressing intra- and inter-coding quantization distortion. Additionally, the work in [21] proposed using partition information to boost the video quality. Our previous work [22] also utilizes mode information from the codec to guide the CNN in the enhancement process. However, these approaches primarily consider spatial information, overlooking the crucial role of temporal information in video quality enhancement.

### B. Multi-frame Quality Enhancement

Yang *et al.* introduced the Multi-Frame Quality Enhancement (MFQE 1.0) approach [12], which utilizes temporal information to enhance video quality. This method uses high-quality frames from compressed video as reference frames to improve the quality of low-quality target frames through a Multi-frame CNN. Subsequently, an updated version, MFQE 2.0 [13], was developed to improve efficiency and achieve better performance. These methods employ dense optical flow

for motion compensation to aggregate information from both target and reference frames. However, optical flow alignment is unsuitable for screen content video quality enhancement, as scene switches can disrupt the pixel-wise correspondence between frames, leading to inaccurate optical flow estimation. In addition to flow-based methods, deformable convolution-based methods have been proposed to learn offsets from compressed frames to obtain aligned features for VQE. An alternative work proposed in [14] is the deformable-based alignment (STDF) approach, which adaptively compensates for sampling positions of frames, capturing the most relevant context and removing artifacts in the target frame. The Spatio-Temporal Detail Information Retrieval (STDR) network in [15] incorporates a multi-path deformable alignment module to enhance the accuracy of offset generation by integrating alignment features from various receptive fields. In a related development, a new end-to-end network, termed Coarse-to-Fine Spatio-Temporal Information Fusion (CF-STIF) [16], has been proposed for enhancing the quality of compressed videos. This network advances the field by predicting more precise offsets, aided by its capability to utilize a larger receptive field. Besides, in the natural video, the motion vector can be utilized to guide the enhancement process in Coding Priors-Guided Aggregation Network (CPGA) [23]. While flow-based, deformable-based, and motion vector-based alignments have primarily been proposed for natural video quality enhancement, they may not effectively compensate for the position of the target frame in screen content videos. To enhance screen content videos, the Content Adaptive Network based on Two Branches (CAT) [24] was proposed to perform specific enhancements on text and graphics separately. Another method, Spatial-Temporal Adaptive (STA) [25], introduced a dual-branch structure for parallel single-frame and multi-frame feature extraction to enhance screen content videos. However, these approaches utilizing deformable convolution may potentially reduce the accuracy of compensating the target frame's position, which reduces their efficiency and practicality. The Quality Enhancement Network using Cross-Frame Information (QECF) [17] introduced a cross-fusion block instead of an alignment-based method. However, QECF was specifically developed for gaming videos, which consist of a series of consistent frames. A temporal group alignment and fusion network (TGAF) [26] was proposed for the quality enhancement of compressed videos by selecting the frames from the video to form a group of pictures according to the temporal distances to the target frame. However, the skipping selection adds irrelated frames when the scene switch occurs. To address the unique characteristics of screen content videos, which often involve dramatic motion and scene switches, we propose a novel network that overcomes the limitations of existing approaches. This new method is designed to handle the specific challenges posed by screen content videos, ensuring more accurate and effective video quality enhancement.

## III. Proposed Method

### A. Motivation

In Section I, we discussed the unique characteristics of screen content videos, such as rapid scene changes and artifacts in high-frequency regions, which make it challenging for existing methods to effectively enhance their quality. This can be observed in Fig. 1(a) and Fig. 1(b), where we compare typical frames from a natural video, and the screen content video, respectively. Fig. 1(c) and Fig. 1(d) then show the artifact distributions of the natural and screen content videos, respectively. The artifact distribution is obtained by calculating the differences between the reconstructed frame and the original frame. We can observe in Fig. 1(c) that the artifact appears throughout the entire area of the natural content due to its diverse range of colors and camera noise. In contrast, we can see in Fig. 1(d) that the artifact mainly occurs in the high-frequency regions of screen content. This highlights the importance of accurately reconstructing high-frequency details in screen content video quality enhancement.

In the temporal domain, we observe that scene switches

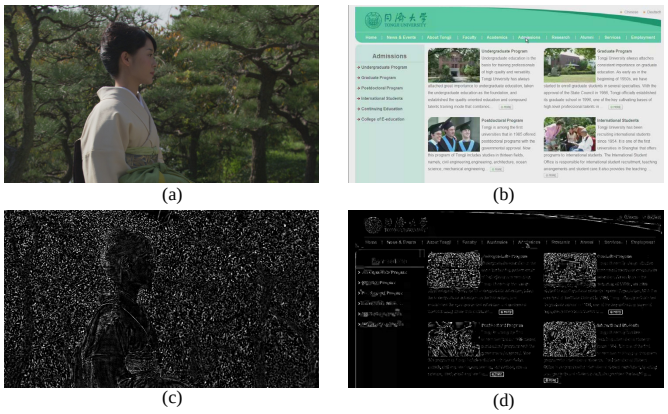



Fig. 1. (a) Original natural frame *Kimono*, (b) original screen content frame 68 *scwebbrowsing*, (c) artifact of natural frame, and (d) artifact of screen content frame.

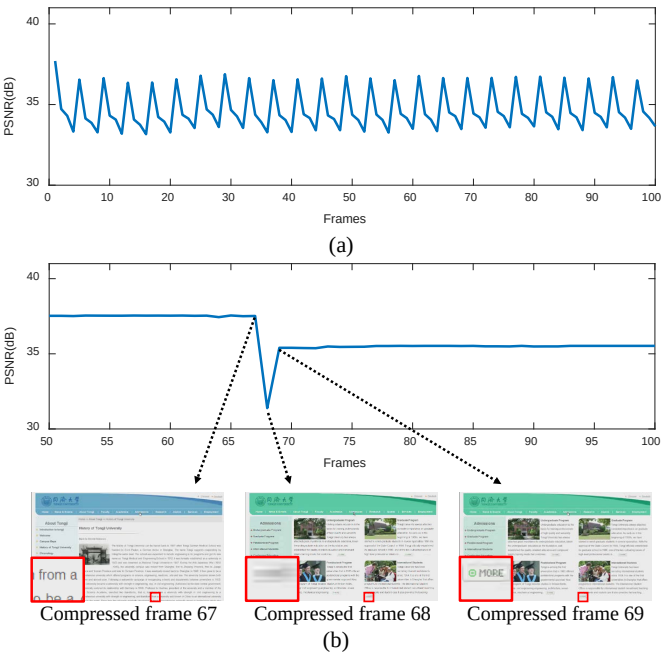



Fig. 2. (a) PSNR statistics for natural video *Kimono*, (b) PSNR statistics for screen content video *scwebbrowsing*.

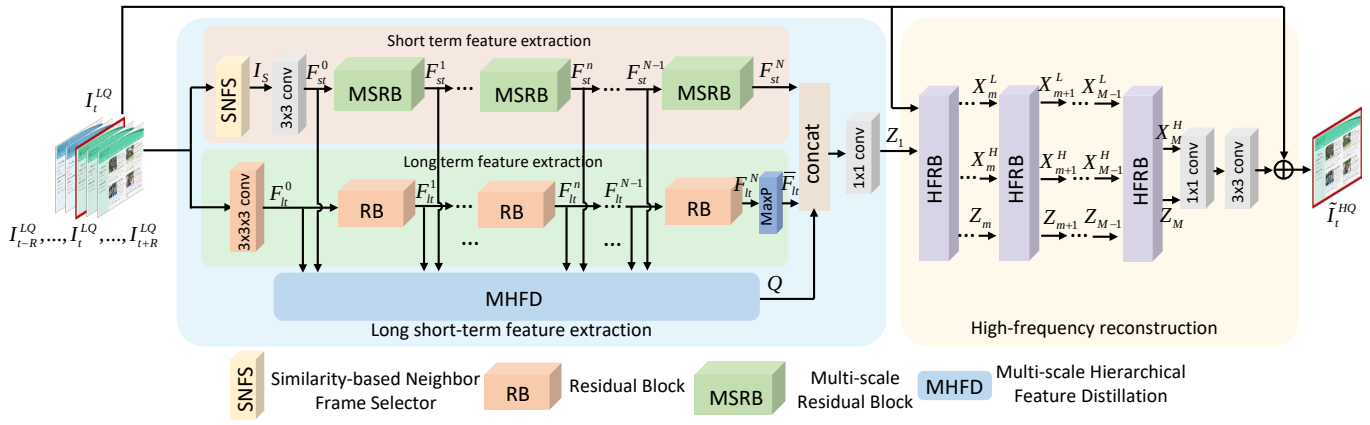



Fig. 3. Our proposed LSFMD structure, which contains long short-term feature extraction, multi-scale hierarchical feature distillation, and high-frequency reconstruction.

in screen content videos can cause significant drops in Peak Signal-to-Noise Ratio (PSNR), leading to noticeable visual quality degradations. To verify this, we encoded the natural video *Kimono* and the screen content video *scwebbrowsing*, and then calculated the PSNR between the reconstructed frames and the original frames as shown in Fig. 2. We observe that the PSNR changes in natural videos (Fig. 2 (a)) tend to remain within a certain range across different frames, while those in screen content videos exhibit significant variations. Besides, we also observe that the compressed screen content video exhibits significant PSNR drops during scene switches resulting in noticeable quality degradations that severely impact the Quality of Experience (QoE). However, existing multiple-frame methods, such as flow-based alignment [12], [13] and deformable-based alignment [14], [24]–[26], are inadequate for compensating for substantial content variations between frames [17], as they rely on a prediction network to compensate for the positions of neighbor frames. Inaccuracies in the prediction network can diminish the performance of the quality enhancement network. In addition, the alignment-free method described in [17] extracts the high-quality region from neighbor frames to enhance the target frame. However, these traditional methods that utilize a fixed set of neighbor frame to enhance the target frame may provide irrelevant information during scene switches, affecting the performance of the model. Consequently, we develop a novel video quality enhancement method to address the challenges of scene switches and dramatic motions in screen content videos.

### B. Overview of the Framework

Our LSFMD model, as shown in Fig. 3, aims to remove artifacts in screen content videos that involve numerous dramatic motion and scene switch scenarios. In this context, we denote a low-quality frame at time $t$ as $I_t^{LQ} \in \mathbb{R}^{H\times W}$, where $H$ and $W$ indicate the vertical and horizontal resolutions of the frame. The main objective of LSFMD is to enhance the quality of $I_t^{LQ}$ by effectively using both short-term and long-term temporal information. To achieve this, our model considers the preceding and succeeding $R = 2$ frames as reference frames to capture the necessary temporal context. The enhanced high-quality frame $\tilde{I}_t^{HQ} \in \mathbb{R}^{H\times W}$ can be expressed as

$$\tilde{I}_t^{HQ} = H_{LSFMD}(\{I_{t-R}^{LQ}, ..., I_t^{LQ}, ..., I_{t+R}^{LQ}\}) \quad (1)$$

where $H_{LSFMD}(\cdot)$ represents the proposed LSFMD, $\{I_{t-R}^{LQ}, ..., I_t^{LQ}, ..., I_{t+R}^{LQ}\}$ represents the group of the $2R+1$ input frames. Next, we will discuss our proposed framework in Fig. 3, which comprises two modules: a long short-term feature extraction module and a high-frequency reconstruction module. In the long short-term feature extraction module, we construct two streams: a short-term feature extraction stream and a long-term feature extraction stream. These streams aim to extract short-term and long-term information from input frames of varying lengths. In addition, a multi-scale hierarchical feature distillation (MHFD) is proposed to enhance the reusability and effectiveness of the short-term and long-term features. This approach enables us to handle the scene switch situation adaptively. By assigning weights to the features in an adaptive manner, our network can effectively learn the correlations between short-term and long-term features. In the reconstruction part, we focus on reconstructing the high-frequency information of the target frame. This component plays a crucial role in enhancing the visual quality of the output, particularly in preserving and enhancing the fine details that are often lost during compression. We will provide a detailed explanation of each component within our LSFMD frame in the following subsections.

