applied sciences

MDPI

Article

Dual-Modality Cross-Interaction-Based Hybrid Full-Frame Video Stabilization

Jaeyoung Jang [1,†], Yuseok Ban [1,†] and Kyungjae Lee [2,*]

[1] Department of Electronics Engineering, Chungbuk National University, 1 Chungdae-ro, Seowon-gu, Cheongju 28644, Republic of Korea; wodud1733@chungbuk.ac.kr (J.J.); ban@cbnu.ac.kr (Y.B.)
[2] School of Artificial Intelligence, Yong In University, 134 Yongindaehak-ro, Cheoin-gu, Yongin 17092, Republic of Korea
* Correspondence: kjlee@yongin.ac.kr; Tel.: +82-31-8020-3696
† These authors contributed equally to this work.

Abstract: This study aims to generate visually useful imagery by preventing cropping while maintaining resolution and minimizing the degradation of stability and distortion to enhance the stability of a video for Augmented Reality applications. The focus is placed on conducting research that balances maintaining execution speed with performance improvements. By processing Inertial Measurement Unit (IMU) sensor data using the Versatile Quaternion-based Filter algorithm and optical flow, our research first applies motion compensation to frames of input video. To address cropping, PCA-flow-based video stabilization is then performed. Furthermore, to mitigate distortion occurring during the full-frame video creation process, neural rendering is applied, resulting in the output of stabilized frames. The anticipated effect of using an IMU sensor is the production of full-frame videos that maintain visual quality while increasing the stability of a video. Our technique contributes to correcting video shakes and has the advantage of generating visually useful imagery at low cost. Thus, we propose a novel hybrid full-frame video stabilization algorithm that produces full-frame videos after motion compensation with an IMU sensor. Evaluating our method against three metrics, the Stability score, Distortion value, and Cropping ratio, results indicated that stabilization was more effectively achieved with robustness to flow inaccuracy when effectively using an IMU sensor. In particular, among the evaluation outcomes, within the "Turn" category, our method exhibited an 18% enhancement in the Stability score and a 3% improvement in the Distortion value compared to the average results of previously proposed full-frame video stabilization-based methods, including PCA flow, neural rendering, and DIFRINT.

Keywords: video stabilization; dual modality; cross-interaction; IMU sensor; neural rendering; full-frame video; augmented reality

Citation: Jang, J.; Ban, Y.; Lee, K. Dual-Modality Cross-Interaction-Based Hybrid Full-Frame Video Stabilization. *Appl. Sci.* **2024**, *14*, 4290. https://doi.org/10.3390/app14104290

Academic Editor: Andrea Prati

Received: 20 February 2024
Revised: 6 April 2024
Accepted: 13 May 2024
Published: 18 May 2024

1. Introduction

As Augmented Reality (AR) technologies, particularly those utilizing Head Mounted Displays (HMDs), become increasingly integrated with our reality, the necessity for video stabilization is becoming more pronounced [1,2]. This technology, essential for correcting the shake induced by user-handled cameras, has been progressively developed since the early 2000s alongside the rapid advancement of camera technology. The need for video stabilization stems from the advancements in image sensor fabrication technology, which, while improving resolution, also amplifies the challenge of identifying images without post-correction due to significant shaking. Employed not only in everyday smartphone cameras but also in action cams that are mounted on displaying devices through modules known as image stabilizers, the demand for this technology continues to grow, reflecting an expanding market size. This trend underscores the critical role of video stabilization in enhancing the user experience in AR environments, ensuring seamless integration of virtual and physical worlds.

Highlighting the necessity of research, this paper addresses the limitations of existing video stabilization techniques. Video stabilization can be divided into three main processes: motion estimation that estimates the movement of objects, motion smoothing that smoothens the path of object movement, and stable frame generation that creates stabilized frames using calculated values. Previously, visual tracking technologies like the KLT tracker [3] based on good features to track have enhanced motion estimation performance, and motion smoothing methods such as robust L1 optimal camera paths [4] or Kalman filters have been developed. The process involves warping fields, calculated to transform frames, leading to two primary issues: cropping, which reduces resolution by cutting off image edges during the warping process, and distortion, which causes pixel value distortions leading to visual artifacts like blur and wobble.

