IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 November 2023, accepted 18 November 2023, date of publication 21 November 2023,
date of current version 29 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3335352

APPLIED RESEARCH

Video Quality Assessment System Using Deep Optical Flow and Fourier Property

DONGGOO KANG[1], YEONGJOON KIM[2], SUNKYU KWON[2], HYUNCHEOL KIM[3], JINAH KIM[3], AND JOONKI PAIK[1,2], (Senior Member, IEEE)
[1]Department of Image, Chung-Ang University, Seoul 06974, South Korea
[2]Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, South Korea
[3]4by4 Inc., Seoul 06541, South Korea

Corresponding author: Joonki Paik (paikj@cau.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) funded by Korean Government for the Ministry of Science & Information and Communications Technology (MSIT) through the Artificial Intelligent Graduate School Program, Chung-Ang University, under Grant 2021-0-01341; and in part by the Field-Oriented Technology Development Project for Customs Administration through National Research Foundation of Korea (NRF) funded by MSIT and Korea Customs Service under Grant 2021M3I1A1097911.

ABSTRACT Ensuring superior video quality is essential in various fields such as VFX film production, digital signage, media facades, product advertising, and interactive media, as it directly elevates the viewer's engagement and experience. The ability to accurately quantify a video's visual quality not only influences its valuation but is pivotal in maintaining high standards. Among the attributes influencing video quality, subjective quality stands out, however, several other elements also contribute significantly. Although automated video evaluations offer efficiency, there are situations necessitating expert editorial insight to measure nuanced subjective attributes. Our research primarily focuses on two prevalent issues undermining video quality: erratic camera motions and suboptimal focus. We employed a deep learning-driven optical flow technique to quantify inconsistent camera movements and adopted a Fast Fourier Transform (FFT)-based algorithm for blur detection. Moreover, our proposed adaptive threshold, grounded in statistical analysis, effectively delineates scenes as either desirable or substandard. Testing this framework on a diverse set of videos, we found it proficiently assessed video quality within a practical threshold range.

INDEX TERMS Deep learning, computational photography, image quality assessment, video stabilization, optical flow.

I. INTRODUCTION

This Video quality assessment (VQA) is the process of evaluating the perceived visual quality of a video. It is a multidisciplinary field that combines techniques from image processing, computer vision, human vision, and statistics. The primary objective of VQA is to measure the quality of a video in a manner that is perceptually relevant to humans. With the increasing availability of high-definition video content and the growing demand for video streaming services, the need for accurate and efficient VQA methods has become increasingly important.

Video quality assessment involves evaluating various factors such as resolution, frame rate, compression, and color accuracy. However, the most crucial factor in determining video quality is subjective quality, which is based on the viewer's perception of how the video appears and feels. This encompasses elements such as sharpness, noise, and overall visual appeal. There are several works that design features in consideration of these factors [1], [2], [3], [4], [5], [6]. These works focused on the specific problem, such as undesired blur [2], [3], [7], [8] and noise [5], [6], [9]. Some approaches propose a system that integrates multiple methods [10] A more comprehensive explanation of hand-crafted feature-based VQA can be found in [11].

Recently, deep learning-based VQA methods are presented to estimate quality score [13], [14], [15], [16], [17], [18], [19], [20]. Learning-based methods are more effective and robust than a hand-crafted feature on the given dataset distribution. However, there are certain limitations to using datasets

The associate editor coordinating the review of this manuscript and approving it for publication was Shovan Barma.

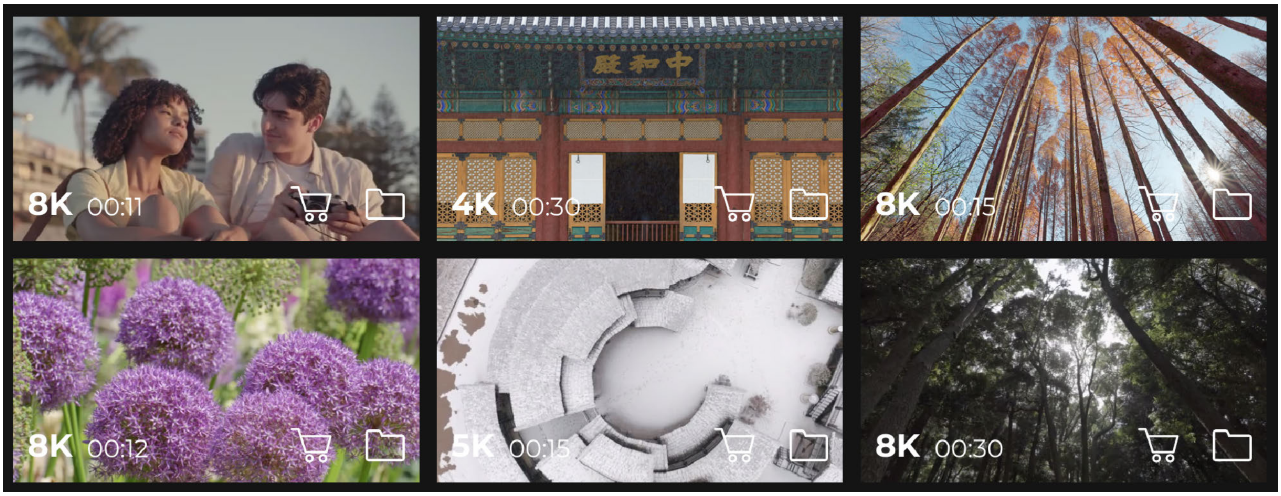


FIGURE 1. An example of actual video clips available for purchase on a commercial site is KEYCUTstock [12]. The clips offered on this site are captured from various domains and are predominantly edited as short clips.

collected by specific groups, as they may not represent the entire distribution of videos found in the world. Additionally, the process of labeling videos can be both expensive and time-consuming.

TABLE 1. Bad clips collected during the actual editing process.

	Undesired Camera Movement	Incomplete Focus	ETC(camera error)
Clips	2,587	1,019	93
Ratio	**69.94**%	27.55%	2.51%

The qualitative nature of video content often makes it the subject of subjective assessments. Typically, this involves a cohort of viewers grading a video based on their personal preferences, usually on a scale ranging from 1 to 5 or 1 to 10. Such evaluations offer an in-depth understanding of video quality, taking into account individual biases and preferences, which are vital since they ultimately dictate viewer satisfaction and engagement.

However, there are inherent challenges. Manual evaluations are labor-intensive, time-consuming, and expensive. Solely relying on algorithmic predictions for video quality often results in a mismatch with genuine human perceptions, especially when dealing with prevalent issues like camera instability or focus discrepancies. Additionally, the lack of diversity in extant datasets and models poses a hindrance to the broad application of such methods on a variety of videos.

Consequently, our primary challenge is the creation of a comprehensive VQA system that adeptly emulates human assessments, especially concerning prevalent video quality anomalies. We aim to establish a pragmatic VQA methodology optimized for real-world applications by incorporating human validation. A notable flaw in current methods is the cumbersome nature of manual evaluations needed for subjective ground truth labels. Our proposition seeks to remedy this through a streamlined collaborative system that garners pertinent human feedback.

To check what issues are major to solve, we interact with editorial experts. Table 1 and Fig. 1 present the collection of video clips obtained through the editing process of high-resolution videos captured by photographers from around the world [12]. These video clips are classified into three categories of imperfections: undesired camera motion, incomplete focus, and camera errors. Since most of the videos are recorded by a hand-held camera, the major problem that degrades video quality is undesired camera motion. Also, incomplete focus caused by the expert manual mode of the camera is the second problem.

Motivated by observed challenges, we crafted a unique VQA system aimed predominantly at identifying and rectifying unintentional camera movement and imprecise focus. Our two-pronged VQA system comprises: i) video quality estimation and ii) adaptive threshold calculation via an iterative algorithm. Our strategy centers on appraising video quality primarily by probing camera motion and blur attributes. We leverage a sophisticated optical flow network to approximate camera motion and employ the Fourier transform mechanism to gauge the degree of video blurriness. Our tailored blurriness index is primed to detect widespread defocus blur stemming from camera focus errors or subject motion dynamics. Utilizing the ascertained video quality, our iterative algorithm computes an adaptive threshold, enabling us to categorize frames as either satisfactory or suboptimal.

Specifically, we focus on an automated algorithm to assess quality degradation from camera motion and blur. Human experts will validate the results to retain perceptual alignment. In this work, we present such a system using optical flow and Fourier analysis to quantify these common artifacts. We believe this efficient hybrid approach can help overcome key challenges in prevailing objective VQA methods to better match subjective human assessments.

The contributions of this paper are summarized as follows:

- A VQA system that accurately quantifies video quality using a deep optical flow network and Fourier property.

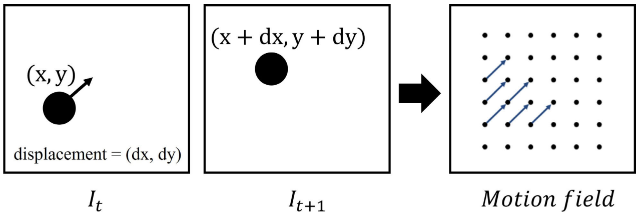


FIGURE 2. **Optical flow in the motion field. In the *t*-th frame of a video, denoted by I_t, a pixel located at coordinates (x, y) undergoes a displacement of (dx, dy), resulting in the new position of $(x + dx, y + dy)$ in the next frame I_{t+1}. A collection of displacement of pixels across consecutive frames is referred to as the motion field.**

- An iterative search algorithm that calculates a video adaptive threshold, which helps to overcome issues caused by a constant threshold approach.
- Extensive experiments that demonstrate the effectiveness of the proposed VQA system with both subjective and objective quality measures showing favorable results.

II. RELATED WORKS

A. CAMERA MOTION

In computer vision algorithms, camera motion estimation involves determining the precise location of specific pixels in the next frame relative to their location in the current frame. By analyzing the changes in pixel position, velocity, and acceleration over time, the camera's motion can be estimated. The following subsection presents three different approaches to measuring camera motion in a video clip.