### C. Long Short-term Feature Extraction

The utilization of single-stream deep neural networks has been widely used for video quality enhancement [14], [17], [24]. However, as the depth of the neural network increases, the presence of unrelated features from neighbor frames can hinder the model's ability to effectively learn and extract the relevant information related to the target frame. This issue becomes particularly problematic in scenes with rapid motion and frequent scene switches, as the unrelated features can have a detrimental impact on the quality enhancement

of the target frame. To tackle this challenge, it is crucial to focus on the useful and related features of the target frame, especially during situations involving rapid motion and scene switches. Consequently, in contrast to the traditional single-steam method, we propose a long short-term feature extraction stream, where the short-term stream provides the relevant features to assist the long-term stream, enabling a more focused and effective analysis of the target frame.

Within the long-term feature extraction stream, using the 3D Residual Block to extract the long-term feature allows the network to understand the video content in spatial and temporal domain which is crucial for maintaining the integrity of text and graphics across consecutive frames. The structure of the long-term feature extraction stream is shown in Fig. 3. We first transform the input sequence to the feature domain by applying a 3D convolution layer to obtain the initial feature $F_{lt}^0$ as:

$$F_{lt}^0 = Conv_{3\times3\times3}(\{I_{t-R}^{LQ}, ..., I_t^{LQ}, ..., I_{t+R}^{LQ}\}) \tag{2}$$

where $Conv_{3\times3\times3}(\cdot)$ denotes the $3 \times 3 \times 3$ convolution layer. Then stacked Residual Blocks compute the features as:

$$F_{lt}^n = H_{RB}^n(F_{lt}^{n-1}), n \in [1, N] \tag{3}$$

$$\bar{F}_{lt} = MaxP(F_{lt}^N) \tag{4}$$

where $N$ is the total number of residual blocks in the long-term feature extraction, $F_{lt}^n$ represents the extracted features after the $n^{th}$ residual blocks $H_{RB}^n(\cdot)$, and $MaxP(\cdot)$ denotes the maxpooling, which is utilized to transform the feature domain. The output of the long-term feature extraction stream, denoted as $\bar{F}_{lt}$, is obtained by passing the features through the $N^{th}$ residual block $H_{RB}^n(\cdot)$, followed by $MaxP(\cdot)$.

This output, $\bar{F}_{lt}$, encapsulates the contextual information, but special attention must be given to flat areas and repetitive text regions, which are commonly found in screen content videos. These regions require different scale filters to effectively capture more useful information. Inspired by the Multi-Scale Residual Block (MSRB) [27] used in image super-resolution, the short-term information extraction stream utilizes the MSRB to adaptively detect features at different scales. For flat areas, larger filters can be used to capture broader context, which is significant for these regions. In contrast, sharp edges, such as those found in text, are critical features that need to be preserved in screen content videos to maintain readability. To address this, the MSRB employs smaller kernels to capture the high-frequency details associated with text edges, ensuring that the sharpness and clarity of text are retained in the reconstructed frame. The structure of the short-term feature extraction stream is summarized as follows:

$$I_S = H_{SNFS}(\{I_{t-R}^{LQ}, ..., I_t^{LQ}, ..., I_{t+R}^{LQ}\}) \tag{5}$$

$$F_{st}^0 = Conv_{3\times3}(I_S) \tag{6}$$

$$F_{st}^n = H_{MSRB}^n(F_{st}^{n-1}), n \in [1, N] \tag{7}$$

where $Conv_{3\times3}(\cdot)$ denotes the $3 \times 3$ convolution layer. Moreover, $H_{SNFS}$ represents the Similarity-based Neighbor Frame Selector, which identifies the most relevant short-term neighbor frames to the target frame $I_S$, as discussed in the next subsection, and $F_{st}^n$ represents the extracted features after the $n^{th}$ MSRB $H_{MSRB}^n(\cdot)$, respectively. We can obtain the output $F_{st}^N$ of the short-term feature extraction stream by stacking the MSRB.

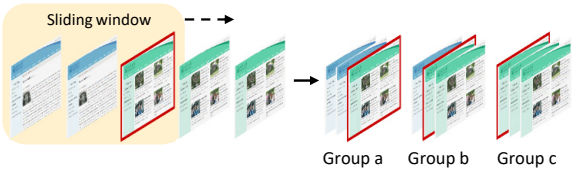



Fig. 4. SNFS, Group a: $\{I_{t-2}^{LQ}, I_{t-1}^{LQ}, I_t^{LQ}\}$, Group b: $\{I_{t-1}^{LQ}, I_t^{LQ}, I_{t+1}^{LQ}\}$, Group c: $\{I_t^{LQ}, I_{t+2}^{LQ}, I_{t+2}^{LQ}\}$

**Similarity-based Neighbor Frame Selector:** Extracting the short-term feature from the shorter input can reduce the disturbance from unrelated neighbor frames. However, during scene transitions, the fixed window for choosing neighbor frames may introduce irrelevant information. To further enhance the frame quality during scene switches, the proposed method incorporates a similarity-based neighbor frame selector (SNFS) in the short-term feature extraction stream. In the SNFS, we employ a sliding window to separate the input frames $\{I_{t-2}^{LQ}, ..., I_t^{LQ}, ..., I_{t+2}^{LQ}\}$ to different groups as shown in Fig. 4. Group a denotes $\{I_{t-2}^{LQ}, I_{t-1}^{LQ}, I_t^{LQ}\}$, group b denotes $\{I_{t-1}^{LQ}, I_t^{LQ}, I_{t+1}^{LQ}\}$, and group c denotes $\{I_t^{LQ}, I_{t+2}^{LQ}, I_{t+2}^{LQ}\}$, SNFS, then calculates the pearson correlation coefficient [28], [29] between each neighbor frame and target frame in each group. This calculation allows for the selective identification of frames that are most relevant and pertinent to the target frame. A larger pearson correlation coefficient indicates a greater degree of similarity. In summary, the working process of the proposed SNFS is operated as:

$$\begin{cases} P_a = pearson(I_{t-2}^{LQ}, I_t^{LQ}) + pearson(I_{t-1}^{LQ}, I_t^{LQ}), \\ P_b = pearson(I_{t-1}^{LQ}, I_t^{LQ}) + pearson(I_{t+1}^{LQ}, I_t^{LQ}), \\ P_c = pearson(I_{t+1}^{LQ}, I_t^{LQ}) + pearson(I_{t+2}^{LQ}, I_t^{LQ}), \\ P = \max(P_a, P_b, P_c) \end{cases} \tag{8}$$

where $pearson(\cdot)$ denotes the operation to calculate the pearson correlation coefficient between the neighbor frames and the target frame, and $P$ denotes the maximum value among $P_a, P_b$, and $P_c$. Once $P$ is determined, the SNFS chooses the group with the larger pearson correlation as the input $I_S$ in Eq. (5). This adaptive selection enables quality enhancement, especially in the context of scene switches and fast motion, where the fixed-window approach may introduce irrelevant information. By adopting the pearson correlation-based frame similarity evaluation, the SNFS can effectively identify the most relevant neighbor frames to the target frame, ensuring that the short-term feature extraction stream has access to the most pertinent information for improving the overall video quality.

### *D. Multi-scale Hierarchical Feature Distillation*

As the depth of the network increases, the extracted short-term and long-term features will gradually disappear during

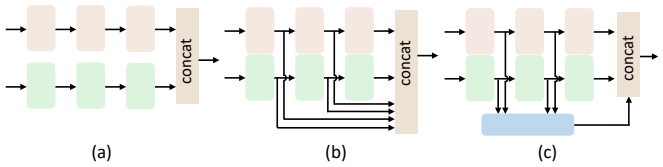



Fig. 5. Comparisons of different hierarchical feature utilization methods, (a) Structure A, (b) Structure B, and (c) Structure C.

the conduction process. Therefore, taking advantage of hierarchical features becomes crucial for significantly improving model performance. However, many existing models overlook the importance of hierarchical features, as shown in Fig. 5 (a), resulting in sub-optimal results. Moreover, simply concatenating all hierarchical features, as in Fig. 5 (b), fails to eliminate redundant features, resulting in inefficient video reconstruction. Therefore, an effective method that can exploit hierarchical features and eliminate redundant features is crucial for screen content video quality enhancement. To address this, our proposed MHFD introduces two key components: the Feature Transformation Strategy and the Local-global Channel Attention Mechanism.

**Feature Transformation Strategy:** The feature transformation component is specially designed to refine the hierarchical features at different network depths through a series of nonlinear transformations. This process aims to enhance the representational power of the network. To achieve this, we employ a series of convolutional layers:

1) Initial feature combination: After obtaining the hierarchical feature from the first convolution layer, we utilize a $1 \times 1$ convolution layer to combine the short-term and long-term features.
2) Shallow feature extraction: Subsequently, a $5 \times 5$ convolution layer is employed to extract the shallow features. The use of a larger $5 \times 5$ receptive field ensures the retention and amplification of salient features.
3) Deeper feature processing: The remaining hierarchical features are processed by a combination of $1 \times 1$ and $3 \times 3$ convolution layers. This combination allows for capturing finer details by utilizing a smaller receptive field.

The process can be summarized as:

$$\begin{aligned} \tilde{F}^n &= Conv_{k\times k}(Conv_{1\times 1}([MaxP(F_g^n), F_l^n])) \\ k &= \begin{cases} 5, & n = 0 \\ 3, & otherwise \end{cases} \end{aligned} \tag{9}$$

where $n = 0, \cdots, N-1$, $\tilde{F}^n$ denotes the feature obtained from the $n^{th}$ feature transformation branch, $Conv_{k\times k}(\cdot)$ presents the $k \times k$ convolution layer, and $[\cdot, \cdot]$ denotes the concatenation operation.

**Local-global Channel Attention Mechanisms:** As in Fig. 6, the inputs of MHFD are the hierarchical features obtained through convolution at different scales. However, these features may contain redundancy information. To further distillate the useful information in the target frame, we design a local-global attention mechanism that combines the benefits of local

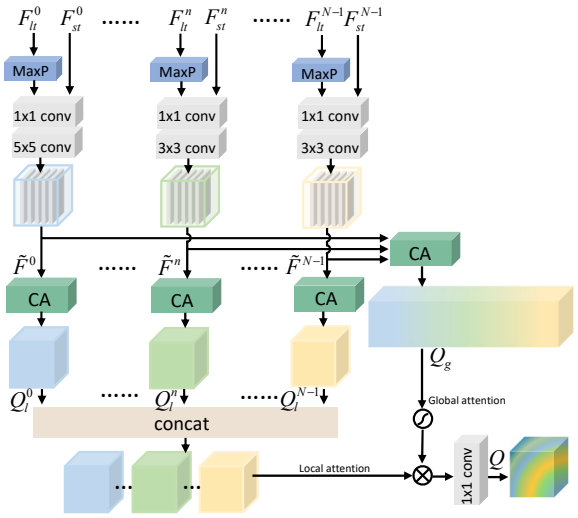



Fig. 6. Multi-scale hierarchical feature distillation (MHFD).

attention and global attention. The local channel attention mechanism is tailored to focus on feature maps specific to certain channel locations. This allows the model to prioritize local patterns and textures that are essential for high-quality video reconstruction in each hierarchical feature branch. We utilize the channel attention [30] to generate the attention weight $\tilde{f}^n$, which can be obtained as:

$$H_{CA}(\cdot) = \sigma(Conv_{1\times 1}(ReLU(Conv_{1\times 1}(AvgP(\cdot))))) \tag{10}$$

$$\tilde{f}_l^n = H_{CA}(\tilde{F}^n) \tag{11}$$

where $H_{CA}(\cdot)$ denotes the channel attention operation, $\sigma(\cdot)$ is the sigmoid function, $ReLU(\cdot)$ is the ReLU [31] activation function, and $AvgP(\cdot)$ represents the average pooling operation. The attention weights $\tilde{f}^n$ indicate the sensitivity of different features in the $n^{th}$ feature transformation branch. Hence, local attention feature $Q_l^n$ can be computed as:

$$Q_l^n = \tilde{f}_l^n \cdot \tilde{F}^n \tag{12}$$

On the other hand, the global channel attention mechanism offers a broader perspective by considering the entire channel extent of the feature maps. By assigning attention weights across different hierarchical feature branches, we can prevent the loss of high-frequency hierarchical features as the network depth increases. The synergy between local and global channel attention mechanisms facilitates a more dynamic and context-aware feature distillation. The global attention feature $Q_g$ can be obtained as:

$$\tilde{f}_g = H_{CA}([\tilde{F}^0, \cdots, \tilde{F}^{N-1}]) \tag{13}$$

$$Q_g = \tilde{f}_g \cdot [\tilde{F}^0, \cdots, \tilde{F}^{N-1}] \tag{14}$$

where $\tilde{f}_g$ denotes the attention weight assigned for all hierarchical features.