Firstly, the rolling shutter effect, due to the type of a recording mechanism of image sensors, causes the top of the frame to be recorded slightly earlier than the bottom one recording mechanism, when scanning from the top to the bottom of a sensor (see Figure 1). This results in a stretching appearance and can lead to more pronounced shake in the footage when the camera is in motion. The rolling shutter effect introduces various visual artifacts, including wobble, skew, and blur, necessitating the need for motion compensation.

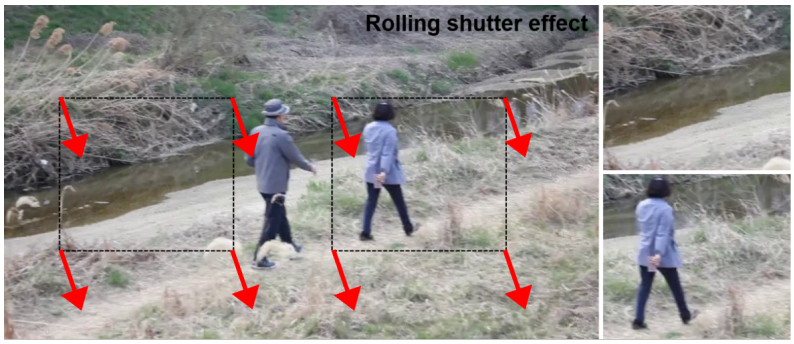


Figure 1. Example of rolling-shutter effect.

Secondly, cropping occurs, a phenomenon where the edges of the video are cut off, reducing the resolution, due to the warping process based on the pixel values between adjacent frames (see Figure 2). This reduction in resolution implies a loss of information at the video's periphery.

Figure 2. Example of cropping: (**left**) input, (**right**) cropped.

Thirdly, distortion manifests as a spatial warping or stretching of the image, creating an illusion of fluttering, primarily due to camera movement, vibration, or the rolling-shutter effect (see Figure 3). This is further compounded by visual artifacts such as shakiness, blur, and wobble.

Figure 3. Example of distortion, (**left**) input, (**right**) distorted.

To alleviate those issues, video stabilization techniques are broadly used and can be categorized into mechanical and digital methods, with mechanical stabilizers like gimbals and Optical Image Stabilizers (OIS). These methods have drawbacks related to cost and bulkiness. Despite the surge in technology development with the advent of deep learning, digital methods still face performance limitations. Our approach uses only the camera's visual sensor in a digital method, enhanced by the Inertial Measurement Unit (IMU) sensor found in most mobile electronic devices today, reducing additional costs and improving performance over existing digital methods to provide visually useful imagery. To mitigate performance degradation caused by the aforementioned issues, our method first uses an IMU sensor to reduce the rolling-shutter effect due to dynamic motion. This advantage allows motion compensation using input frame and sensor data to lessen the rolling-shutter effect. Secondly, to address cropping, a process is required to fill unknown pixel value areas with surrounding regions to create a full-frame image of the same size as the input frame. However, dynamic scenes introduce additional factors like lighting changes that cause pixel value changes, and panorama image stitching for frame rendering can severely exacerbate visual artifacts, necessitating the application of neural rendering [5] using convolutional and Encoder–Decoder-based networks for performance enhancement. Previous research studies have developed technologies for video stabilization using IMU sensors or creating full-frame images with a Cropping ratio of 1 using deep learning-based methods. Our video stabilization algorithm first synchronizes IMU sensor data with the video, receiving timestamp (s), gyroscope (rad/s), accelerometer (m/s^2), and magnetometer (μT) values, and applies motion compensation to input video frames using the Versatile Quaternion-based Filter (VQF) [6] algorithm and AKAZE-based [7] optical flow. Then, to improve cropping, PCA-flow-based video stabilization [8] is performed. Final stabilized frames are produced by applying neural rendering to address distortion occurring during the full-frame video creation process. Our method, which performs motion compensation using sensor data followed by deep learning-based full-frame video stabilization, represents a novel hybrid approach to video stabilization not previously explored. To quantitatively assess our method, we employ various metrics: the Stability score, indicating video stability; the Distortion value, depicting the extent of video deformation or alteration; the Cropping ratio, denoting the proportion of peripheral areas removed during video stabilization. Furthermore, we evaluated visual quality with the following metrics: the LPIPS assesses alignment with human visual perception; the SSIM measures structural similarity between images; the PSNR gauges image quality loss. The specifics of these evaluation metrics are elaborated in Section 4.2. The applicability and expected effects of our proposed technique include the following: First, correcting video shake to contribute to an accurate stabilization making it valuable across various industries. Second, it offers the advantage of producing visually useful imagery without the bulkiness and cost issues associated with the usages of gimbals and OIS.