1) OPTICAL FLOW

Optical flow is a technique that uses the motion of pixels in consecutive frames of a video to estimate the movement of the camera. It can be used to measure the speed and direction of camera movement and the amount of rotation. Optical flow can be applied to the video stabilization property, and this method enables camera movement estimation. Optical flow is the motion of the pixel between consecutive frames, caused by the camera movement or object movement.

As shown in Fig. 2, A collection of displacement of pixels across consecutive frames is referred to as the motion field. An arrow in the motion field is the corresponding optical flow vector, which can estimate the information of the moving pixels between two consecutive frames as:

$$I_t = I(x, y, t), \ D_t = (dx, dy), \tag{1}$$

where (x, y) represents the coordinate of a pixel in the image I_t, t the frame number, and the vector D_t the displacement of the pixel across consecutive frames. From (1), the next frame is formulated as

$$I_{t+1} = I(x + dx, y + dy, t + dt). \tag{2}$$

Representative works to estimate optical flow include Lucas-Kanade and Horn-Schunck methods. They assumed that the flow is essentially constant in a local neighborhood of the pixel, and solved using the least-squares criterion for all the pixels. While these methods show reasonable results under certain constraints, such as small and approximately constant displacement of pixels within a neighborhood, they can be computationally expensive due to the process of finding the best match between corresponding points in adjacent frames. To address this issue, there are feature tracking methods that track sparse pixels to reduce the computational cost and deep learning-based methods that perform vector operations parallel using GPU memory.

Deep-learning based optical flow estimation methods are faster and more accurate than the conventional method but need large datasets to optimize millions of network parameters [21], [22], [23], [24], [25]. collecting optical flow data is challenging as ground-truth motion fields would require annotations for every pixel in a video clip. To overcome this problem, most datasets [21], [26], [27], [28] synthesize motion fields using virtual tools.

2) FEATURE TRACKING

Feature tracking is a technique that uses distinctive features in a video, such as corners or edges, to track the camera's movement. It can be used to determine the camera's position and orientation of the camera in each frame of the video. Representative future detection and tracking methods include Harris Corner Detector [29], Scale Invariant Feature Transform (SIFT) [30], and Speeded-Up Robust Features (SURF) [31]. These methods detect invariant feature points and track feature points at consecutive frames. Based on this measurement, we can compute the affine transformation matrix and use it to stabilize the video. Feature tracking methods are faster and memory efficient than optical flow estimation methods. Also, deep-learning-based methods [32], [33], [34] that learn the fine-grained features using CNNs were introduced recently. However, if feature points are not detected due to motion blur and camera distortion, the predicted consecutive transformation matrix may become unstable.

3) STABILIZATION

Camera stabilization is a technique used to remove unwanted camera movement from a video, and it involves measuring the amount and type of camera movement present in the video. Most of works use a combination of feature tracking and optical flow methods for camera stabilization [35], [36], [37], [38]. Deep learning-based methods also use these methods as prior knowledge and constraints.

B. BLUR DETECTION

Blur detection is a technique used to determine whether an image or video is blurred or not. Blurry images or videos have edges and details that are not sharp and clear, and the overall image appears to be out of focus. Blur detection is crucial in a variety of applications, including image and video processing, computer vision, and computational photography.

Fourier transform-based blur detection is a method for detecting blur in an image or video by analyzing the frequency content of the image [3], [4], [7]. The Fourier transform is a mathematical technique that decomposes a signal into its constituent frequencies, which can then be analyzed to determine the frequency content of the signal. In the case of blur detection, the Fourier transform is applied to the image to convert it into the frequency domain. Once in the frequency domain, the image is analyzed to determine the frequency content and the distribution of the energy across different frequencies.

Blur detection is also a challenging task because the blur can be caused by different factors, such as camera shake, low light, compression artifacts, and user-intentional out-of-focus. Actually, the last mentioned factor is the most difficult case in that It should consider the photographer's intention. Tang proposed a deep learning-based segmentation method, DefocusNet, for detecting out-of-focus regions [39]. This method requires a segmentation mask and generates a blur on the background region. While it shows favorable results, it only optimizes background region blur and may not work well for images with different distributions than those in the training datasets.

III. PROPOSED METHOD

A. SYSTEM OVERVIEW

The proposed VQA system consists of 3-steps: i) blurriness estimation using the Fourier property, ii) camera movement estimation using the video stabilization method, and iii) calculation of statistical adaptive threshold. Fig. 3 shows an overview of the proposed VQA system. To estimate blurriness, we adopt the Fourier transform and calculate their high-frequency portion. We also estimate camera motion from optical flow obtained using a deep learning-based neural network to estimate camera motion.

Based on the estimated blurriness and camera motion, we compute the video adaptive threshold using iterative search and statistical algorithms. In the following subsections, we will provide a step-by-step detailed explanation of the three steps.

B. BLURRINESS ESTIMATION USING THE FOURIER PROPERTY

The two-dimensional (2D) Fourier transform is a mathematical technique that is used to decompose an image into its constituent frequencies. It is used to analyze the frequency content of an image, which can be useful for various image processing tasks, such as image enhancement, filtering, and compression. The 2-D Fourier transform is applied to an image by converting the image into the frequency domain, which represents the image in terms of its constituent frequencies rather than its spatial coordinates.

As a theoretical basis, the one-dimensional (1D) Fourier transform is calculated using the following formula:

$$F(u) = \int_{-\infty}^{\infty} f(x)e^{-2j\pi ux}dx, \tag{3}$$

where $F(u)$ represents the transformed signal in the frequency domain, $f(x)$ the input 1D signal, and u the frequency coordinates. Applying the Fourier transform to 2D image input is also possible using the following formula:

$$F(u, v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)e^{-2j\pi(ux+vy)}dx\, dy, \tag{4}$$

where $F(u, v)$ is the transformed image in the frequency domain, $f(x, y)$ the original image in the spatial domain, and (u, v) the frequency coordinates.

The result of the 2D Fourier transform is a complex-valued image, where the magnitude, $|F(u, v)|$, represents the strength of the frequency and the phase represents the position of the frequency. The magnitude of the transformed image is often used to analyze the frequency content of the image, as it represents the strength of the different frequencies present in the image. The 2D Fourier transform can be used to analyze the frequency content of an image in both the horizontal and vertical directions. This is useful for image processing tasks such as image enhancement, filtering, and compression.

We used Richard's method to estimate the degree of blurriness, as described in [40]. This method employs the Fourier transform to measure no-reference noise and blur. Initially, the input image is transformed into the frequency domain using the Fourier transform. To evaluate the energy of a specific frequency in all directions, we create n equi-spaced circles with radius r_i for $i = 1, \ldots, n$ in the 2D Fourier transform domain, identified by the (u, v) coordinate system as shown in Fig. 4(a). Subsequently, we calculate the normalized power within the i-th ring, which is the area between adjacent circles with radii r_i and r_{i-1}, where $r_0 = 0$.

$$p_i = \frac{1}{P}\left\{\sum_{u^2+v^2\leq r_i} F(u, v) - \sum_{u^2+v^2\leq r_{i-1}} F(u, v)\right\}, \quad \text{for} \quad i = \{1, \ldots n, \} \tag{5}$$

where $P = p_1$. The normalized power in each ring is shown in Fig. 4(b).

Fig. 5 shows sets of images, Fourier transform magnitude, and the corresponding normalized power. As shown in Fig. 5(a), the natural scene has a linear and balanced shape on the histogram. On the other hand, the noisy scene has relatively more high-frequency components, and the blurry scene lacks both low- and high-frequency components as shown in Fig. 5(b) and 5(c), respectively.

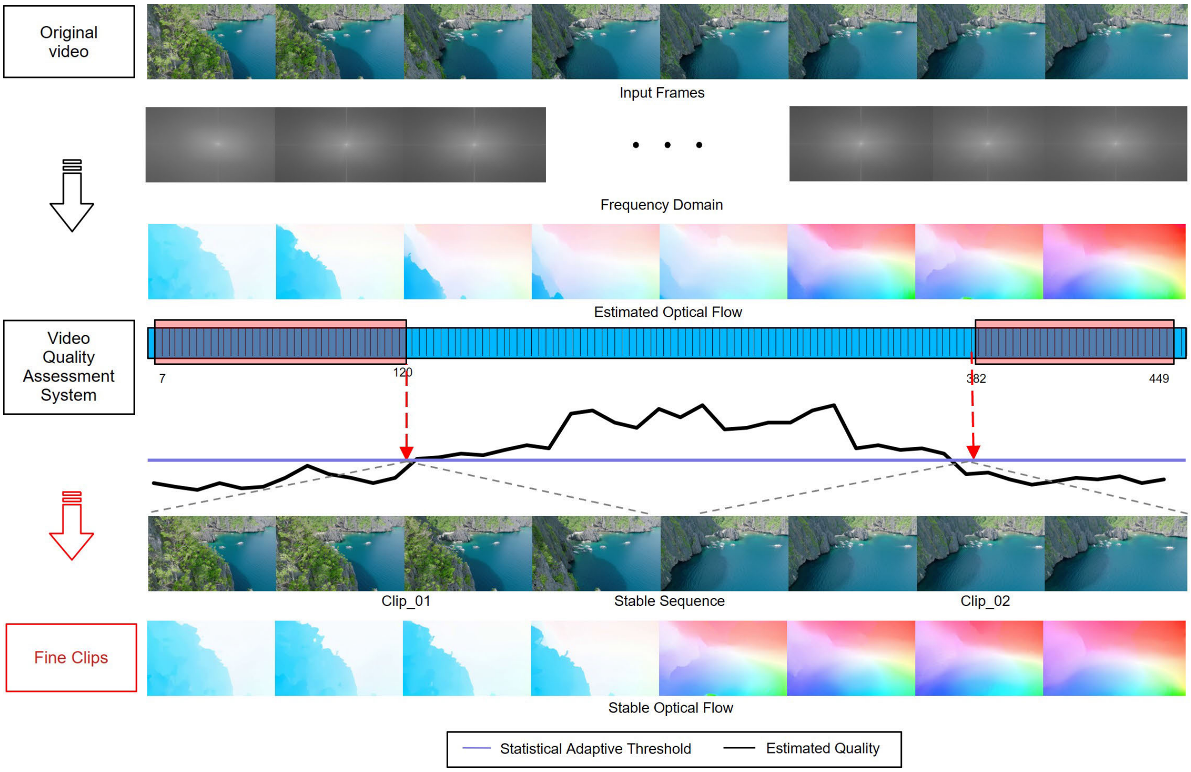


FIGURE 3. (From top to bottm) Optical flow and the corresponding video frames, blurriness, and camera motion score graph with an adaptive threshold, a collection of video frames in a stable sequence, and their optical flows. We first estimate blurriness using the Fourier property and camera motion based on optical flow. After that, we calculate the video adaptive threshold based on the statistical adaptive threshold algorithm.