Finally, the output feature map $Q$ of the MHFD can be obtained by combining the local and global attention features:

$$Q = Conv_{1\times 1}([Q_l^0, \cdots, Q_l^{N-1}] \otimes \sigma(Q_g)) \tag{15}$$

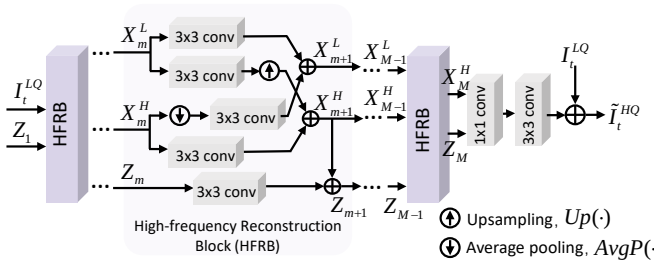



Fig. 7. The structure of HFRB in the high-frequency reconstruction module.

where $\otimes$ denotes elementwise multiplication. Here, the output $Q$ of MHFD is the local-global attention-weighted feature which contains the refined information from each scale of long short-term feature extraction. After we obtain the distillated feature $Q$, we can fuse it with the short-term and long-term features, as depicted in Fig. 3, as:

$$Z_1 = Conv_{1\times1}([F_{st}^N, \bar{F}_{lt}, Q]) \tag{16}$$

where $Z_1$ is the output of the long short-term feature extraction module in Fig. 3, which contains the long short-term features and multi-scale hierarchical features. Increasing the depth of the network causes the extracted short-term and long-term features to diminish in prominence as they propagate through the model. Therefore, taking advantage of hierarchical features will greatly improve model performance.

### *E. High-frequency Reconstruction*

Conventional reconstruction blocks using vanilla convolution typically focus on reconstructing the frame as a whole. However, this approach sometimes overlooks finer details that contribute significantly to the perceived sharpness and clarity of the screen content image. High-frequency components of this image, such as edges, textures, fonts, and fine structures, contain crucial details. By explicitly incorporating this information into the reconstruction block, we can restore fine details that are often lost during the compression process, which is vital for delivering an enhanced visual experience in screen content videos. To extract the high-frequency feature adaptively, we utilize the scale-space theory introduced in [18], [19] to factorize the feature map tensors into low- and high-frequency groups. The HFRB is then designed in this paper to effectively integrate this high-frequency information into the reconstruction module. With the stacking of the HFRB in Fig. 7, the interaction between the extracted high-frequency features and the reconstructed features dynamically fine-tunes the high-frequency details in parallel. The synergistic effect of this parallel interaction not only aids in restoring the completeness of textures and edges but also enhances the overall quality of the reconstructed feature.

Different from the traditional reconstruction part using a single input from the previous layer, our HFRB in Fig. 7 combines features from the previous layer and extra features extracted from the target frame. Let us denote the input feature tensors to the $m^{th}$ HFRB, where $m = 1, \cdots, M$, as:

1) $X_m^H \in \mathbb{R}^{(1-\alpha)c_f \times h \times w}$: high-frequency feature maps extracted from the target frame.
2) $X_m^L \in \mathbb{R}^{\alpha c_f \times \frac{h}{2} \times \frac{w}{2}}$: low-frequency feature maps extracted from the target frame.
3) $Z_m \in \mathbb{R}^{c \times h \times w}$: features from the long short-term extraction module and the high-frequency feature from the target frame excepting the input of the first HFRB, which is $Z_1$ in Eq. (16).

where $h$ and $w$ denote the spatial dimensions, and $c$ and $c_f$ denote the channel number where $c_f = 2c$ apart from the last HFRB. In the last HFRB, $c_f$ is equal to $c$ for the channel matching. Here, $\alpha \in [0, 1]$ denotes the ratio of channels allocated to the low-frequency part. The setting of the $\alpha$ will be introduced in Section IV-A.

By explicitly incorporating the target frame information into the reconstruction block, our model can more effectively restore fine details often lost during the training process. The output feature tensors of the $m^{th}$ HFRB is denoted as $X_{m+1}^H \in \mathbb{R}^{(1-\alpha)c_f \times h \times w}$, $X_{m+1}^L \in \mathbb{R}^{\alpha c_f \times \frac{h}{2} \times \frac{w}{2}}$, and $Z_{m+1} \in \mathbb{R}^{c \times h \times w}$.

The process of HFRB is represented as:

$$\{X_{m+1}^H, X_{m+1}^L, Z_{m+1}\} = H_m^{HFRB}(X_m^H, X_m^L, Z_m) \tag{17}$$

where

$$\begin{cases} X_{m+1}^H = Conv_{3\times3}(X_m^H) + Up(Conv_{3\times3}(X_m^L)), \\ X_{m+1}^L = Conv_{3\times3}(X_m^L) + Conv_{3\times3}(AvgP(X_m^H)), \\ Z_{m+1} = Conv_{3\times3}(Z_m) + X_{m+1}^H \end{cases} \tag{18}$$

where $H_m^{HFRB}(\cdot)$ denotes the $m^{th}$ HFRB and $Up(\cdot)$ denotes the upsampling operation by a scale factor of 2. $Up(\cdot)$ operation denotes the upsampling of the input by a scale factor of 2. The $Up(\cdot)$ and $AvgP(\cdot)$ operations are used for communication between the low-frequency and high-frequency feature groups, which helps adjust the feature dimensions. In the HFRB, the output feature $Z_{m+1}$ encapsulates the high-frequency information from the target frame. This feature is subsequently fed into the next layer to extract deeper features. Concurrently, the high-frequency details from the target frame are fed into the subsequent layer for analysis and extraction of the most pertinent features. The HFRB's capability to handle multiple inputs allows for the parallel extraction and integration of high-frequency information. This enables the LSFM model to dynamically shift its focus towards these crucial details as it progresses deeper into the network. This adaptive mechanism ensures that the essential high-frequency characteristics from the target frame are not overlooked but are instead emphasized throughout the reconstruction process.

Finally, the reconstructed frame can be represented as:

$$\tilde{I}_t^{HQ} = Conv_{3\times3}(Conv_{1\times1}([Z_M, X_M^H])) + I_t^{LQ} \tag{19}$$

where the $Z_M$ and $X_M^H$ are the output of the last HFRB. In the reconstruction module, this high-frequency information is progressively integrated with the major features in the HFRB. The parallel extraction and integration of high-frequency details enable the model to dynamically adjust its focus on these components as the network deepens. This newly adaptive mechanism ensures that the essential details from the target frame are not lost but rather emphasized, leading to improved frame quality.

### F. Training Scheme

To effectively handle the high-frequency information and improve the performance, we adopt the robust Charbonnier loss function in [32], [33] to train our model in an end-to-end manner. The loss function $L$ is represented as:

$$L = \sqrt{\left\| I_t^{HQ} - \tilde{I}_t^{HQ} \right\|^2 + \varepsilon^2} \tag{20}$$

where $I_t^{HQ}$ is the ground truth frame at time $t$, $\tilde{I}_t^{HQ}$, represents the enhanced frame generated at time $t$ by our model, and $\varepsilon = 10^{-3}$ is a constant value used across all experiments.

## IV. Experimental Results

### A. Implementation Details

Our proposed LSFMD model mainly focuses on enhancing the video quality of screen content sequences. In our LSFMD framework, each convolutional layer, except for the final convolutional layer, is followed by a ReLU activation function [31] to introduce non-linearity into the model. Due to the limited number of available screen content sequences within the Common Test Condition (CTC) [34], we gathered additional screen content sequences from other sources [35]–[37]. Our dataset consists of 41 video sequences with various resolutions, including 2560×1440, 1920×1080, and 1280×720. The lengths of these videos range from 300 to 600 frames, with frame rates varying between 20 and 60 fps. Among these sequences, 28 videos were adopted for training and the remaining 13 videos were for model testing. In the test set, 10 video sequences are provided from the CTC [34], that is a common dataset to exemplify various challenges in video quality enhancement. Notably, the CTC dataset contains only 3 videos characterized by frequent scene switches and dramatic motions. To make a robust assessment of the model's capabilities in handling real-world scenarios that feature rapid scene changes and motion complexities, we added 3 self-capture sequences to introduce more variations with scene transitions and dynamic motions. The video sequences were encoded using the HEVC reference software HM16.20-SCM8.8 under Low Delay Main SCC (LDMS) configuration as the network inputs, while the uncompressed raw video sequences were used as the ground-truths. We utilized four Quantization Parameters (QPs) of 22, 27, 32, and 37 for encoding the sequences and training a separate model for each QP. During training, only the luminance channel (Y channel) of each frame was considered as input. Model construction and training were implemented using PyTorch. The patch size of each input image and its corresponding ground truth was 128 × 128. To augment our dataset, we randomly selected 300 patches from one frame for each iteration. In our experiments, the learning rate was set to 0.0001 for all QPs. The adaptive moment estimation (Adam) optimization method [38] was used to train the model for 300000 iterations. A computer equipped with Ubuntu 20.04 operating system, an Intel i9-10900K CPU, 64 GB RAM, and NVIDIA 3090Ti GPUs, was used to perform the model training.

In the LSFMD model, the number of MSRB and RB are set as 3 ($N = 3$) and the number of HFRB is also set as 3 ($M = 3$). The $\alpha$ in HFRB was set as 0.5 throughout the module for the channel matching between the extracted high-frequency feature and the reconstructed feature within the HFRB, apart from the first and the last HFRB. To convert a vanilla feature representation to a low-frequency and high-frequency feature representation, we set $\alpha$ in the first HFRB to 0. In this case, the low-frequency input of the first HFRB is disabled. To convert the low-frequency and high-frequency feature representation back to vanilla feature representation, we set $\alpha$ in the last HFRB to 0, disabling the low-frequency output in the HFRB, and resulting in a single output.