In summary, the contributions of this paper are as follows:

- For the first time, we combined an IMU sensor with a deep learning-based full-frame video stabilization method, demonstrating an increase in stability.
- To address the main issues of video stabilization, such as cropping and distortion degradation, we integrated PCA flow and neural rendering.
- Our technology contributes to correcting video shake for accurate target detection and tracking and has the advantage of generating visually high-quality videos at low cost.

2. Background

2.1. Motion Estimation

Optical flow is a technique commonly used to estimate the motion of objects between video frames. Based on the calculated warping field, it contributes to compensating for this information in subsequent processing stages, making it widely employed across various fields. Motion estimation techniques employing optical flow can be broadly categorized into sparse and dense approaches. Sparse optical flow defines and detects features such as ORB corner points [9], subsequently estimating motion using a KLT tracker based on these detected outcomes. Conversely, dense optical flow provides information about the magnitude and direction of pixel movement, performing motion estimation without relying on feature bases. Although dense optical flow boasts high accuracy, its comprehensive computation across unnecessary areas results in slow processing speeds. To overcome these limitations, RAFT (Recurrent All-Pairs Field Transforms) optical flow [10], which utilizes dense optical flow and R-CNN (Region-based Convolutional Neural Network) features [11], has been developed. This method, structured around an R-CNN involving feature extraction, visual similarity, and iterative updates, enhances accuracy with each learning phase by repetitively updating the flow vector. Recently, research has been conducted on Gyroflow+ [12], which integrates gyroscope data with optical flow and homography. For this purpose, a self-guided fusion module and a homography decoder have been proposed. Attempts to overcome the limitations of the dense optical flow approach have continued; among them, Xiao et al. [13] experimented with a method using a module that deep-couples optical flow with deformable convolution. Specifically, they proposed a method capable of robust motion estimation even in scenarios with large motion. Additionally, in scenarios such as satellite video, an efficient computation method was developed by applying temporal difference for temporal compensation [14] as an alternative to optical flow for motion compensation.

Another method of motion estimation employs an affine transform, utilizing matrix operations for coordinate transformation between input and output. This encompasses techniques such as homography, per-pixel warp fields, and multi-grid methods. Homography applies perspective transformation matrices derived from features extracted across two planes. Per-pixel warp fields generate warp fields at the pixel level based on histogram differences, identifying similarities between feature trajectories and pixel profiles in static backgrounds and differences in dynamic objects. Multi-grid techniques learn a series of set mesh-grid transformations from previous stabilized camera frames to generate camera paths. Bundled camera paths define a bundled of spatially variant camera paths through measured local homographies, optimizing these paths for video stabilization after motion smoothing. IMU sensor-based motion estimation selects between optical flow (KLT tracker) and IMU-aided motion estimator based on the camera's angular velocity threshold. However, a limitation exists as cropping may occur in all scenarios regardless of the threshold applied. Lately, GlobalFlowNet [15], an unsupervised method for performing video stabilization, has been developed. It utilizes a foreground mask in preprocessing for robust homography-based motion estimation and employs low-level confidence features. This approach enhances the capture of consistent spatial correspondence.