In Fig. 4(b), the blurriness of an image is measured using the ℓ_1 distance between a diagonal line L and its pixels p_i.

$$e_i = p_i - \ell_i. \tag{6}$$

The overall blurriness distance error E' is computed as

$$E' = \sum_{i=0}^{n} e_i. \tag{7}$$

A value of E' greater than zero denotes a noisy frame, whereas a value less than zero indicates blurring. In this context, we are singularly concerned with blurring, hence utilize the negative E' value to deduce the image's blurriness, represented by B'.

$$B' = \begin{cases} |E'|, & \text{if } E' < 0 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Our custom blurriness metric is explicitly devised to discern defocus blur induced by errors in camera focusing or subject movements. This metric sidesteps other quality impediments such as noise or compression-related artifacts and contrasts the scrutinized image with a pristine, sharp counterpart devoid of any noise.

C. CAMERA MOVEMENT ESTIMATION USING VIDEO STABILIZATION PROPERTY

Optical flow is designed to determine the displacement between two consecutive image frames captured at times t and $t + dt$. Each pixel's flow vector indicates its relative movement between these frames, facilitating the estimation of inter-frame camera motion. To validate its accuracy, optical flow estimations can be juxtaposed with ground truth data, typically sourced from motion capture systems or dedicated sensors that monitor camera's spatial and angular coordinates. Such a comparison offers a verifiable account of genuine camera motion. As elucidated by the research in [27], optical flow has been proven efficacious in deciphering both rotational and translational aspects of camera motion, particularly for marginal inter-frame displacements.

For our optical flow estimation, we adopted FlowNet 2.0 as the cornerstone—a deep learning methodology proficient in addressing both minuscule and extensive shifts. Deep learning strategies for video stabilization leverage the principle of optical flow to execute grid warps, ensuring a commendable stabilization accuracy [41], [42]. Such reliance underscores the pivotal role of optical flow in characterizing both object trajectories and camera dynamics, underscoring its pertinence in the realm of video stabilization.

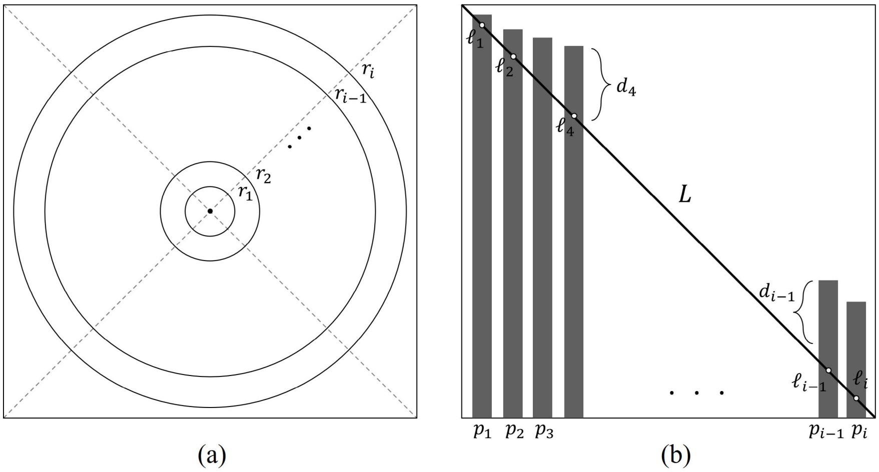


FIGURE 4. (a) Circles in the Fourier domain with (u, v) coordinates and (b) the normalized power within each ring [40].

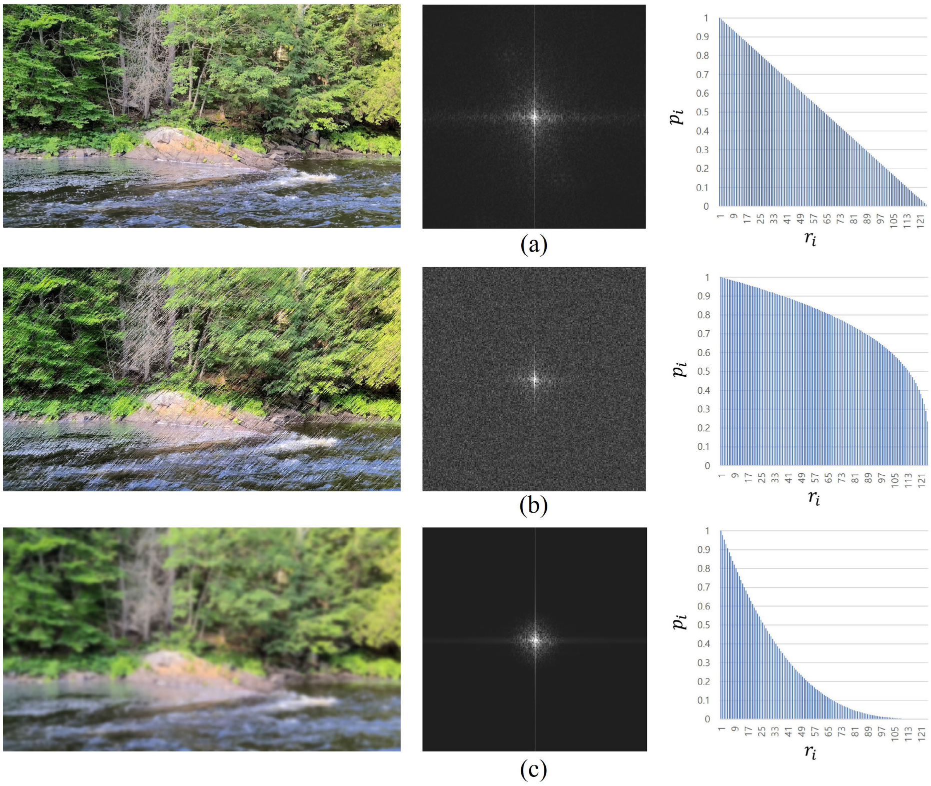


FIGURE 5. Visualization of (a) natural, (b) noisy, and (c) blurry images, with their corresponding Fourier transform magnitudes and histograms.

Fig. 6 describes the architecture of FlowNet 2.0. The design features bifurcated pathways: one addressing large displacements and the other tailored for smaller shifts. The former amalgamates FlowNetC and FlowNetS modules, which respectively ascertain the correlation of input visuals or features and warp the preceding frame $I_{t-1} = (x, y)$ based on the intermediate optical flow $F_t = (\delta x_{t-1}, \delta y_{t-1})$. The consequent warped frame $\tilde{I}_{t-1}$ aids in error computation.

$$e_i = \|\tilde{I}_{t-1} - I_t\|. \tag{9}$$

The computed error is fed as input to optimize the network. This proposed warping operation significantly improves the results, allowing for the use of a shallow layer compared to previous methods. To handle small displacement, the network

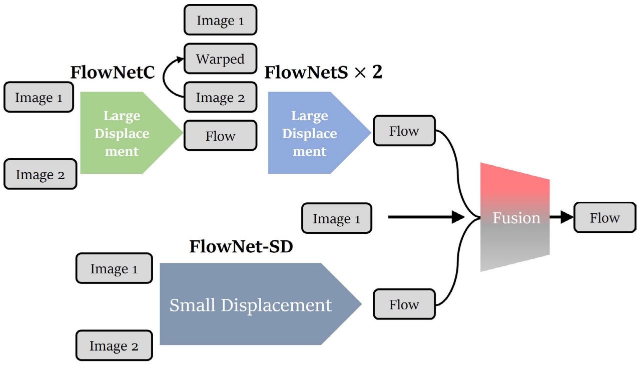


FIGURE 6. Structure of the FlowNet 2.0 model [22].

employs a smaller stride at the beginning and convolutions between up-convolutions in the FlowNetS architecture. The resulting flows from both paths are fused using the small fusion network to provide the final estimation.

Fig. 7 demonstrates that the original FlowNet yields noisy results when dealing with small movements and details. In contrast, the FlowNet 2.0 with a two-path architecture produces optical flow estimates that are robust in handling small displacements.

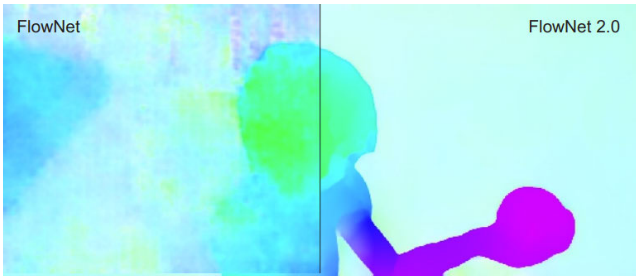


FIGURE 7. Visual comparison between the original FlowNet and FlowNet2.0 using our proposed system.