### B. Overall Performance

**Objective Visual Quality Assessment:** In this section, we compare the proposed LSFMD method with the state-of-the-art video quality enhancement methods, STDF-R3 [14], QECF [17], CAT [24], TGAF [26], STA [25], CF-STIF-M [16], and STDR [15]. To evaluate the quality enhancement performance of each quality enhancement method, the Peak Signal-to-Noise Ratio (PSNR) improvement (ΔPSNR) and the Structural Similarity Index (SSIM) improvement (ΔSSIM) are used. Table I shows the average ΔPSNR and the average ΔSSIM, respectively, over all frames of each test sequence. The best ΔPSNR/ΔSSIM is highlighted in bold. We can see that our proposed LSFMD outperforms other methods in most cases, highlighting the effectiveness of our approach. For instance, when using a QP of 37, our LSFMD achieves the highest ΔPSNR of 1.915 dB for the *paperpdf* sequence, which contains text and graphics. The average ΔPSNR of our LSFMD is 0.938 dB, which is 46.33% higher than that of CAT (0.641 dB), 52.52% higher than that of QECF (0.615 dB), 48.42% higher than that of STDF-R3 (0.632 dB), 16.09% higher than that of TGAF (0.808 dB), 8.31% higher than that of STA (0.866 dB), 17.25% higher than that of CF-STIF-M (0.800 dB), 20.72% higher than that of STDR (0.777 dB), and 21.19% higher than that of EAST-LITE (0.774 dB). For other QPs ( 22, 27, and 32), our LSFMD approach also outperforms other state-of-the-art video quality enhancement approaches. A similar trend can be found for ΔSSIM. This demonstrates that our LSFMD approach not only performs well in reducing pixel-level differences but also enhances the visual quality perceived by the human visual system. To further evaluate the performance, BD-rate [34] is used to indicate the bitrate savings achieved by these models under the equivalent PSNR. The experimental results are compared and tabulated in Table II. Our LSFMD obtains an average BD-rate savings of 7.53%. For the test sequence *scSlideShow* with dramatic motion and scene switch, our LSFDM achieves up to 12.50% BD-rate saving for the Y component under LDMS configuration. We conjecture that our LSFMD effectively removes the artifacts and restores the high-frequency information, thereby enhancing the quality of decoded frames and reducing the BD-rate.

**Subjective Visual Quality Comparison:** This section compares the subjective quality of different models. Fig. 8 shows the subjective visual quality performance of various models on the sequences *ChineseEditing*, *MissionControlClip3*, and *scwebbrowsing*, all encoded with QP = 37. From this figure,

TABLE I
OVERALL ΔPSNR AND ΔSSIM ($\times 10^{-3}$) OF DIFFERENT MODELS AT QP=22,27,32,37

| QP | Seq. | STDF-R3 [14] | | QECF [17] | | CAT [24] | | TGAF [26] | | STA [25] | | CF-STIF-M [16] | | STDR [15] | | EAST-LITE [37] | | Proposed LSFMD | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ΔPSNR | ΔSSIM | ΔPSNR | ΔSSIM | ΔPSNR | ΔSSIM | ΔPSNR | ΔSSIM | ΔPSNR | ΔSSIM | ΔPSNR | ΔSSIM | ΔPSNR | ΔSSIM | ΔPSNR | ΔSSIM | ΔPSNR | ΔSSIM |
| | 1 | 0.327 | 3.18 | 0.325 | 3.23 | 0.318 | 4.19 | 0.432 | 4.11 | 0.477 | 4.92 | 0.369 | 3.98 | 0.408 | 4.33 | 0.411 | **6.02** | **0.451** | 4.59 |
| | 2 | 0.273 | 2.22 | 0.244 | 1.63 | 0.200 | 1.24 | 0.480 | 3.79 | 0.382 | 2.12 | 0.360 | 2.41 | 0.321 | 1.77 | 0.525 | **5.55** | **0.529** | 4.54 |
| | 3 | 0.867 | 2.78 | 0.770 | 2.7 | 0.951 | 3.27 | 1.101 | 3.63 | 1.134 | 3.62 | 1.243 | 3.41 | 1.266 | 4.01 | 0.991 | **4.82** | **1.356** | 4.25 |
| | 4 | 0.492 | 4.17 | 0.503 | 4.12 | 0.477 | 4.00 | 0.610 | 4.30 | 0.672 | 4.74 | 0.625 | 4.72 | 0.546 | 4.59 | 0.573 | 4.17 | **0.679** | **5.12** |
| | 5 | 1.281 | 2.87 | 1.225 | 2.67 | 1.421 | 3.08 | 1.728 | 3.18 | 1.771 | 3.38 | 1.718 | 3.28 | 1.718 | 3.31 | 1.500 | 3.36 | **1.915** | **3.49** |
| 37 | 6 | 0.779 | 2.38 | 0.831 | 2.34 | 0.864 | 2.79 | 1.233 | 4.24 | 1.254 | 3.90 | 1.127 | 3.32 | 1.189 | 3.81 | 1.069 | **4.56** | **1.299** | 4.08 |
| | 7 | 0.301 | 3.46 | 0.365 | 3.51 | 0.329 | 3.05 | 0.516 | 4.13 | 0.538 | 3.85 | 0.278 | 3.85 | 0.306 | 4.16 | 0.528 | 3.07 | **0.589** | **4.78** |
| | 8 | 0.914 | 4.02 | 0.910 | 3.98 | 0.878 | 4.21 | 0.866 | 4.11 | **1.166** | **4.62** | 1.054 | 4.30 | 1.094 | 4.52 | 1.076 | 3.85 | 1.165 | 4.52 |
| | 9 | 0.453 | 5.71 | 0.373 | 3.53 | 0.416 | 6.26 | 0.463 | 6.44 | 0.529 | 6.61 | 0.526 | 5.97 | 0.408 | 5.04 | 0.476 | 4.35 | **0.56** | **6.83** |
| | 10 | 0.406 | 4.90 | 0.427 | 4.93 | 0.403 | 4.86 | 0.545 | 4.97 | 0.597 | 5.72 | 0.520 | 5.51 | 0.514 | 5.11 | 0.545 | 4.59 | **0.635** | **5.77** |
| | 11 | 1.008 | 3.28 | 0.907 | 3.38 | 0.969 | 3.56 | 1.137 | 3.78 | 1.292 | 4.02 | 1.286 | 3.93 | 1.046 | 3.72 | 1.107 | **6.91** | **1.494** | 4.23 |
| | 12 | 0.569 | 5.33 | 0.563 | 5.19 | 0.568 | 5.18 | 0.721 | 5.73 | 0.754 | 5.93 | 0.672 | 5.67 | 0.689 | 6.08 | 0.647 | 5.82 | **0.799** | **6.67** |
| | 13 | 0.545 | 5.06 | 0.551 | 4.96 | 0.535 | 4.98 | 0.677 | 5.38 | 0.698 | 5.55 | 0.625 | 5.25 | 0.591 | 5.51 | 0.612 | 3.69 | **0.720** | **6.01** |
| | Avg. | 0.632 | 3.80 | 0.615 | 3.55 | 0.641 | 3.90 | 0.808 | 4.45 | 0.866 | 4.54 | 0.800 | 4.28 | 0.777 | 4.30 | 0.774 | 4.67 | **0.938** | **4.99** |
| 32 | Avg. | 0.533 | 2.09 | 0.531 | 2.12 | 0.541 | 2.04 | 0.656 | 2.19 | 0.790 | 2.66 | 0.704 | 2.54 | 0.655 | 2.37 | 0.684 | 2.37 | **0.798** | **2.67** |
| 27 | Avg. | 0.467 | 0.91 | 0.495 | 1.07 | 0.429 | 0.91 | 0.586 | 0.99 | 0.626 | 1.15 | 0.608 | 1.22 | 0.588 | 1.10 | 0.548 | 1.12 | **0.692** | **1.27** |
| 22 | Avg. | 0.417 | 0.53 | 0.470 | 0.54 | 0.426 | 0.55 | 0.550 | 0.61 | 0.603 | 0.66 | 0.533 | 0.64 | 0.537 | 0.62 | 0.496 | 0.61 | **0.611** | **0.68** |

1: BigBuck(1920×1080, 404 frames, 60 fps) 2: ChineseEditing(1920×1080, 600 frames, 60 fps) 3: EnglishDocumentEditing(1920×1080, 300 frames, 30 fps) 4: MissionControlClip3(1920×1080, 600 frames, 60 fps) 5: Paperpdf(1920×1080, 300 frames, 60 fps) 6: Sephora(1920×1080, 300 frames, 60 fps) 7: mixvideo(1920×1080, 300 frames, 60 fps) 8: scSlideShow(1280×720, 500 frames, 20 fps) 9: scmap(1280×720, 600 frames, 60 fps) 10: scprogramming(1280×720, 600 frames, 60 fps) 11: scwebbrowsing(1280×720, 300 frames, 30 fps) 12: MissionControlClip1(2560×1440, 600 frames, 60 fps) 13: MissionControlClip2(2560×1440, 600 frames, 60 fps).

TABLE II
OVERALL BD-RATE(%) OF DIFFERENT MODELS AT QP=22,27,32,37

| Sequences | STDF-R3 [14] | QECF [17] | CAT [24] | TGAF [26] | Proposed LSFMD |
|---|---|---|---|---|---|
| BigBuck | -5.33 | -6.09 | -6.13 | -7.24 | **-7.90** |
| ChineseEditing | -1.39 | -1.44 | -1.25 | -2.06 | **-3.03** |
| EnglishDocumentEditing | -2.67 | -2.59 | -2.68 | -3.53 | **-4.33** |
| MissionControlClip3 | -6.02 | -6.24 | -5.91 | -7.17 | **-8.21** |
| Paperpdf | -4.17 | -4.53 | -4.19 | -5.81 | **-7.03** |
| Sephora | -5.81 | -6.07 | -5.78 | -7.25 | **-10.07** |
| mixvideo | -2.05 | -2.25 | -2.00 | -2.69 | **-3.34** |
| scSlideShow | -9.94 | -10.05 | -9.85 | -11.19 | **-12.50** |
| scmap | -7.22 | -6.42 | -6.00 | -7.58 | **-8.56** |
| scprogramming | -5.81 | -6.39 | -6.33 | -8.02 | **-8.66** |
| scwebbrowsing | -3.14 | -2.91 | -2.90 | **-3.67** | -3.66 |
| MissionControlClip1 | -7.27 | -7.44 | -7.44 | -9.09 | **-10.87** |
| MissionControlClip2 | -7.25 | -7.54 | -7.19 | -8.65 | **-9.70** |
| **Average** | -5.24 | -5.38 | -5.20 | -6.46 | **-7.53** |

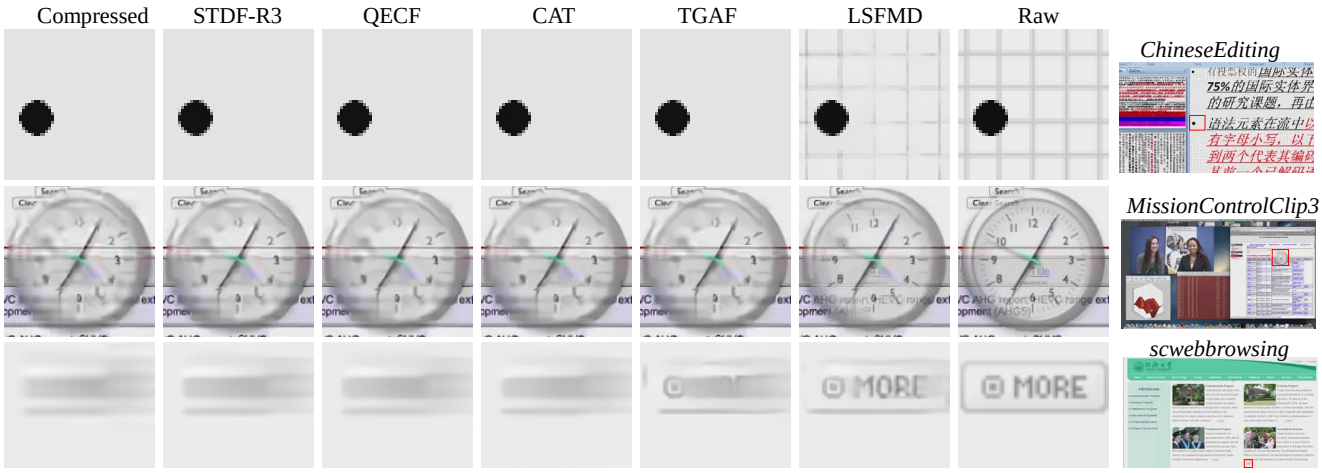



Fig. 8. Subjective visual quality comparison at QP = 37 on *ChineseEditing*, *MissionControlClip3*, and *scwebbrowsing*.