2.2. Video Completion

Full-frame video stabilization with motion inpainting addresses missing areas in stabilized videos by constructing image mosaics using neighboring frames. It utilizes

local motion estimation for applying global transformation only to common coverage areas between frames and calculates optical flow between frames to eliminate unwanted motion fluctuations, thereby achieving stabilized motion paths. However, this method may produce visible artifacts in non-planar and dynamic scenes and create visible color seams when combining propagated color from different frames due to effects like lighting changes, shadows, and vignetting.

Temporally coherent completion of dynamic video presents an automatic video completion algorithm that synthesizes missing areas in videos in a temporally coherent manner. Despite limitations in handling discrepancies due to dynamically changing video frames and mismatched image-space motion vectors, this algorithm is well-suited for processing dynamic scenes captured with moving cameras. It utilizes optical flow and color, matching colors temporally using pixel-wise forward/backward flow fields, although it may not accurately repaint the screen using motion-based features. However, in videos containing rapid movements, there are difficulties in accurately estimating the flow, which results in a decline in the quality of color completion.

Flow-edge-guided video completion improves upon traditional issues of being unable to synthesize sharp flow edges and often producing over-smoothed results by jointly synthesizing colors and flow, propagating color along flow trajectories to enhance temporal consistency. This approach alleviates memory issues, allows for a high-resolution output, and avoids visible seams by operating in the gradient domain, performing video completion through dense flow fields. In some cases, such as with dynamic textures, optical flow estimation can be inaccurate, leading to visual artifacts. Additionally, image composition becomes challenging when large areas are obscured throughout the entire sequence.

2.3. View Synthesis and Rendering

Deep blending for free-viewpoint image-based rendering (IBR) addresses the difficulties of traditional IBR when moving far from input frames due to numerous visible artifacts. It employs novel view synthesis using held-out real image data to learn blending weights for combining input photo contributions. Accurate geometry provision is crucial for CNNs to find correct blending weights, yet direct blending in image space can result in visible artifacts and glitches, especially when flow estimates are unreliable. For instance, there is a limitation of flickering occurring in the resultant image when composing images with significant or inconsistent lighting differences.

Free view synthesis overcomes the limitations of traditional methods that rely on camera grid and stereo matching, which restrict the layout of input views. It generates a free view synthesis from unstructured input images of general scenes by correcting input images with Structure from Motion (SfM) and calculating 3D proxy geometry through Multi-View Stereo (MVS). Utilizing depth maps and 3D proxy geometry, it maps encoded features to the target view, blending them using an Encoder–Decoder network. Since it only composes images on a frame-by-frame basis, there is a drawback of lacking temporal consistency. Additionally, visual artifacts occur when the proxy 3D model used for mapping misses significant parts of the scene.

Out-of-boundary view synthesis towards full-frame video stabilization [16] significantly improves upon traditional grid-based and pixel-based warping methods through a two-stage coarse-to-fine method. It notably contributes to minimizing cropping in the boundary areas and reducing jitter. However, in cases of significant movement of dynamic objects between adjacent frames, the accurate generation of out-of-boundary regions may be challenging due to discontinuities.

In addition, methods based on progressive fusion and temporal fusion have been explored to reduce distortion. Jiang et al. [17] proposed a Multi-Scale Progressive Fusion Network (MSPFN) to eliminate various degrees of blurring. They generated a Gaussian pyramid of rain images and employed a coarse fusion module with Conv-LSTM to capture global textures. Subsequently, a fine fusion module was introduced to fuse correlated information in a cascading manner, forming progressive multi-scale fusion. Ultimately,

the utilization of a residual module facilitated the generation of high-quality images. Xiao et al. [18] addressed the challenge of limited and difficult-to-extract information provided by frames by proposing temporal grouping projection fusion and Multi-Scale Deformable (MSD) convolution alignment. Temporal fusion was applied to regroup continuously input frames into different poses, thereby reducing the complexity of projection while enabling the learning of more complementary information from frames. Following this, a multi-scale residual block was utilized to learn complex motion information for accurate frame alignment. Finally, a temporal attention module was employed to generate images that maintained a high level of consistency with the reference frame.