We estimated camera movement M_t by utilizing the magnitude of optical flow F_t. The algorithm for estimating the camera movement is illustrated in Fig. 8. The video sequence pair I_t and I_{t-1} are fed into the FlowNet 2.0 network to compute optical flow F_t. We then apply principal component analysis (PCA) to F_t, assuming that the main principal components correspond to camera movements as they represent global motion, while the movements of objects within the scene represent local motion.

$$M_t = \arg\max \text{PCA}(o_t), \tag{10}$$

where o_t is each optical flow value on the corresponding pixel.

Finally, we computed the magnitude of camera motion C_t by summing over the camera motion vectors.

$$C_t = \sum_{i}^{n} m_i, \tag{11}$$

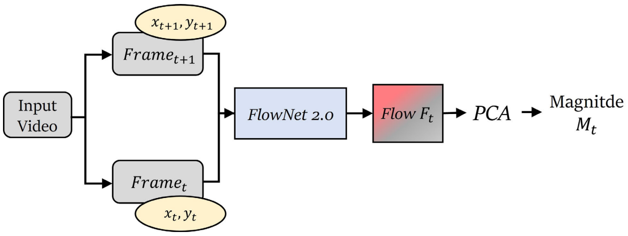


FIGURE 8. Camera movement estimation network.

where m_i is each magnitude of the major principal components.

D. STATISTICAL ADAPTIVE THRESHOLD

The proposed method estimates the blurriness B and camera movement C of a video. Based on these results, an adaptive threshold is calculated using statistical and iterative search algorithms. However, it is challenging to distinguish stable images from videos captured in various domains. For example, while a particular input video may remain stable within a predetermined range, videos of the sea or mountain areas are significantly impacted by waves or the shaking of leaves. Furthermore, different domains have distinct motion patterns, which makes it impractical to use a fixed threshold. Additionally, differentiating intentional motion from actual unstable motion in an input video is also challenging. As a result, it is crucial to discover a threshold that can adapt to the input video.

To address this problem, we propose the Statistical Adaptive Threshold (SAT), which separates the stable sequences from the videos. An adaptive threshold is a threshold that is calculated based on the characteristics of the input data. This is in contrast to a fixed threshold, which is the same for all input data. Adaptive thresholding is useful in cases where the input data has a wide range of values, or where the threshold value needs to be adjusted depending on the specific input.

In the context of video stabilization, adaptive thresholding can be used to distinguish between stable and unstable video frames. The threshold value can be adjusted based on the estimated blurriness and camera movement of the video, as well as the domain in which the video was captured.

SAT utilizes FlowNet 2.0, a deep learning method for motion estimation in videos, as a backbone to extract stable images. Optical flow-based camera motion estimation is the foundation for image stabilization, making it a crucial component of SAT, and providing improved performance. In addition, we use the blurriness value derived from the Fourier property. SAT is unique compared to other methods as it incorporates statistical techniques such as mean, standard deviation, and variance to evaluate the motion of an object in an input image. Additionally, SAT is distinct in the following ways:

- It incorporates statistical techniques such as mean, standard deviation, and variance to evaluate the motion of an object in an input image. This allows SAT to more accurately distinguish between stable and unstable video frames.

- SAT uses an adaptive threshold to separate stable sequences. This adaptive threshold is calculated based on the estimated blurriness and camera movement of the video, as well as the domain in which the video was captured. This allows SAT to be more robust to a wide range of input data.
- SAT is able to handle a wide range of video inputs, including videos with different levels of blurriness and camera movement, as well as videos from different domains.

To accurately determine the camera motion between frames and blurriness in various domains, the algorithm is computed by the following formula:

$$V_n' = \sum_{i=1}^{N} \Big[\sqrt{|f_{c'}^{C}(I_i, I_{i-1})|^2} + f_{c'}^{B}(I_i) \Big], \tag{12}$$

where I represents an image frame of input video, and values of video quality is stored as a list form. We also utilize a stabilization range variable to identify which frame is stable. A pseudo code for calculating stabilization range variables is shown in Algorithm1.

When a fixed threshold is employed, there's a risk of misclassifying minor camera movements as undesirable in video clips. Conversely, if we establish a threshold based solely on statistical methods without a constant, it can lead to misclassifying stable videos as problematic since the threshold must fall within the range of values observed in stable videos. To address this challenge, we adopt a combined approach that leverages both the stabilization variable and the stabilization constant.

To generate a stabilization variable, we compute the value of $S_{\min}^*$ and sort it based on the average of C_n' using S':

$$S' = \frac{\min(V_n') + \max(V_n')}{2}. \tag{13}$$

In the version of (14), which accounts for the relocation of outliers, we identify the value that minimizes the difference between the mean of C_n' and S' as followed:

$$S_{\min}^* = \arg\min_{V_i \in V_n} \sum_{i=1}^{N} \Big[\frac{V_i' - Q_2^{\text{median}}(V_n')}{Q_3(V_n') - Q_1(V_n')} \Big], \tag{14}$$

where, $Q^{\text{median}}2$ represents the median of $C'n$ as defined in (12). Q_3 corresponds to the third quartile (75 The distance from the center to zero is measured as a quartile [43]. C_i' denotes the size of each object movement. Therefore, robust normalization mitigates the effect of the outliers. The stabilization range variable is denoted as θ and is computed as:

$$\theta = f_{\alpha^*}(S_{\min}^*). \tag{15}$$

The optimal stabilization variable θ is determined by the size of camera motion computed by equation (14), the maximum-minimum value of motion with outliers removed by equation (14), and the value α^* that minimizes $S_{\min}^*$ and S' in each equation. Once the stabilization variable θ is computed, it is used to estimate the threshold using equations (18).

Algorithm 1 SV-Setup: Stabilizing Variable Setup in SAT

```
Require: $D$: input video, $\alpha$: initial stabilization variable, $\beta$: step size
Ensure: $\theta$: final stabilization variable(SV)

for each frame $I_i$ in input video $D$ do
    Compute the video quality $V^*$ using camera motion $C'$ and the blurriness $B'$
    $C_i' = C(I_i, I_{i-1})$ with flownet2.0 movement
    $C_n' = [\sqrt{|C_1'|^2}, \ldots, \sqrt{|C_i'|^2}]$
    $B_n' =$ CDF of Fourier rings $|D|$
    $V^* = C' + B'$
end for
$S_n' = S_{\text{scaler}}(V^*)$ with outlier removal and normalization

$\alpha = 0.85, \beta = 0.05$
for step $i = \alpha$ to $\beta$ do
    if $S_n' \leq \alpha$ then
        $S_{\min}' = \min(S_n'), S_{\max}' = \max(S_n')$
        $\bar{S}_{\text{median}} = \frac{1}{2}(S_{\max}' + S_{\min}')$
        $\tilde{S}_{\text{mean}} = \frac{1}{n}\sum S_n^*$
        if $\tilde{S}_{\text{mean}} \leq \bar{S}_{\text{median}}$ then
            $S_{\text{ascending}}^* = [S_1', \ldots, S_n']$ is an ascending list
            $\alpha_{\text{ascending}}^* = [\alpha_1', \ldots, \alpha_n']$ is an ascending list
            $\theta = \alpha_{\text{ascending}}^*(\min(S_{\text{ascending}}^*))$
            return $\theta$
        else
            $\theta = \alpha$
            return $\theta$
        end if
        Update $\alpha \leftarrow \alpha - \beta$
    end if
end for
```

To estimate the threshold for separating stable sequences, we use a statistical method commonly used in a normal distribution, such as mean and standard deviation. We assume that the computed values C_n' follow a normal distribution. In statistics of normal distribution, 2σ covers their 97.9% distribution. Based on these observations, it became evident that most issues occur in just a few seconds within the entire video. Taking into account this observation and statistical considerations, we compute the average T_{mean}, standard deviation T_{std}, and margin values T_{margin} as follows:

$$T_{\text{mean}} = \frac{1}{n} \sum_{k=0}^{\theta} V_n'(k), \tag{16}$$

$$T_{\text{std}} = \frac{\sum_{k=0}^{\theta}(V_i'(k) - T_{\text{mean}})}{n}, \tag{17}$$

$$T_{\text{margin}} = 1 * \frac{Q_3(V_n') - Q_1(V_n')}{\theta}. \tag{18}$$

The computed margin T_{margin} is initially set to 1, but in some domains, the constant value 1 can have a relatively large effect on the total threshold. To address this issue, we multiply stabilization variable θ to the initial margin (18). The final adaptive threshold is then calculated as the sum of T_{mean}, T_{std}, and T_{margin}, taking into account the stabilization variable.

This adaptive thresholding method is effective because it takes into account the statistical properties of the data, the severity of the camera motion, and the fact that most issues occur in just a few seconds within the entire video.

$$SAT = T_{\text{mean}} + T_{\text{std}} + T_{\text{margin}}. \tag{19}$$

IV. EXPERIMENTAL RESULTS

A. DATASET

We evaluated the proposed method on a dataset collected from the actual editing process [12]. Fig. 9 shows a collection of video data from the KEYCUTstock website. It provides footage for Premiere in 4K, 8K, and higher resolutions. The collection includes various domains, such as cityscape, nature, timelapse, etc.

Table 1 shows the ratio of bad clips detected by an editing specialist in the collected dataset. The bad clips are mainly caused by unwanted camera movements, which can result from shaking hands while filming with a handheld camera, or the effects of wind or panning on a drone. Another significant issue is incomplete focus due to being out-of-focus. These issues were observed to occur in a few seconds of a long clip.

B. EVALUATION METRICS

We evaluated the proposed method with various metrics that are widely used for image quality assessment. Peak Signal-to-Noise Ratio (PSNR), Mean-Squared Error (MSE), Structural Similarity (SSIM) [44], and Visual Information Fidelity (VIF) [45] are used as metrics. To adapt the metrics to the proposed method, we compare the results of the original video and video clips which are extracted by the proposed threshold. To evaluate the quality of predicted video clips from various perspectives, we focused on the following features:

- Loss of image quality between frames of the same image.
- Similarity between frames by considering luminance, contrast, and structure of the image.
- Comparison of the amount of information present in the image with the reference image.