we can clearly see that the reconstructed frames of HM16.20-SCM8.8 exhibit noticeable compression artifacts and suffer from significant loss of high-frequency information details. These artifacts and details cannot be effectively restored by STDF-R3 [14], QECF [17],CAT [24], or TGAF [26]. As depicted in Fig. 8, our proposed LSFMD removes the artifacts and restores the content more effectively than the other models. Taking the *ChineseEditing* sequence as an example, it can be observed that the edges of the background still disappear in other methods, but they are successfully restored by our LSFMD. For *MissionControlClip3*, the clock's numbers are blurry and the words in the background under the clock are unreadable. After being processed by our LSFMD, the details of the clock are restored clearly, and the content of the

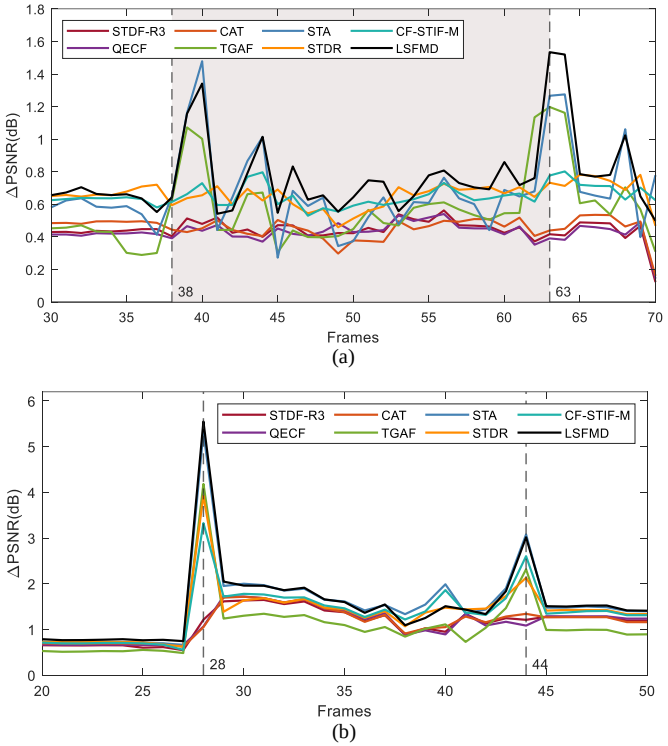



Fig. 9. ΔPSNR curves of STDF-R3, QECF, CAT, and our LSFMD method for sequences, (a) *scprogramming* and (b) *scSlideShow*.

background under the clock is clear. For *scwebbrowsing*, we visualize the frame during the scene switch situation. There is a loss of high-frequency information, resulting in blurry text and icons. However, when applying our proposed approach, these elements become clearer. The examples presented in Fig. 8 collectively demonstrate the superiority of LSFMD over other models in terms of subjective visual quality. Once again, this showcases the ability of our LSFMD model to effectively restore high-frequency information and handle scenarios involving scene switches.

Additionally, to further substantiate the effectiveness of our proposed method, we conducted a subjective video quality assessment (VQA) study on five videos with scene switches and fast motion as shown in Table III We invited 14 reviewers, including both professionals and non-professionals with backgrounds in computer vision, to evaluate the enhanced videos according to industry recommendations and the VQA guidelines outlined in [39] using the double stimulus method. Subsequently, we calculated the mean opinion score (MOS) for each video using the method described in [40], [41]. The MOS results in Table III demonstrate that our proposed method outperforms other state-of-the-art video quality enhancement methods. These findings confirm that our method not only enhances video quality at the pixel level, as shown in Table I, but also more closely aligns with human perceptual preferences in terms of visual quality. This highlights the ability of our approach to effectively handle screen content videos featuring scene transitions and dramatic motion, which can significantly influence the Quality of Experience (QoE) for viewers.

**Quality Evaluation on Dramatic Motion and Scene Switches:** To evaluate the capability of our proposed LSFMD in handling dramatic motion and scene switches, two different types of screen content videos were selected to compute the ΔPSNR curves for STDF-R3, QECF, CAT, TGAF, STA, CF-STIF-M, STDR, and our proposed method. The *scprogramming* sequence involves pop-up windows and window switching. These dynamic motions are commonly seen in daily life and can pose difficulties for video quality enhancement algorithms. Additionally, the *scSlideShow* sequence is composed of spliced videos from CTC [34], allowing us to evaluate the performance of our method in scenarios involving abrupt scene transitions. The results are shown in Fig. 9, where dashed lines indicate scene switch frames and gray shadow regions distinguish the frames exhibiting dynamic motion. The result in Fig. 9(a) demonstrates that our proposed LSFMD mostly outperforms the others from frame 38 to frame 63 in the *scprogramming* sequence. This shadow region encompasses window switches and a pop-up window. It can demonstrate that our proposed method can achieve significant ΔPSNR during periods of dramatic motion. While the STA only utilizes the single frame to handle the scene switch which does not perform well in dramatic motion. In Fig. 9(b), frame 28 and frame 44 represent the switch points between two PowerPoint slides in the *scSlideShow* sequence. Notably, our proposed method demonstrates an improvement during most of the transition points, highlighting its effectiveness in handling abrupt scene transitions. In summary, our approach can take the balance between the performance in dynamic content and scene transitions but also proves effective in enhancing the quality of videos with slight motion. This robustness to screen content videos highlights the versatility and reliability of our method.

TABLE III
PERFORMANCE EVALUATION OF MEAN OPTION SCORE (USER STUDY) ACROSS VARIOUS VIDEO ENHANCEMENT METHODS

| Sequences | STDF-R3 | QECF | CAT | TGAF | LSFMD |
|---|---|---|---|---|---|
| ChineseEditing | 50.981 | 40.364 | 39.907 | 48.659 | **61.938** |
| scwebbrowsing | 44.173 | 54.943 | 50.282 | 58.994 | **64.764** |
| EnglishDocumentEditing | 39.741 | 37.361 | 54.422 | 45.547 | **59.476** |
| scmap | 52.067 | 52.233 | 56.432 | 51.794 | **57.057** |
| scprogramming | 54.529 | 48.091 | 51.833 | 45.942 | **57.343** |

**Model Size and Computational Complexity:** Table IV displays the average ΔPSNR in relation to the model parameters and floating point operations (FLOPs) for various methods including LSFMD, STDF-R3, QECF, CAT, and TGAF. These results are averaged over all test sequences. Our RB, MSRB, and MHFD modules in the LSFMD lead to increased consumption of FLOPs and require more parameters, as shown in Table IV. However, these modules are specifically designed to learn contextual information, capture high-frequency details, and efficiently remove redundant hierarchical features, respectively. This is further supported by the results of our ablation study, which will be discussed in the next section. As a result, the performance of LSFMD significantly surpasses other methods, as in Table IV. In addition, our LSFMD is a modular network, allowing for easy adjustment of the model size by varying the number of RB, MSRB, and HFRB blocks.

TABLE IV
COMPARISION OF MODEL SIZE AND COMPUTATIONAL COMPLEXITY

| Model | STDF-R3 | QECF | CAT | TGAF | LSFMD-N2M2 | LSFMD |
|---|---|---|---|---|---|---|
| Δ PSNR (dB) | 0.632 | 0.615 | 0.641 | 0.808 | 0.883 | 0.938 |
| Parameters (KB) | 364.51 | 773.313 | 848.546 | 1403.1 | 1244.029 | 1903.075 |
| FLOPs (G) | 3.856 | 3.292 | 7.541 | 20.344 | 36.383 | 54.449 |

Therefore, in applications with computational limitations, we can use a lightweight structure, such as LSFMD-N2M2, with fewer blocks ($N = 2$, $M = 2$). The LSFMD-N2M2 requires fewer model parameters than TGAF, as shown in Table IV, yet still achieves 0.883 dB ΔPSNR, which is 9.28% higher than that of TGAF (0.808 dB). This highlights the efficiency and effectiveness of our proposed method.

**Quality Enhancement at Different QPs:** To verify the generalization ability of the LSFMD model across different QPs, we conducted additional encoding of all test sequences at QPs of 24, 29, 34, and 39, while training the model at different QPs: QP = 22, 27, 32, and 37. The performance in terms of ΔPSNR is presented in Fig. 10. Fig. 10(a) shows the ΔPSNR of the model trained at QP = 22 and tested at QP = 22 and 24. In Fig. 10(b), the model is trained at QP = 27 and tested at QP = 27 and 29. Similarly, Fig. 10(c) and Fig. 10(d) show ΔPSNR of the model trained at QP =32 and 37, respectively, and tested at different QPs = 32 and 34, 37 and 39. As shown in this figure, each trained model can obtain good quality enhancement on decoded videos at adjacent QPs, thereby verifying the model's generalization ability at various QPs.

**Transfer to Natural Video Domain:** To further demonstrate the generalization capabilities of our proposed LSFMD approach, we retrained it on naturally compressed sequences. To ensure a fair comparison, we used the same dataset and experimental setup as in MFQE 2.0 [13], compressing at QP = 37 to retrain the QECF, CAT, STA, and our proposed LSFMD. The results of MFQE2.0, STDF-R3, TGAF, CF-STIF-M, and STDR are extracted from the original papers. While our method is specifically designed for screen content video, the results demonstrate its commendable in natural domains as well. In Table V, our findings show that the average PSNR/SSIM of the test sequences increased by 0.729 dB/0.01378, surpassing the MFQE 2.0 [13] method, which achieved an increase of 0.562 dB/0.01090. In addition, our approach outperforms other screen content-oriented methods such as QECF and CAT. These findings confirm the robustness of our LSFMD approach.

### C. Ablation Study

In this section, we conducted several ablation experiments on the LSFMD model to analyze its effectiveness in handling scene switches and reconstructing high-frequency details. To evaluate the performance, we present the ΔPSNR curve for frames affected by scene switches and visualize the frame that loses high-frequency information. Other ablation studies are evaluated by calculating the average PSNR improvement across all test sequences.