1) PEAK SIGNAL-TO-NOISE RATIO

PSNR represents the power of noise relative to the maximum power that a signal has. It is used to evaluate information loss and image quality on images that are contaminated by distortion. It is computed as the following formula:

$$PSNR = 20 * log_{10}(\frac{MAX_I}{\sqrt{MSE}}) \tag{20}$$

$$= 20 * log_{10}(MAX_I) - 10 * log_{10}(MSE), \tag{21}$$

where, MAX_I is the maximum value of the image. For 8-bit grayscale images, it ranges from 0 to 255.

2) MEAN SQUARED ERROR

MSE is a commonly used metric to evaluate the similarity between two images. The purpose of using MSE is to calculate the average squared difference between the pixel values of two images. When comparing two images using MSE, the pixel values of the images are compared, and the differences between the pixel values are squared. The squared differences are then averaged to produce a single value that represents the overall difference between the two images. The lower the MSE value, the more similar the two images are considered to be. The function of MSE is as follows:

$$MSE = \frac{1}{m*n}\sum_{0}^{m-1}\sum_{0}^{n-1}||I(i,j) - I_{n-1}(i,j)||^2, \tag{22}$$

where I is an image of size $m*n$, and I_{n-1} is a distorted image by addtive noise to I. Here, I represents the current frame of the input video, and I_{n-1} represents the previous frame. The smaller the value of Mean Squared Error (MSE), as shown in Equation (22), the higher the value of Peak Signal-to-Noise Ratio (PSNR) because it appears in the denominator. Therefore, lower MSE and higher PSNR indicate better image quality. However, PSNR has certain limitations in terms of human-perceived image quality, such as the inability to detect information loss and blurriness in the image.

3) STRUCTURAL SIMILARITY INDEX MEASURE

SSIM is a perceptual metric that quantifies image quality degradation caused by processing such as data compression or by losses in data transmission. It is designed to better imitate the human perception of image quality than PSNR. SSIM is calculated based on the luminance, contrast, and structure of an image. The SSIM score ranges from 0 to 1, with a score of 1 indicating that the two images are identical. A score of 0 indicates that the two images are completely different.

$$SSIM(I, I_{n-1}) = L(I, I_{n-1}) \times C(I, I_{n-1}) \times S(I, I_{n-1}), \tag{23}$$

where $L(I, I_{n-1})$ compares the average luminance of two images, $C(I, I_{n-1})$ is the contrast function, and $S(I, I_{n-1})$ is the structure function formulated as

$$L(I, I_{n-1}) = \frac{2\mu_I\mu_{I_{n-1}} + c_1}{\mu_I^2 + \mu_{I_{n-1}}^2 + c_1},$$

$$C(I, I_{n-1}) = \frac{2\sigma_I\sigma_{I_{n-1}} + c_2}{\sigma_I^2 + \sigma_{I_{n-1}}^2 + c_2},$$

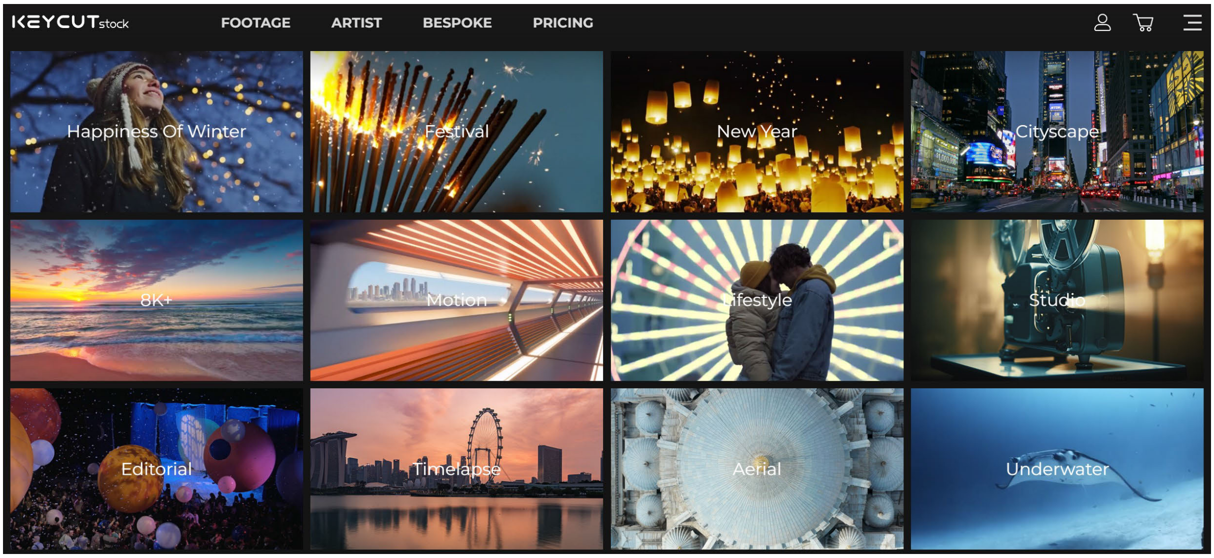


FIGURE 9. Collection of selected video data from the KEYCUTstock. Since the domains of the videos are diverse, we designed an adaptive threshold using statistical methods to accurately identify small camera motion.

$$S(I, I_{n-1}) = \frac{\sigma_{I*I_{n-1}} + c_3}{\sigma_I \sigma_{I_{n-1}} + c_3}, \tag{24}$$

where μ, σ, and $\sigma_{I*I_{n-1}}$ respectively represent the mean, variance, and covariance of $\{I, I_{n-1}\}$. c_1, c_2, c_3 are stabilizing variables that compensate for the problem of the denominator. $c_1 = (k_1*L)^2, c_2 = (k_2*L)^2$, and $c_3 = c_2/2$. $k_1 = 0.01, k_2 = 0.03$ by default.

4) VISUAL INFORMATION FIDELITY

VIF is a full-reference image quality assessment metric based on natural scene statistics and the notion of image information extracted by the human visual system. VIF is calculated by comparing the statistical properties of the test image to the reference image. These statistical properties are designed to capture the key features of an image that are important for human perception, such as the distribution of luminance, contrast, and structure. Same with SSIM, VIF scores range from 0 to 1, with a score of 1 indicating that the two images are identical. A score of 0 indicates that the two images are completely different..

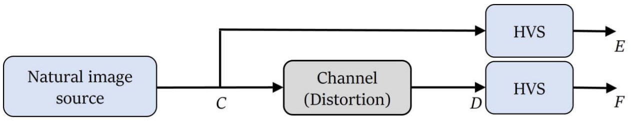


FIGURE 10. Visualization of VIF. [45].

As shown in Fig. 10, The original image is a clean image signal without distortion and is represented by C The human visual system (HVS) perceives a signal represented by E, while the image signal affected by the distortion is represented by D through the channel C. The distortion function is a zero mean Gaussian by default. The signal F recognized by the HVS through the channel C can be expressed as:

$$VIF = \frac{\sum I(\bar{C}; \bar{F}|S)}{\sum I(\bar{C}; \bar{E}|S)}, \tag{25}$$

where $I_{n-1}(C; F)$ and $I(C; E)$ are the amount of mutual information. S represents the positive scalar value computed by the Gaussian scale mixture (GSM) model to C.

5) PRECISION, RECALL, AND DICE SCORE

Precision is the percentage of the actual Ground Truth (GT) in the predicted positive value. It formulated as:

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{TP}{\#\,\text{ground truth}}. \tag{26}$$

Recall, also called sensitivity, is the percentage of the predicted positive range for the positive range of GT, which can be formulated as

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{TP}{\#\,\text{prediction}}. \tag{27}$$

Dice score is well-known as F1 score as the harmonic mean of recall and precision. Score 1 refers to the best precision and recall score, which can be formulated as

$$\text{Dice} = \frac{2 \times TP}{(TP + FP) + (TP + FN)} \tag{28}$$

$$= \frac{2(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}. \tag{29}$$

C. QUANTITATIVE RESULTS

We used PSNR, SSIM, and VIF metrics to evaluate the quality of the processed videos. We used different methods to find thresholds for detecting stable clips in the source videos, including the mean, mean + standard deviation, mean

TABLE 2. Experimental results of Video 1 (PSNR, SSIM, VIF score).

Method	Extraction	PSNR	SSIM	VIF
Ground Truth	**2**	**32.35**	**0.9491**	**0.6955**
Original	1	30.56	0.9098	0.6188
Average	2	32.38	0.9495	0.6966
Average+Std	1	31.38	0.9310	0.6506
Avg+Std(30%)	2	31.15	0.9256	0.6397
Proposed method(SAT)	**2**	**32.41**	**0.9499**	**0.7011**

TABLE 3. Experimental results of Video 2 (PSNR, SSIM, VIF score).

Method	Extraction	PSNR	SSIM	VIF
Ground Truth	**2**	**27.32**	**0.8841**	**0.4641**
Original	1	23.51	0.8333	0.2893
Average	2	26.08	0.8673	0.4294
Average+Std	4	25.98	0.8601	0.3568
Avg+Std(30%)	2	25.85	0.8648	0.4206
Proposed method(SAT)	**2**	**27.19**	**0.8829**	**0.4646**

TABLE 4. Experimental results of Video 3 (PSNR, SSIM, VIF score).

Method	Extraction	PSNR	SSIM	VIF
Ground Truth	**3**	**32.24**	**0.8989**	**0.3986**
Original	1	31.45	0.8804	0.3489
Average	5	31.44	0.8830	0.3697
Average+Std	3	30.71	0.8728	0.3421
Avg+Std(30%)	5	31.43	0.8831	0.3692
Proposed method(SAT)	**3**	**32.31**	**0.8992**	**0.4092**

TABLE 5. Experimental results of Video 4 (PSNR, SSIM, VIF score).

Method	Extraction	PSNR	SSIM	VIF
Ground Truth	**3**	**45.69**	**0.9592**	**0.7787**
Original	1	43.86	0.9643	0.7271
Average	3	45.38	0.9536	0.7659
Average+Std	2	44.12	0.9693	0.7399
Avg+Std(30%)	3	45.52	0.9580	0.7776
Proposed method(SAT)	**3**	**45.53**	**0.9580**	**0.7779**

+ bottom 30% of standard deviation, and our proposed SAT algorithm. We then compared the stable clips from each threshold to the original video. Our SAT threshold outperformed the other methods in terms of clip stability. Additionally, the stable clips from the proposed threshold had higher quality than the original video, as objectively measured by the chosen metrics. The detail of the video used in the evaluation is shown in Fig. 11. The detailed results are summarized in Tables 2, 3, 4, 5. Additionally,

TABLE 6. Experimental results of Video 1, Video 2, Video 3, Video 4 (Dice, Precision, Recall score).

video	Method	Dice	Precision	Recall	Mean
video1	Average	0.9727	0.9979	0.9487	0.9731
	Average+Std	0.9345	0.8880	0.9862	0.9362
	Avg+Std(30%)	0.9452	0.9074	0.9862	0.9463
	Proposed method(SAT)	**0.9901**	**0.9940**	**0.9862**	**0.9901**
video2	Average	0.8409	0.8810	0.9635	0.8951
	Average+Std	0.7341	0.5929	0.9635	0.7635
	Avg+Std(30%)	0.8915	0.8296	0.9535	0.8949
	Proposed method(SAT)	**0.9166**	**0.9821**	**0.8594**	**0.9194**
video3	Average	0.6972	0.5643	0.9120	0.7245
	Average+Std	0.6778	0.5126	1.0000	0.7301
	Avg+Std(30%)	0.7191	0.5732	0.9648	0.7641
	Proposed method(SAT)	**0.8885**	**1.0000**	**0.7993**	**0.8959**
video4	Average	0.9524	0.9836	0.9231	0.9530
	Average+Std	0.7602	0.6132	1.0000	0.7911
	Avg+Std(30%)	0.8159	0.6890	1.0000	0.8350
	Proposed method(SAT)	**0.9572**	**1.0000**	**0.9179**	**0.9584**

Table 8 presents extensive results for a variety of domains. While the proposed SAT method performs slightly worse than other methods in some cases, it outperforms them in most cases.

Next, we used the dice, precision, and recall metrics to measure the overall quality of the video. As shown in Table 6, the proposed method produced clips that were most similar to the ground truth clips, except in terms of recall. The reason for the lower recall is that the proposed method divided the clips slightly more narrowly than the ground truth clips. However, the overall dice, precision, and recall scores of the proposed method were higher than those of the other methods.

We quantified the overall video quality using metrics such as dice, precision, and recall. As illustrated in Table 6, the proposed method's outputs closely resemble the reference clips in most metrics, with an exception in recall. This slight dip in recall arises because our method partitions clips somewhat more finely compared to the ground truth. Despite this, the aggregate scores of dice, precision, and recall surpass other methods.

To benchmark our proposed technique, we applied both the camera motion and blurriness indices on two experimental videos: one solely exhibiting camera movement and another with combined camera movement and blur. As Table 7 indicates, the disparity in performance between our method and a motion-only index approach was marginal for the camera movement-only video. Conversely, in the video displaying both phenomena, the performance gap was substantial, underlining the efficacy of our dual-index approach over the sole reliance on the motion index.

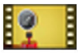
Video 1				Video 2				Video 3				Video 4			
Information		Ground Truth		Information		Ground Truth		Information		Ground Truth		Information		Ground Truth	
Image		Clip	Frame	Image		Clip	Frame	Image		Clip	Frame	Image		Clip	Frame
		GT 1	0~47			GT 1	0~102			GT 1	165~335			GT 1	113~133
Domain	Museum	GT 2	100~558	Domain	Machine	GT 2	391~439	Domain	River	GT 2	362~405	Domain	Grass	GT 2	201~244
Size	91.4 MB			Size	73.6 MB			Size	127.1 MB	GT 3	444~512	Size	45.7 MB	GT 3	304~433
Time	24 sec			Time	19 sec			Time	25 sec			Time	18 sec		
Fps	23.98 fps			Fps	25 fps			Fps	25 fps			Fps	23.98 fps		
Frame	588			Frame	480			Frame	631			Frame	439		

FIGURE 11. Detailed description of Video 1 to Video 4 used in the quantitative evaluation.

TABLE 7. A quantitative comparison of the proposed motion + blurriness index and the motion index alone on Video 4, which contains both camera movement and blur.

Method		Video 3(Camera motion)			Video 4(Camera motion + Blurriness)		
Ground Truth	Clipped GT	GT1	GT2	GT3	GT1	GT2	GT3
	Frame	165-335	362-405	444-512	113-133	201-244	304-433
Motion index	Frame(Predict)	171-320	362-404	456-490	114-133	198-240	304-418
	Loss(Missing frame)	**12.28%** (21/171)	**4.55%** (2/44)	**49.28%** (34/69)	**4.76%** (1/21)	**14.89%** (7/47)	**11.54%** (15/130)
	Average	**22.04%**			**10.40%**		
Motion + Blurriness index (Proposed method)	Frame(Predict)	171-320	362-404	456-490	114-133	200-243	304-425
	Loss(Missing frame)	**12.28%** (21/171)	**4.55%** (2/44)	**49.28%** (34/69)	**4.76%** (1/21)	**4.26%** (2/47)	**6.15%** (8/130)
	Average	**22.04%**			**5.07%**		

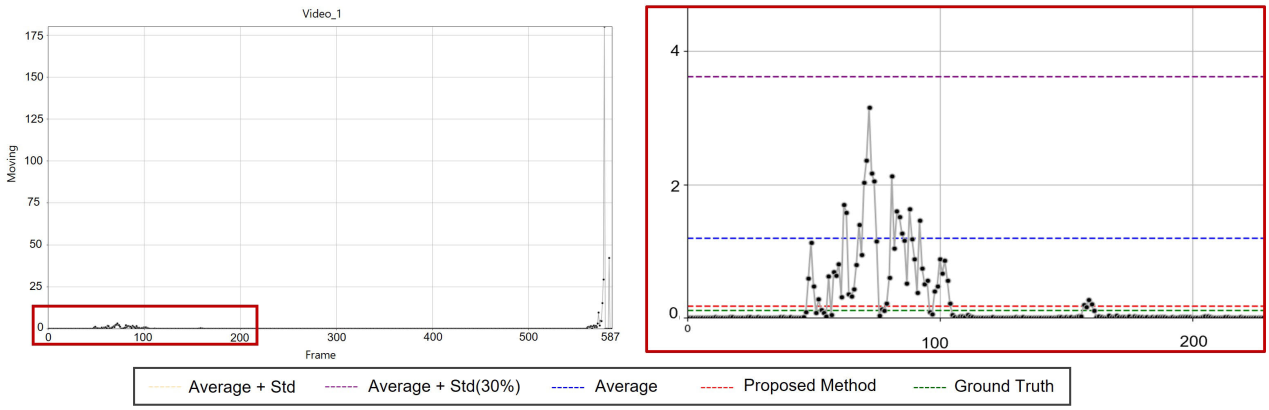


FIGURE 12. Estimated graph and thresholds of video 1. Ground truth is manually set by an editorial specialist.

D. QUALITATIVE RESULTS

We conducted a qualitative experiment using graphs for each threshold, as illustrated in Fig. 12, 13, and 14.

Despite the presence of various types of input images in various domains, the proposed method shows the most similar threshold to the ground truth threshold. Even though there are various types of input images in various domains, our proposed method crops stable images. Fig. 12 is the case that the last few seconds have extremely large movements. Due to this problem, statistical methods are highly affected by outliers. However, the proposed method accurately finds the optimal threshold than others. Fig. 13 is the case that has some movements in nature. In this case, some movements are acceptable but it is ambiguous. The editorial specialist

TABLE 8. Quantitative comparison on the various domains in real-world video with the PSNR, SSIM, VIF metrics. GT means clips extracted by human expert-defined threshold, Avg means clips extracted by threshold computed by average of index. Avg+Std means clips extracted by threshold computed by average and standard deviation. SAT means clips extracted by the proposed statistical adaptive threshold.