**Study of Long Short-term Feature Extraction:** As discussed in Section III-B, the long short-term feature extraction

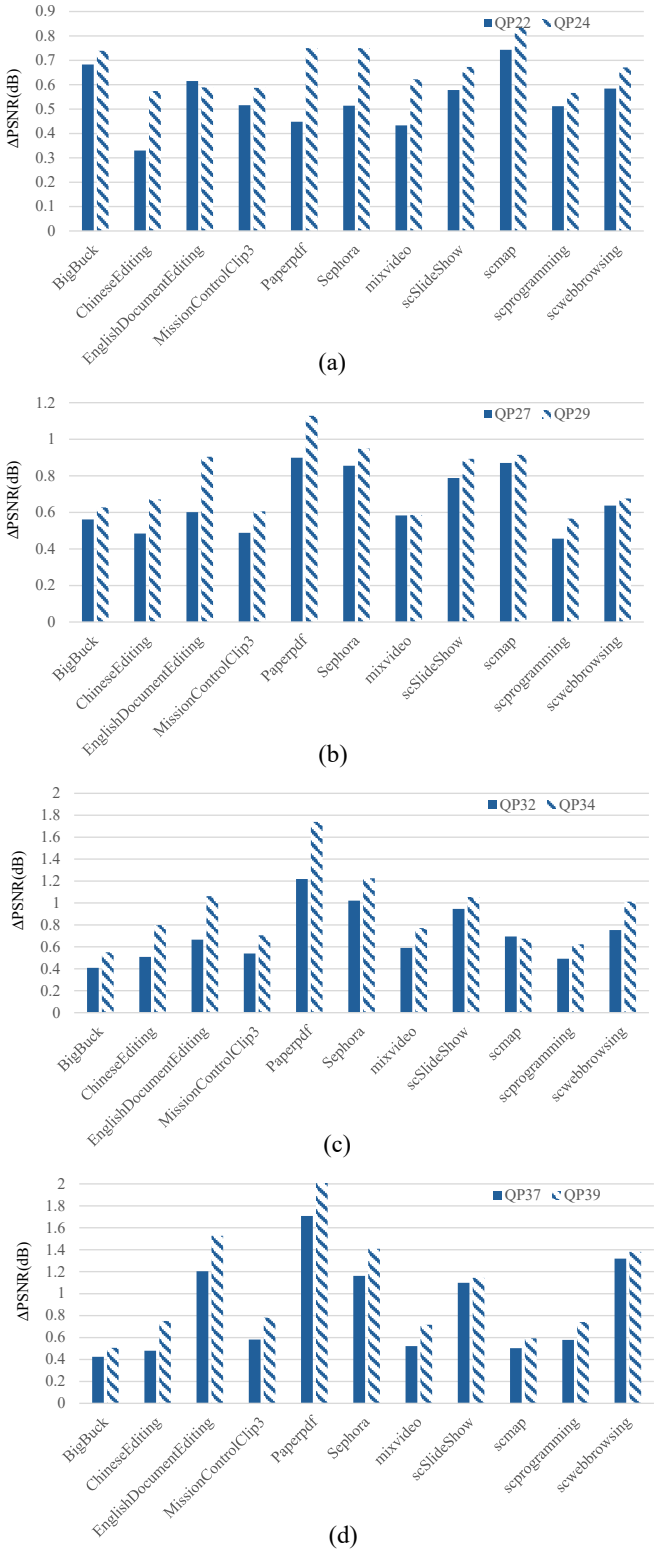



Fig. 10. ΔPSNR of the model trained and tested at different QPs under LDMS configuration. (a) Trained at QP=22, Tested at QP=22 and 24, (b) Trained at QP=27, Tested at QP=27 and 29, (c) Trained at QP=32, Tested at QP=32 and 34, and (d) Trained at QP=37, Tested at QP=37 and 39.

consists of short-term feature extraction, long-term feature extraction, and MHFD. These components can adaptively handle scene switches to achieve better performance in screen

TABLE V
OVERALL ΔPSNR AND ΔSSIM ($\times 10^{-3}$) OF DIFFERENT MODELS AT QP=37 IN MFQE2.0 DATASET # THE RESULT OF EACH SEQUENCE OF CF-STIF-M IS NOT PROVIDED IN THE ORIGINAL PAPER

| Class | Sequences | MFQE2.0 | | CF-STIF-M | | STDF-R3 | | TGAF | | STDR | | STA | | QECF | | CAT | | Proposed LSFMD | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| A | PeopleOnStreet | 0.92 | 15.7 | - | - | 0.65 | 10.4 | 0.84 | 13.2 | 1.53 | 23.4 | 1.22 | 19.6 | 1.16 | 19.1 | 0.89 | 15.8 | 1.22 | 19.8 |
| | Traffic | 0.59 | 10.2 | - | - | 1.18 | 18.2 | 1.53 | 23.3 | 0.85 | 13.4 | 0.67 | 11.5 | 0.60 | 11.2 | 0.51 | 10.4 | 0.66 | 11.3 |
| B | BasketballDrive | 0.47 | 8.3 | - | - | 0.77 | 14.7 | 1.06 | 19.4 | 0.94 | 15.0 | 0.71 | 12.7 | 0.69 | 12.3 | 0.57 | 11.1 | 0.69 | 11.9 |
| | BQTerrace | 0.40 | 6.7 | - | - | 0.54 | 13.2 | 0.71 | 17.9 | 0.72 | 12.3 | 0.56 | 9.7 | 0.53 | 9.3 | 0.42 | 8.1 | 0.54 | 9.3 |
| | Cactus | 0.50 | 10.0 | - | - | 0.7 | 12.3 | 0.82 | 15.2 | 0.85 | 15.2 | 0.67 | 12.3 | 0.61 | 11.9 | 0.45 | 10.1 | 0.66 | 12.1 |
| | Kimono | 0.55 | 11.8 | - | - | 0.58 | 9.3 | 0.68 | 11.6 | 1.05 | 19.1 | 0.77 | 15.4 | 0.68 | 14.3 | 0.48 | 11.9 | 0.73 | 14.6 |
| | ParkScene | 0.46 | 12.3 | - | - | 0.66 | 10.7 | 0.92 | 15.0 | 0.70 | 16.9 | 0.50 | 12.8 | 0.42 | 11.6 | 0.29 | 9.8 | 0.49 | 12.9 |
| C | BasketballDrill | 0.58 | 12.0 | - | - | 0.48 | 10.9 | 0.64 | 16.4 | 0.99 | 18.9 | 0.77 | 15.4 | 0.73 | 15.0 | 0.61 | 13.9 | 0.70 | 14.4 |
| | BQMall | 0.62 | 12.0 | - | - | 0.90 | 16.1 | 1.13 | 20.7 | 1.19 | 21.2 | 0.86 | 17.4 | 0.81 | 16.6 | 0.62 | 14.4 | 0.84 | 16.7 |
| | PartyScene | 0.36 | 11.8 | - | - | 0.60 | 16.0 | 0.81 | 22.9 | 0.79 | 22.4 | 0.59 | 17.3 | 0.45 | 14.1 | 0.39 | 13.8 | 0.58 | 16.5 |
| | RaceHorses | 0.39 | 8.0 | - | - | 0.70 | 12.6 | 0.89 | 17.1 | 0.55 | 15.3 | 0.47 | 11.5 | 0.39 | 10.4 | 0.30 | 8.5 | 0.36 | 9.4 |
| D | BasketballPass | 0.73 | 15.5 | - | - | 0.73 | 17.5 | 0.97 | 24.8 | 1.26 | 25.1 | 0.93 | 19.3 | 0.83 | 17.3 | 0.64 | 14.5 | 0.93 | 18.9 |
| | BlowingBubbles | 0.53 | 17.0 | - | - | 0.91 | 11.3 | 1.22 | 15.9 | 0.86 | 26.7 | 0.68 | 21.6 | 0.57 | 19.2 | 0.49 | 17.6 | 0.64 | 19.7 |
| | BQSquare | 0.34 | 6.5 | - | - | 0.68 | 19.6 | 0.83 | 26.3 | 1.28 | 17.2 | 0.90 | 12.5 | 0.43 | 7.7 | 0.61 | 9.5 | 0.82 | 12.5 |
| | RaceHorses | 0.59 | 14.3 | - | - | 0.95 | 18.2 | 1.23 | 25.1 | 0.95 | 24.4 | 0.70 | 17.6 | 0.60 | 15.9 | 0.46 | 12.4 | 0.64 | 16.7 |
| E | FourPeople | 0.73 | 9.5 | - | - | 0.92 | 10.7 | 1.02 | 13.0 | 1.12 | 13.7 | 1.01 | 12.9 | 0.90 | 12.2 | 0.84 | 11.8 | 0.92 | 12.4 |
| | Johnny | 0.60 | 6.8 | - | - | 0.69 | 7.3 | 0.83 | 8.9 | 0.89 | 9.8 | 0.84 | 9.2 | 0.81 | 9.3 | 0.73 | 9.4 | 0.77 | 8.8 |
| | KristenAndSara | 0.75 | 8.5 | - | - | 0.94 | 8.9 | 1.11 | 11.3 | 1.18 | 11.4 | 1.01 | 10.7 | 0.94 | 10.1 | 0.78 | 9.7 | 0.95 | 10.3 |
| Average | | 0.56 | 10.9 | 0.89 | 15.9 | 0.75 | 13.2 | 0.96 | 17.7 | 0.98 | 17.9 | 0.77 | 14.4 | 0.68 | 13.2 | 0.56 | 11.8 | 0.73 | 13.8 |

TABLE VI
COMPARISONS OF DIFFERENT STRUCTURES IN OUR PROPOSED LSFMD AT QP=37

| Structure | A | B | C | D | E | **F** | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SNFS | ✓ | ✓ | ✓ | ✓ | – | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| The number "N" of MSRBs in short term feature extraction | 3 | 3 | 3 | – | 3 | **3** | 2 | 3 | 3 | 4 | 4 |
| The number "N" of RBs in long term feature extraction | 3 | 3 | – | 3 | 3 | **3** | 2 | 3 | 3 | 4 | 4 |
| MHFD | – | – | – | – | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Hierarchical feature concatenation | – | ✓ | – | – | – | – | – | – | – | – | – |
| The number "M" of HFRBs in high-frequency reconstruction | 3 | 3 | 3 | 3 | 3 | **3** | 2 | 2 | 4 | 3 | 4 |
| ΔPSNR (dB) | 0.899 | 0.899 | 0.839 | 0.727 | 0.930 | 0.938 | 0.883 | 0.903 | 0.897 | 0.945 | 0.933 |
| Parameters (KB) | 1783.825 | 1790.689 | 1780.177 | 569.521 | 1903.075 | 1903.075 | 1244.029 | 1799.347 | 2006.803 | 2458.969 | 2562.697 |
| Time consumption (ms) | 482.913 | 493.575 | 167.801 | 382.764 | 508.190 | 521.824 | 411.629 | 530.556 | 546.855 | 705.566 | 718.751 |

content videos. To verify the effectiveness of these structures, we remove the short-term feature extraction stream, long-term feature extraction stream, or MHFD from LSFMD. The ablation results are shown in Table VI. We also compare the overall time consumption for enhancing a single frame at $1280 \times 720$ resolution using different structures in this table. When we remove the MHFD, as shown in Fig. 5(a), the "Structure A" column in Table VI reveals a ΔPSNR loss of approximately 0.039 dB compared to our method. This indicates that the inclusion of MHFD improves the performance of our model. Furthermore, we also note that the inclusion of MHFD adds 38.911ms to the overall time consumption for enhancing a single frame at $1280 \times 720$ resolution, as shown in the "Time consumption" column of the table. This indicates that the local-global channel attention effectively balances time consumption and performance, further illustrating our model's efficiency. We also compare MHFD with the hierarchical feature utilization methods mentioned in Fig. 5(b) and presented the result in the "Structure B" column, demonstrating a ΔPSNR drop of about 0.039 dB. This suggests that distilling the useful hierarchical features makes our model pay more attention to the features of the target frame. The "Structure C" and "Structure D" demonstrate the results of using only short-term feature extraction and long-term feature extraction, respectively. We observe a significant drop in ΔPSNR, which clarifies the importance of the combination of these two streams.

To further validate that our modules meet their design objectives, we visualize the extracted features from Fig. 11(a) when different modules are adopted. The feature maps from the short-term feature extraction module, highlighted in Fig. 11(b) and Fig. 11(d), primarily focus on high-frequency information of the target frame. On the other hand, the long-term feature extraction module captures more extensive features from neighbor frames, as illustrated in Fig. 11(c) and Fig. 11(e). It is worth noting that in the region highlighted by the red rectangle in Fig. 11(c), we can see the features from the neighbor frame are also introduced. This observation verifies that our SNFS in the short-term feature extraction stream ensures that short-term information is extracted from frames with similar content, enhancing the accuracy of the reconstruction. After the SNFS, the MSRB captures high-frequency details associated with text edges, preserving the sharpness and clarity of the text in the reconstructed frame, as we claim. Compared to the features extracted from the short-term stream, the long-term feature extraction integrates information from neighbor frames, enriching the feature set and maintaining the integrity of text and graphics across consecutive frames, as detailed in Section III-C.