Video	Domain	PSNR				SSIM				VIF			
		GT	Avg	Avg+ Std	SAT	GT	Avg	Avg+ Std	SAT	GT	Avg	Avg+ Std	SAT
Video1	Museum	32.35	32.38	31.38	**32.41**	0.9491	0.9495	0.9310	**0.9499**	0.6955	0.6966	0.6506	**0.7011**
Video2	Machine	27.32	26.08	25.98	**27.19**	0.8841	0.8673	0.8601	**0.8829**	0.4641	0.4294	0.3568	**0.4646**
Video3	River	32.24	31.44	30.71	**32.31**	0.8989	0.8830	0.8728	**0.8992**	0.3986	0.3697	0.3421	**0.4092**
Video4	Grass	45.69	45.38	44.12	**45.53**	0.9592	0.9536	0.9693	**0.9580**	0.7787	0.7659	0.7399	**0.7779**
Video5	Tree	42.07	**42.64**	40.97	41.60	0.9375	**0.9416**	0.9398	0.9380	0.7563	**0.7674**	0.7204	0.7402
Video6	Fountain	34.84	32.46	31.86	**36.31**	0.8701	0.8147	0.8043	**0.9035**	0.5042	0.3931	0.3612	**0.5753**
Video7	Park	30.88	29.46	29.52	**31.03**	0.7912	0.7539	0.7490	**0.7964**	0.3231	0.3100	0.2721	**0.3310**
Video8	Paddy	40.00	39.48	**40.12**	39.48	0.9312	0.9192	**0.9287**	0.9192	0.6192	0.6094	**0.6130**	0.6094
Video9	Sky&Tree	30.91	**32.21**	30.64	30.93	0.8785	**0.9066**	0.8745	0.8783	0.4349	**0.4651**	0.4221	0.4365
Video10	Insect	43.55	42.47	38.92	**43.85**	0.9693	0.9567	0.9434	**0.9768**	0.7569	0.7086	0.5760	**0.7821**
Video11	Lion	43.11	40.26	40.97	**42.50**	0.9556	0.9461	0.9447	**0.9501**	0.7064	0.6072	0.6466	**0.6918**
Video12	Town	42.31	42.17	**42.49**	**42.49**	0.9724	0.9608	**0.9728**	**0.9728**	0.7385	0.7349	**0.7483**	**0.7483**
Video13	Bird&Wolf	34.39	31.99	32.97	**34.83**	0.9269	0.8777	0.8985	**0.9383**	0.5003	0.3969	0.4374	**0.5359**
Video14	Bear	34.00	33.54	33.49	**34.50**	0.9171	0.9148	0.9145	**0.9284**	0.6769	0.6304	0.6284	**0.6746**
Video15	Apartment	33.58	32.88	31.77	**34.11**	0.9037	0.8964	0.8783	**0.9081**	0.4443	0.4105	0.3640	**0.4597**
Video16	Women	39.53	38.71	38.71	**39.97**	0.9577	0.9550	0.9550	**0.9595**	0.5365	0.5078	0.5078	**0.5573**
Video17	Architecture	25.61	**25.23**	**25.23**	**25.23**	0.8384	**0.8183**	**0.8183**	**0.8183**	0.2434	**0.2380**	**0.2380**	**0.2380**
Video18	Clock tower	45.27	45.79	42.88	**45.81**	0.9650	0.9668	0.9591	**0.9671**	0.8147	0.8198	0.7706	**0.8199**
Video19	Bridge	34.79	33.58	34.32	**34.51**	0.9128	0.9041	0.9085	**0.9120**	0.4608	0.4337	0.4348	**0.4444**
Video20	Seaport	46.09	42.82	41.87	**46.69**	0.9753	0.9604	0.9548	**0.9796**	0.7876	0.7224	0.6874	**0.8008**

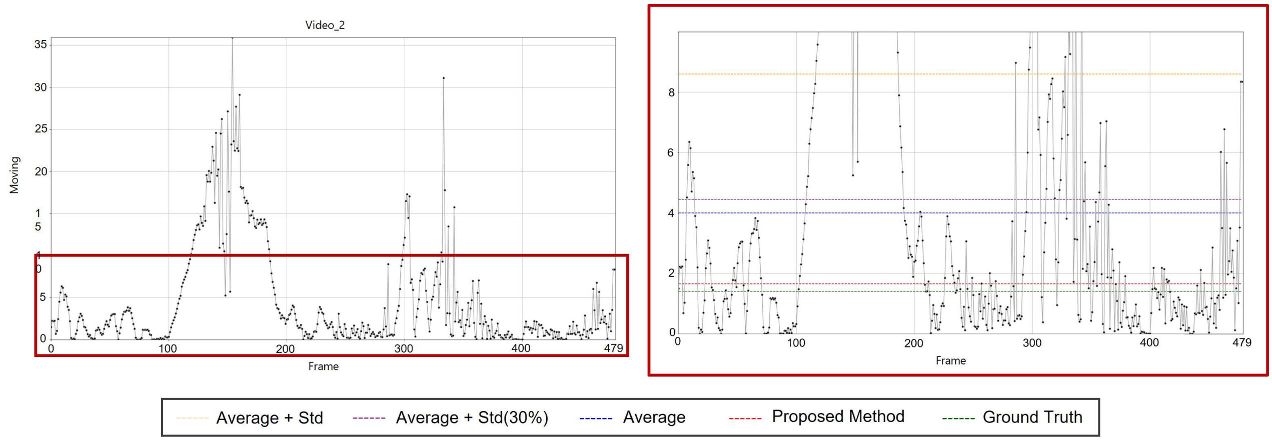


FIGURE 13. Estimated graph and thresholds of video 2.

judged that only a few seconds of the middle and a few seconds of the last are stable clips. The other methods have too high a threshold than ground truth, but the proposed method has a similar threshold than others. Fig. 14 is a video that has tiny movements and large movements in the last few seconds. Tiny movements are acceptable motions but the first and the last movements are not acceptable. The other methods have a relatively higher threshold than the ground truth, but the proposed method has a similar threshold to the ground truth threshold. Overall, the proposed method shows a lower threshold than the others and is similar to the ground truth threshold. Therefore, experiments show that the proposed method is similar to the editorial specialist.

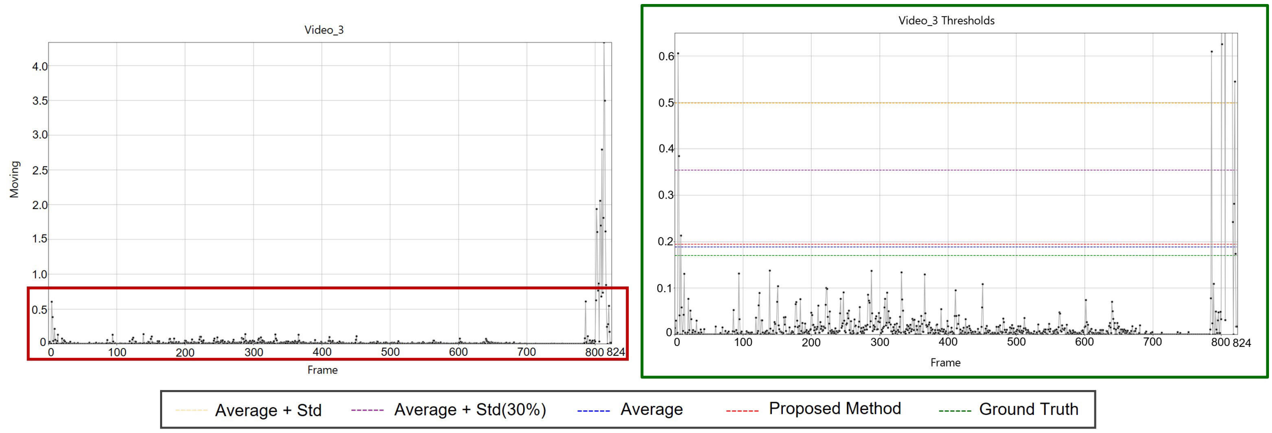


FIGURE 14. Estimated graph and thresholds of video 3.

TABLE 9. Conventions.

Variable	Definition
$I_t, I(x, y, t)$	Pixel value on the t-th frame of input video
$D(\cdot)$	Displacement function across consecutive frames
dx, dy	Transformed signal in the frequency domain
I_{t+1}	Next frame shifted by $F(t)$
$F(u)$	Transformed signal in the frequency domain
$f(\cdot)$	Input 1D signal
u	Frequency Coordinates
$F(u, v)$	Transformed image in the frequency domain
r	radius of frequency domain
p, P	normalized power
e, E'	blurriness distance error
B'	blurriness index
o, O	Intermediate optical flow
$\tilde{I}_{t-1}$	Image warped by optical flow O
m, M	Magnitude of camera movement
C	Magnitude of principal component of M
SAT	Statistical adaptive threshold of proposed method
V'	Video quality index
S', S_{scaler}	Temporal mean for scaling threshold
S^*_{min}	Temporal threshold that sorted ascending form
$Q_{\text{median}}, Q_1, Q_3$	median, first quartile, third ratio
θ	Stabilization range for robust normalization
T_{mean}	Mean vale of entire video quality
T_{std}	Standard deviation value of video quality
T_{margin}	Adaptive minimum margin

V. LIMITATION AND FUTURE WORK

Our proposed VQA system primarily grapples with the challenges posed by inadvertent camera movement and blurriness. Evaluations through an array of objective metrics coupled with visual juxtapositions exhibit commendable efficacy. Nonetheless, our approach has its limitations. Firstly, when an object's motion magnitude exceeds that of the camera's, the PCA operation can misconstrue the object's trajectory as stemming from the camera. Secondly, our blur metric is squarely aimed at detecting full-frame defocus blur instigated by lapses in camera focus or subject motion. Pinpointing and demarcating specific regions of sporadic blur within frames would entail the integration of intricate segmentation and spatial methodologies. Lastly, even with our statistical adaptive threshold mechanism, intentional camera movements remain challenging to discern accurately. The first two issues arise from inherent challenges in image processing. Nonetheless, further exploration into techniques like object segmentation, requiring fine-grained annotated datasets, might offer mitigation for these challenges.

VI. CONCLUSION

In this paper, we introduce a comprehensive Video Quality Assessment (VQA) system designed to address prevalent quality challenges such as unintentional camera movements and defocused blur. Fundamentally, the proposed approach integrates deep learning-driven optical flow estimation—specifically utilizing FlowNet 2.0—to capture and quantify camera motion dynamics across sequential frames. For the identification and quantification of blurriness, we utilized the Fourier transform to evaluate high-frequency content. Our system also introduces a unique iterative search mechanism that determines an adaptive threshold tailored to each video, grounded in its inherent statistical characteristics. Our research stands out due to the development of an autonomous VQA system proficient in mimicking human judgment concerning notable anomalies like erratic camera movements and focus discrepancies. Distinct from previous models and datasets, our methodology sidesteps the need for vast labeled training data. The proposed method not only enhances established objective quality benchmarks like PSNR, SSIM, and VIF but, more crucially, demonstrates scores that align closely with human subjective evaluations—affirming their perceptual accuracy and relevance.

ACKNOWLEDGMENT

(Donggoo Kang and Yeongjoon Kim contributed equally to this work.)

REFERENCES

[1] S. Winkler, *Digital Video Quality: Vision Models and Metrics*. Hoboken, NJ, USA: Wiley, 2005.

[2] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Proc. Int. Conf. Image Process.*, Sep. 2002, p. 3.

[3] R. Liu, Z. Li, and J. Jia, "Image partial blur detection and classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[4] A. Maalouf and M.-C. Larabi, "A no reference objective color image sharpness metric," in *Proc. 18th Eur. Signal Process. Conf.*, Aug. 2010, pp. 1019–1022.