Therefore, our proposed MHFD effectively leverages these insights by combining the advantages of both long-term and short-term feature extractions. This successful integration is demonstrated in Fig. 11(f), where the region corresponding to the red rectangle in Fig. 11(f) shows more relevance to the target frame than the correlated region in Fig. 11(c). This observation further verifies that MHFD integrates both low and high-frequency features while filtering redundant features of neighbor frames. This balanced feature integration enhances the overall effectiveness of our approach to video quality enhancement.

In summary, the ablation study highlights the effectiveness of the long short-term feature extraction components, includ-

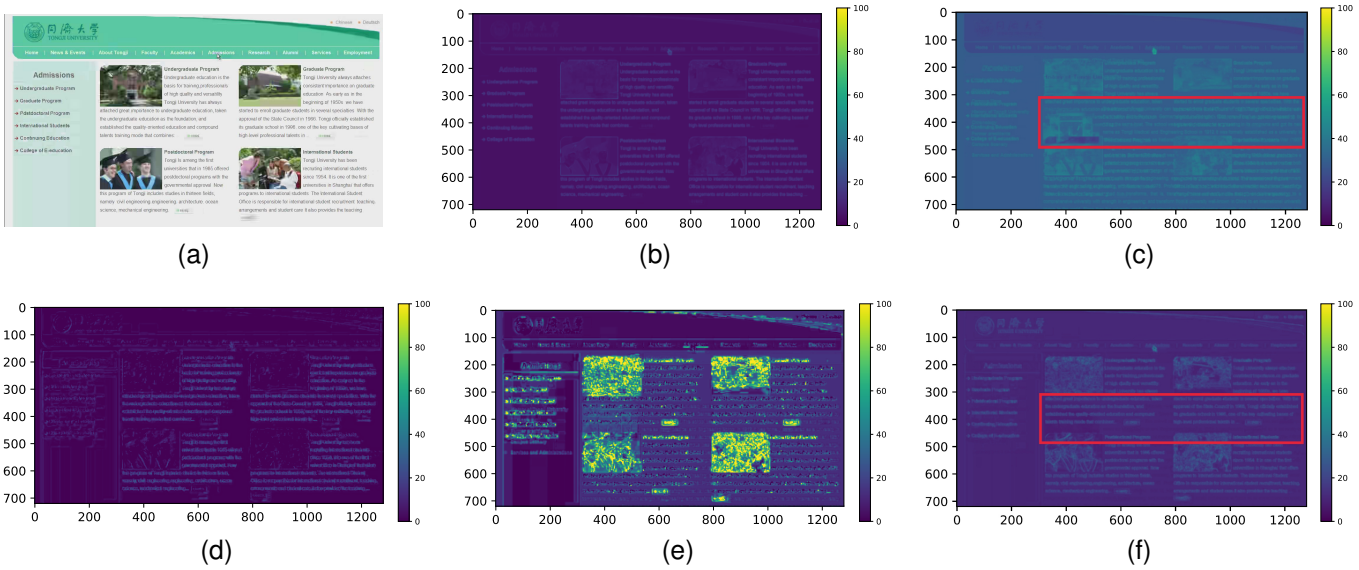



Fig. 11. Visualization of feature maps produced by different modules of the proposed LSFMD. (a) Enhanced frame of our proposed LSFMD, (b) feature map $F_{st}^{0}$ in short-term feature extraction, (c) feature map $F_{lt}^{0}$ in long-term feature extraction, (d) feature map $F_{st}^{N}$ in short-term feature extraction, (e) feature map $F_{lt}^{N}$ in long-term feature extraction, and (f) feature map $Q$ of MHFD.

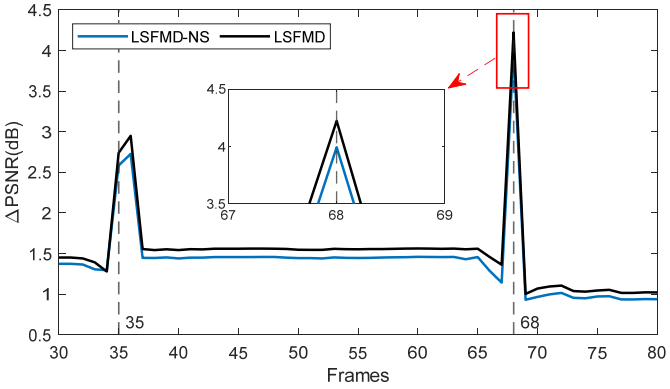



Fig. 12. Δ PSNR curves of screen content video *scwebbrowsing*.


Fig. 13. Subjective visual quality comparison at QP37 on *scmap*.

ing short-term feature extraction, long-term feature extraction, and MHFD, in achieving better performance for screen content videos.

**Study of SNFS:** As discussed in Section III-B, the SNFS enables the model to adaptively handle scene switches and enhances performance in screen content videos. To verify the effectiveness of the SNFS, we remove the SNFS from LSFMD. The ablation results are shown in Table VI and Fig. 12, where dashed lines indicate scene switch points. In the "Structure E" column of Table VI, it is evident that incorporating the SNFS adds an additional 13.634ms for improving each frame. However, this increase in processing time is considered acceptable given the improvement in quality it delivers. In Fig. 12, "LSFMD-NS" represents the LSFMD model without SNFS. The results reveal that the LSFMD-NS model experiences a slight decrease in PSNR during the scene switch compared to our proposed LSFMD model. This means that using the similarity frame to extract short-term information can further improve the quality of frames in the presence of scene switches. In other words, the SNFS component plays a crucial role in the LSFMD model's ability to adaptively handle scene switches and maintain high-quality performance. By using the similarity frame, the SNFS helps the model better extract short-term information, leading to improved reconstruction quality during scene transitions.

**Study of High-frequency Reconstruction Block (HFRB):** To verify the efficiency of our proposed HFRB in restoring high-frequency information in the target frame, we conducted a visual analysis of the high-frequency details in a frame from the "scmap" sequence. As illustrated in Fig. 13, the model labeled "LSFMD-NH" represents the LSFMD model without the incorporation of the HFRB. The absence of the HFRB in this model leads to a noticeable blurriness in the text, underscoring the importance of high-frequency detail preservation for maintaining text clarity and overall image sharpness. This comparison highlights the critical role of the HFRB in enhancing the visual quality of the reconstructed frames. The HFRB effectively restores the fine details that are often lost during the compression process, resulting in sharper and clearer images, especially in the text regions. To examine how the number of HFRB blocks affects performance, we varied the quantity of these blocks. The outcomes of these adjustments are detailed in columns "Structure F", "Structure H", and "Structure I" of Table VI. The results indicate that

the optimal number of HFRB blocks is "3".

**Influence of the Number of Blocks:** The LSFMD features a modular network design that facilitates simple tuning of the model size through the adjustment of MSRB, RB, and HFRB block quantities. From columns "Structure F" to "Structure K" in Table VI, we observe that increasing the number of these blocks significantly enhances the PSNR gain. However, beyond a certain depth, the performance begins to decline. An excessive number of blocks not only hampers training but also leads to the loss of useful information. To strike a balance between performance and model size, we set $N = 3$ and $M = 3$ in the final LSFMD model.

This modular architecture enables fine-tuning of network complexity to achieve the desired performance-complexity trade-off. For instance, in Table IV, the LSFMD with fewer blocks ($N = 2$, $M = 2$) requires fewer model parameters than TGAF, but still outperforms TGAF in terms of $\Delta$PSNR. This design flexibility ensures that the LSFMD can be optimized for different application scenarios and computational constraints, making it a versatile and adaptable solution for screen content video reconstruction.

## V. Conclusion

In this paper, we propose a novel method tailored for handling scene switches and reconstructing high-frequency information in screen content videos. Our approach includes a long short-term feature extraction module, consisting of three components: the long-term feature extraction stream, which learns contextual information; the short-term feature extraction stream, which selects relevant features from shorter inputs to better manage fast motion and scene switches; and the multi-scale feature distillation mechanism, which adaptive fuse the short-term and long-term features. Meanwhile, we introduce the SNFS into the short-term feature stream to further enhance the quality of scene switch frames. In the reconstruction phase, we propose the HFRB, which guides the model to focus on restoring high-frequency components. This is crucial for preserving the sharpness and clarity of text and other fine details in screen content videos. The novel contributions of our work, including the modular feature extraction module, the SNFS mechanism, and the HFRB, have collectively led to substantial improvements in screen content video reconstruction quality. Experimental results demonstrate that our proposed LSFMD significantly enhances the quality of compressed videos, surpassing the current state-of-the-art methods. Moreover, we conduct thorough ablation studies to verify the effectiveness of the designed network structure and its individual components. Currently, the high computational demands of our proposed method may limit its suitability for resource-constrained devices. Moving forward, we plan to explore the use of teacher-student techniques [42], [43] to fine-tune a more lightweight version of the model. This strategy has the potential to significantly reduce computational complexity by allowing the lightweight student model to learn from a more complex teacher model, distilling essential knowledge into a simpler form that requires fewer computational resources. Additionally, we aim to further enhance computational efficiency through model compression techniques such as pruning and quantization [44]. To improve the model's generalizability, we will also broaden our dataset to include a diverse range of screen content videos, dynamic scenarios, and varied noise patterns. To optimize the model for real-time applications, our focus will be on minimizing frame dependencies, utilizing parallel processing technologies like GPUs and FPGAs, and deploying adaptive complexity mechanisms that can adapt to varying hardware specifications and content types. By incorporating these ideas, we aim to enhance the model's suitability for real-time applications, thereby extending its practical use across various smart devices.

## References


[1] J. Xu, R. Joshi, and R. A. Cohen, "Overview of the emerging HEVC screen content coding extension," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 50–62, 2015.

[2] J. Chen, J. Ou, H. Zeng, and C. Cai, "A fast algorithm based on gray level co-occurrence matrix and gabor feature for HEVC screen content coding," *Journal of Visual Communication and Image Representation*, vol. 78, p. 103128, 2021.

[3] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[4] J. Lainema, F. Bossen, W.-J. Han, J. Min, and K. Ugur, "Intra coding of the HEVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1792–1801, 2012.

[5] Z. Ma, W. Wang, M. Xu, and H. Yu, "Advanced screen content coding using color table and index map," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4399–4412, 2014.

[6] X. Xu, S. Liu, T.-D. Chuang, Y.-W. Huang, S.-M. Lei, K. Rapaka, C. Pang, V. Seregin, Y.-K. Wang, and M. Karczewicz, "Intra block copy in HEVC screen content coding extensions," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 6, no. 4, pp. 409–419, 2016.

[7] W.-S. Park and M. Kim, "CNN-based in-loop filtering for coding efficiency improvement," in *2016 IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE, 2016, pp. 1–5.

[8] Y. Dai, D. Liu, and F. Wu, "A convolutional neural network approach for post-processing in HEVC intra coding," in *MultiMedia Modeling: 23rd International Conference, MMM 2017, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part I 23*. Springer, 2017, pp. 28–39.

[9] T. Wang, M. Chen, and H. Chao, "A novel deep learning-based method of improving coding efficiency from the decoder-end for HEVC," in *2017 Data Compression Conference (DCC)*. IEEE, 2017, pp. 410–419.

[10] S. Kuanar, C. Conly, and K. Rao, "Deep learning based HEVC in-loop filtering for decoder quality enhancement," in *2018 Picture Coding Symposium (PCS)*. IEEE, 2018, pp. 164–168.