[5] K. Rank, M. Lendl, and R. Unbehauen, "Estimation of image noise variance," *IEE Proc. Vis., Image, Signal Process.*, vol. 146, no. 2, p. 80, 1999.

[6] A. Amer and E. Dubois, "Fast and reliable structure-oriented video noise estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 113–118, Jan. 2005.

[7] N. D. Narvekar and L. J. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (CPBD)," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2678–2683, Sep. 2011.

[8] A. R. Reibman, S. Sen, and J. Van der Merwe, "Analyzing the spatial quality of internet streaming video," in *Proc. Int. Workshop Video Process. Quality Metrics Consum. Electron.*, 2005, pp. 1–6.

[9] K. Zhu, K. Hirakawa, V. Asari, and D. Saupe, "A no-reference video quality assessment based on Laplacian pyramids," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 49–53.

[10] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5923–5938, Dec. 2019.

[11] M. Shahid, A. Rossholm, B. Lövström, and H.-J. Zepernick, "No-reference image and video quality assessment: A classification and review of recent approaches," *EURASIP J. Image Video Process.*, vol. 2014, no. 1, pp. 1–32, Dec. 2014.

[12] *Stock Footage, 4K Stock Footage, 8K Stock Footage High Picture Quality Stock, Keycutstock*. Accessed: Jan. 2023. [Online]. Available: https://www.keycutstock.com/

[13] D. Li, T. Jiang, and M. Jiang, "Unified quality assessment of in-the-wild videos with mixed datasets training," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1238–1257, Apr. 2021.

[14] F. Götz-Hahn, V. Hosu, H. Lin, and D. Saupe, "KonVid-150k: A dataset for no-reference video quality assessment of videos in-the-Wild," *IEEE Access*, vol. 9, pp. 72139–72160, 2021.

[15] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, "Patch-VQ: 'Patching up' the video quality problem," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14014–14024.

[16] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2351–2359.

[17] J. You and J. Korhonen, "Deep neural networks for no-reference video quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2349–2353.

[18] J. Korhonen, Y. Su, and J. You, "Blind natural video quality prediction via statistical temporal features and deep spatial features," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 3311–3319.

[19] F. Yi, M. Chen, W. Sun, X. Min, Y. Tian, and G. Zhai, "Attention based network for no-reference UGC video quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1414–1418.

[20] W. Shen, M. Zhou, X. Liao, W. Jia, T. Xiang, B. Fang, and Z. Shang, "An end-to-end no-reference video quality assessment method with hierarchical spatiotemporal feature representation," *IEEE Trans. Broadcast.*, vol. 68, no. 3, pp. 651–660, Sep. 2022.

[21] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.

[22] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1647–1655.

[23] S. Zhao, L. Zhao, Z. Zhang, E. Zhou, and D. Metaxas, "Global matching with overlapping attention for optical flow estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17571–17580.

[24] Z. Huang, X. Shi, C. Zhang, Q. Wang, K. C. Cheung, H. Qin, J. Dai, and H. Li, "FlowFormer: A transformer architecture for optical flow," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 668–685.

[25] S. Zhao, Y. Sheng, Y. Dong, E. I. Chang, and Y. Xu, "MaskFlownet: Asymmetric feature matching with learnable occlusion mask," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6277–6286.

[26] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. 12th Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2012, pp. 611–625.

[27] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *Int. J. Comput. Vis.*, vol. 92, no. 1, pp. 1–31, Mar. 2011.

[28] M. Menze, C. Heipke, and A. Geiger, "Object scene flow," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 60–76, Jun. 2018.

[29] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vis. Conf.*, vol. 15, no. 50, 1988, p. 5244.

[30] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[31] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. 9th Eur. Conf. Comput. Vis.*, vol. 3951, May 2006, pp. 404–417.

[32] F. Bellavia and D. Mishkin, "HarrisZ+: Harris corner selection for next-gen image matching pipelines," *Pattern Recognit. Lett.*, vol. 158, pp. 141–147, Jun. 2022.

[33] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8918–8927.

[34] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4937–4946.

[35] J. Yu and R. Ramamoorthi, "Learning video stabilization using optical flow," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8156–8164.

[36] Y. Xu, J. Zhang, S. J. Maybank, and D. Tao, "DUT: Learning video stabilization by simply watching unstable videos," *IEEE Trans. Image Process.*, vol. 31, pp. 4306–4320, 2022.

[37] D. Cozzolino, G. Poggi, and L. Verdoliva, "Extracting camera-based fingerprints for video forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2019, pp. 1–8.

[38] Y.-L. Liu, W.-S. Lai, M.-H. Yang, Y.-Y. Chuang, and J.-B. Huang, "Hybrid neural fusion for full-frame video stabilization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2279–2288.

[39] C. Tang, X. Zhu, X. Liu, L. Wang, and A. Zomaya, "DeFusionNET: Defocus blur detection via recurrently fusing and refining multi-scale deep features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2695–2704.

[40] R. W. Dosselmann and X. D. Yang, "No-reference noise and blur detection via the Fourier transform," Dept. Comput. Sci., Univ. Regina, Regina, SK, Canada, Tech. Rep. CS-2012-01, 2012.

[41] M. Wang, G.-Y. Yang, J.-K. Lin, S.-H. Zhang, A. Shamir, S.-P. Lu, and S.-M. Hu, "Deep online video stabilization with multi-grid warping transformation learning," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2283–2292, May 2019.

[42] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K., Aug. 2020, pp. 402–419.

[43] V. N. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, "Study the influence of normalization/transformation process on the accuracy of supervised classification," in *Proc. 3rd Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Aug. 2020, pp. 729–735.

[44] Y. Ren, L. Sun, G. Wu, and W. Huang, "DIBR-synthesized image quality assessment based on local entropy analysis," in *Proc. Int. Conf. Frontiers Adv. Data Sci. (FADS)*, Oct. 2017, pp. 86–90.

[45] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

DONGGOO KANG was born in Seoul, South Korea, in 1992. He received the B.S. degree in financial economics from Seokyeong University, South Korea, in 2018, and the M.S. degree in AI imaging from Chung-Ang University, South Korea, in 2020, where he is currently pursuing the Ph.D. degree in AI imaging. His research interests include computational photography and human-object interaction discovery.

YEONGJOON KIM received the B.S. degree in business administration from Kookmin University, South Korea. He is currently pursuing the M.S. degree in artificial intelligence with Chung-Ang University. His research interests include computer vision, image segmentation, and meta-learning (zero/few-shot learning) for domain adaptation.

SUNKYU KWON was born in Seoul, South Korea, in 1997. He received the B.S. degree in software engineering from Catholic Kwandong University, South Korea, in 2022. He is currently pursuing the M.S. degree in artificial intelligence with Chung-Ang University. His research interests include video stabilization and video frame interpolation.

HYUNCHEOL KIM received the B.A. degree from Incheon National University, South Korea, in 1996, the M.S. degree in computer science from the New Jersey Institute of Technology (NJIT), in 2005, and the Ph.D. degree in image engineering from Chung-Ang University, in 2015. His research interests include artificial intelligence, video enhancement and restoration, visual tracking, object detection and recognition, and scene-based video summarization.

JINAH KIM was born in Cheonan-si, South Korea, in 1989. She received the B.A. degree from the Department of Film and Video, Dongguk University, South Korea, in 2012. She started working on improving video quality, in 2014. She is currently working in the field related to video quality for approximately a decade.

JOONKI PAIK (Senior Member, IEEE) was born in Seoul, South Korea, in 1960. He received the B.S. degree in control and instrumentation engineering from Seoul National University, in 1984, and the M.S. and Ph.D. degrees in electrical engineering and computer science from Northwestern University, USA, in 1987 and 1990, respectively.

He began his career at Samsung Electronics, from 1990 to 1993, where he played a key role in designing image stabilization chipsets for consumer camcorders. In 1993, he joined as a Faculty Member of Chung-Ang University, Seoul, where he is currently a Professor with the Graduate School of Advanced Imaging Science, Multimedia, and Film. From 1999 to 2002, he was a Visiting Professor with the Department of Electrical and Computer Engineering, The University of Tennessee, Knoxville. Since 2005, he has been the Director of the National Research Laboratory, South Korea, specializing in image processing and intelligent systems. He was the Dean of the Graduate School of Advanced Imaging Science, Multimedia, and Film, from 2005 to 2007, and concurrently served as the Director of the Seoul Future Contents Convergence Cluster. In 2008, he was a full-time Technical Consultant with the Systems LSI Division, Samsung Electronics. Here, he developed various computational photographic techniques, including an extended depth of field system.

Dr. Paik has had a notable influence in scientific and governmental circles in South Korea. He is a member of the Presidential Advisory Board for Scientific/Technical Policy with the Korean Government and serves as a Technical Consultant for Computational Forensics with the Korean Supreme Prosecutor's Office. His accolades include being a two-time recipient of the Chester-Sall Award from the IEEE Consumer Electronics Society. He has also received the Academic Award from the Institute of Electronic Engineers of Korea and the Best Research Professor Award from Chung-Ang University. He has actively participated in various professional societies. He served the Consumer Electronics Society of the IEEE in several capacities, including as an Editorial Board Member, the Vice President of International Affairs, and the Director of Sister and Related Societies Committee. In 2018, he was appointed as the President of the Institute of Electronics and Information Engineers. Since 2020, he has been the Vice President of Academic Affairs with Chung-Ang University. In an exceptional move, in 2021, he simultaneously assumed the roles of Vice President of Research and Dean of the Artificial Intelligence Graduate School, Chung-Ang University, for a one-year term. Expanding his scope of responsibilities, in 2022, he accepted a five-year appointment as the Project Manager for the Military AI Education Program under Korea's Department of Defense. With a career spanning over three decades, he has made significant contributions to the fields of image processing, intelligent systems, and higher education.