[11] R. Yang, M. Xu, and Z. Wang, "Decoder-side HEVC quality enhancement with scalable convolutional neural network," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 817–822.

[12] R. Yang, M. Xu, Z. Wang, and T. Li, "Multi-frame quality enhancement for compressed video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6664–6673.

[13] Z. Guan, Q. Xing, M. Xu, R. Yang, T. Liu, and Z. Wang, "MFQE 2.0: A new approach for multi-frame quality enhancement on compressed video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 949–963, 2019.

[14] J. Deng, L. Wang, S. Pu, and C. Zhuo, "Spatio-temporal deformable convolution for compressed video quality enhancement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, 2020, pp. 10 696–10 703.

[15] D. Luo, M. Ye, S. Li, C. Zhu, and X. Li, "Spatio-temporal detail information retrieval for compressed video quality enhancement," *IEEE Transactions on Multimedia*, vol. 25, pp. 6808–6820, 2022.

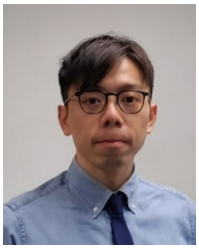
[16] D. Luo, M. Ye, S. Li, and X. Li, "Coarse-to-fine spatio-temporal information fusion for compressed video quality enhancement," *IEEE Signal Processing Letters*, vol. 29, pp. 543–547, 2022.

[17] J. Huang, J. Cui, M. Ye, S. Li, and Y. Zhao, "Quality enhancement of compressed screen content video by cross-frame information fusion," *Neurocomputing*, vol. 493, pp. 486–496, 2022.
[18] T. Lindeberg, *Scale-space theory in computer vision*. Springer Science & Business Media, 2013, vol. 256.
[19] Y. Chen, H. Fan, B. Xu, Z. Yan, Y. Kalantidis, M. Rohrbach, S. Yan, and J. Feng, "Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3435–3444.
[20] R. Yang, M. Xu, T. Liu, Z. Wang, and Z. Guan, "Enhancing quality for HEVC compressed videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 7, pp. 2039–2054, 2018.
[21] W. Lin, X. He, X. Han, D. Liu, J. See, J. Zou, H. Xiong, and F. Wu, "Partition-aware adaptive switching neural networks for post-processing in HEVC," *IEEE Transactions on Multimedia*, vol. 22, no. 11, pp. 2749–2763, 2019.
[22] Z. Huang, Y.-L. Chan, S.-H. Tsang, and K.-M. Lam, "Mode information guided CNN for quality enhancement of screen content coding," *IEEE Access*, vol. 11, pp. 24 149–24 161, 2023.
[23] Q. Zhu, J. Hao, Y. Ding, Y. Liu, Q. Mo, M. Sun, C. Zhou, and S. Zhu, "Cpga: Coding priors-guided aggregation network for compressed video quality enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2964–2974.
[24] Y. Liu, M. Ye, Y. Gao, S. Li, Y. Zhao, and X. Li, "Content adaptive compressed screen content video quality enhancement," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022, pp. 01–06.
[25] C. Shu, M. Ye, H. Guo, and X. Li, "Spatial-temporal adaptive compressed screen content video quality enhancement," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 71, no. 5, pp. 2884–2888, 2024.
[26] Q. Zhu, Y. Qiu, Y. Liu, S. Zhu, and B. Zeng, "Compressed video quality enhancement with temporal group alignment and fusion," *IEEE Signal Processing Letters*, vol. 31, pp. 1565–1569, 2024.
[27] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 517–532.
[28] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Noise Reduction in Speech Processing*, pp. 1–4, 2009.
[29] Y. Liu, J. Wu, L. Li, W. Dong, J. Zhang, and G. Shi, "Spatiotemporal representation learning for blind video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3500–3513, 2022.
[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
[31] K. Fukushima, "Visual feature extraction by a multilayered network of analog threshold elements," *IEEE Transactions on Systems Science and Cybernetics*, vol. 5, no. 4, pp. 322–333, 1969.
[32] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Fast and accurate image super-resolution with deep laplacian pyramid networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 11, pp. 2599–2613, 2018.
[33] P. Yi, Z. Wang, K. Jiang, Z. Shao, and J. Ma, "Multi-temporal ultra dense memory network for video super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 8, pp. 2503–2516, 2020.
[34] K. Sharman and K. Suehring, "Common test conditions, document jctvc-z1100," *Geneva, Switzerland*, 2016.
[35] S.-H. Tsang, Y.-L. Chan, and W. Kuang, "Mode skipping for HEVC screen content coding via random forest," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2433–2446, 2019.
[36] JCT-VC, "Screen content sequences," 2015. [Online]. Available: ftp://mpeg.tnt.uni-hannover.de/testsequences/
[37] Z. Huang, Y.-L. Chan, S.-H. Tsang, N.-W. Kwong, K.-M. Lam, and W.-K. Ling, "Spatio-temporal feature learning for enhancing video quality based on screen content characteristics," *Journal of Visual Communication and Image Representation*, vol. 104, p. 104270, 2024.
[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[39] V. Q. E. Group *et al.*, "Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment," 2000. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI
[40] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.
[41] Q. Jiang, Z. Liu, K. Gu, F. Shao, X. Zhang, H. Liu, and W. Lin, "Single image super-resolution quality assessment: a real-world dataset, subjective studies, and an objective metric," *IEEE Transactions on Image Processing*, vol. 31, pp. 2279–2294, 2022.
[42] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3048–3068, 2021.
[43] J. Zhou, B. Zhang, D. Zhang, G. Vivone, and Q. Jiang, "Dtkd-net: Dual-teacher knowledge distillation lightweight network for water-related optics image enhancement," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024.
[44] A. Kuzmin, M. Nagel, M. Van Baalen, A. Behboodi, and T. Blankevoort, "Pruning vs quantization: Which is better?" *Advances in neural information processing systems*, vol. 36, pp. 62 414–62 427, 2023.

**Ziyin Huang** received the M.Sc. from The Guangdong University of Technology, China, in 2020. She is currently pursuing the Ph.D. degree in the Digital Signal Processing Laboratory, The Hong Kong Polytechnic University, Hong Kong. Her current research interests include computer vision, deep learning, and video enhancement.

**Yui-Lam Chan** (S'94–A'97–M'00) received the B.Eng. (Hons.) and Ph.D. degrees from The Hong Kong Polytechnic University, Hong Kong, in 1993 and 1997, respectively. He joined The Hong Kong Polytechnic University in 1997, where he is currently an Associate Professor with the Department of Electrical and Electronic Engineering. He is actively involved in professional activities. He has authored over 160 research papers in various international journals and conferences. His research interests include multimedia technologies, signal processing, image and video compression, video quality enhancement and assessment, video transcoding, video conferencing, digital TV/HDTV, 3DTV/3DV, multiview video coding, machine learning for video coding and processing, and future video coding standards including screen content coding, light-field video coding, and 360-degree omnidirectional video coding. Dr. Chan served as an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING. He was the Secretary of the 2010 IEEE International Conference on Image Processing. He was also the Special and Demo Sessions Co-Chairs, IEEE International Conference on Visual Communications and Image Processing, the Publications Chairs of the IEEE International Conference on Multimedia and Expo.

**Ngai-Wing Kwong** received the Ph.D. degree from The Hong Kong Polytechnic University (PolyU), Hong Kong, in 2024. Currently, he is a Postdoctoral Fellow at The Hong Kong Polytechnic University. His research interests include Computer Vision (CV), Video Quality Assessment (VQA), and Image, Video, and Signal processing using artificial intelligence (AI) and deep learning.



**Sik-Ho Tsang** received the Ph.D. degree from The Hong Kong Polytechnic University (PolyU), Hong Kong, in 2013. He is an Assistant Professor in Hong Kong Chu Hai College. His research interests involve computer vision (CV), natural language processing (NLP), and acoustic signal processing using artificial intelligence (AI) and deep learning. He is a reviewer of international journals including the IEEE Transactions on Image Processing, IEEE Transactions on Broadcasting, IEEE Transactions on Circuits and Systems for Video Technology, and IEEE Access.

**Kin-Man Lam** received his Associateship in Electronic Engineering with distinction from The Hong Kong Polytechnic University (formerly called Hong Kong Polytechnic) in 1986, his M.Sc. degree in communication engineering from the Department of Electrical Engineering, Imperial College, U.K., in 1987, and his Ph.D. degree from the Department of Electrical Engineering, University of Sydney, Australia, in 1996. From 1990 to 1993, he was a lecturer at the Department of Electronic Engineering of The Hong Kong Polytechnic University. He joined the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University again as an Assistant Professor in October 1996. He became an Associate Professor in 1999, and has been a Professor since 2010. He was actively involved in professional activities. He has been a member of the organizing committee or program committee of many international conferences. Prof. Lam was the Chairman of the IEEE Hong Kong Chapter of Signal Processing between 2006 and 2008, and was the Director-Student Services and the Director-Membership Services of the IEEE SPS between 2012 and 2014, and between 2015 and 2017, respectively. He was also the VP-Member Relations and Development and VP-Publications of the Asia-Pacific Signal and Information Processing Association (APSIPA) between 2014 and 2017, and between 2017 and 2021, respectively. He was an Associate Editor of IEEE Trans. on Image Processing between 2009 and 2014, and Digital Signal Processing between 2014 and 2018. He was also an Editor of HKIE Transactions between 2013 and 2018, and an Area Editor of the IEEE Signal Processing Magazine between 2015 and 2017. Currently, he is the the IEEE SPS VP-Membership. Prof. Lam also serves as a Senior Editorial Board member of APSIPA Trans. on Signal and Information Processing, and an Associate editor of EURASIP International Journal on Image and Video Processing. His current research interests include image and video processing, computer vision, and human face analysis and recognition.

**Wing-Kuen Ling** received the B.Eng. (Hons) and M.Phil. degrees from the department of Electronic and Computer Engineering, the Hong Kong University of Science and Technology, in 1997 and 2000, respectively, and the Ph.D. degree in the department of Electronic and Information Engineering from the Hong Kong Polytechnic University in 2003. In 2004, he joined the King's College London as a Lecturer. In 2010, he joined the University of Lincoln as a Principal Lecturer and promoted to a Reader in 2011. In 2012, he joined the Guangdong University of Technology as a Full Professor. He is a Fellow of the IET, a senior member of the IEEE, a China National Young Thousand-People-Plan Distinguished Professor, Guangdong Province Pearl Scholar and University Hundred-People-Plan Distinguished Professor. He serves in the nonlinear circuits and systems technical committee, the digital signal processing technical committee and the power and energy for circuits and systems technical committee of the IEEE Circuits and Systems Community, as well as the cloud and wireless systems for industrial applications technical committee, the industrial informatics technical committee and the industrial building automation, control and management technical committee of the IEEE Industrial Electronics Society. He was awarded the best reviewer prizes from the IEEE Instrumentation and Measurement Society in 2008 and 2012. He has also served as the guest editor-in-chief of several special issues of highly rated international journals, such as the IET Signal Processing, the Circuits, Systems and Signal Processing, the HKIE Transactions, the Sensors, and the American Journal of Engineering and Applied Sciences. He is currently an associate editor of the IEEE Transactions on Consumer Electronics, the IET Signal Processing, the Circuits, Systems and Signal Processing, the Journal of Franklin Institute, the Measurement, the Measurement: Sensors, the Journal of Industrial Management, the Frontiers in Signal Processing, and the Signal and Information Processing. He has published an undergraduate textbook, a research monograph, five book chapters, 250+ internationally leading journal papers and 150+ highly rated international conference papers as well as owned 70+ China patents. His research interests include the time frequency analysis, the optimization theory, the symbolic dynamics, the biomedical signal processing and the multimedia signal processing.