2.4. Video Stabilization Using IMU Sensors

Image deblurring using IMU sensors estimates blur function from gyroscope and accelerometer data during shooting [19]. Known blur function allows image improvement through non-blind deconvolution for deblurring. Since the algorithm assumes a constant depth of the scene, there is the limitation of an inaccurate blur estimation due to depth differences in real scenes. Digital video stabilization and rolling-shutter correction using gyroscopes measures camera motion with gyroscopes to perform digital video stabilization and rolling-shutter correction efficiently. Despite its strength under poor lighting and significant foreground motion, it may introduce shaky motion-induced visual artifacts. Deep online video stabilization using IMU sensors synthesizes stabilized images through deep motion estimation using data from IMU sensors. It identifies various motion types with a Deep Neural Network (DNN) classifier and employs Long Short-Term Memory (LSTM) [20] for extracting temporal features, effectively removing shaky artifacts, performing strongly across datasets with less time consumption, although it requires sufficient training data for accurate predictions. Deep online fused video stabilization employs both gyroscope sensor data and image content in an unsupervised learning DNN for video stabilization. The network fuses motion representations combining optical flow with real/virtual camera pose histories, where LSTM cells infer new virtual camera poses for generating warping grids to stabilize video frames. Although numerous studies are underway to enhance stabilization performance using IMU sensors, the issue of cropping still persists in video stabilization.

2.5. Limitations

Conventional methods have several limitations. Firstly, shooting with cameras that utilize sensors with a rolling-shutter mechanism, where the shutter closes sequentially, results in stretched and shaken photographs. Secondly, cropping occurs, a phenomenon where the edges of the video are cut off, reducing the resolution, due to the warping process based on the pixel values between adjacent frames. Lastly, distortion exists, a condition where pixel values are distorted due to visual artifacts including shakiness, blur, and wobble caused by the camera's movement or vibration, making the space appear twisted, elongated, and wavering.

3. Proposed Method: Cross-Interaction-Based Video Stabilization

3.1. Methodology

To address the performance issues caused by the major problems identified, we propose a combination of three technologies. This holistic approach integrates hardware and software solutions to significantly enhance the quality of video stabilization, aiming for a comprehensive improvement over existing methods. Firstly, considering the necessity for IMU sensors to perform motion compensation from shaky input data, we sought to improve motion estimation performance. It was essential to establish a portable connection between sensor data and video data, necessitating the integration of a developing board, an IMU sensor, and a camera, which operates within the given power consumption. The data could be collected once the communication between the IMU sensor and the developing board was established. To achieve our research objectives, we developed a technique that began with synchronizing sensor data with video footage using a Raspberry Pi to communicate

with an IMU sensor. This synchronization enabled precise motion compensation in the preprocessing stage for each frame of the input video. To enhance the accuracy of the sensor data, the VQF method was applied alongside a Low-Pass Filter (LPF) to correct the errors in sensor data, followed by the utilization of the AKAZE optical flow. Motion compensation was executed using the collected IMU sensor data with Gyroflow (v1.5.4) [21]. Secondly, to address cropping, we utilized PCA-flow-based video stabilization [8]. This method could also improve the slight cropping caused by motion compensation in our approach. However, applying PCA-flow-based video stabilization [8] has the disadvantage of a low visual quality due to increased distortion. Lastly, to further mitigate distortion that arose during the full-frame video creation process, we applied neural rendering [5], culminating in the production of the stabilized frames. Therefore, we achieved a final full-frame stabilization video with improved stability and distortion. This comprehensive approach aimed at addressing the key challenges in video stabilization, significantly improving the quality and usability of the output video footage. Our method represents a novel approach, utilizing sensor data for initial motion compensation followed by a deep learning-based full-frame video stabilization process, a strategy not previously explored in existing research. Additionally, our technique offers the advantage of being applicable to videos of any type, providing a versatile solution to video stabilization challenges. This comprehensive method integrates hardware and software to refine the stabilization process, ensuring the output video maintains high fidelity to the original scene dynamics while correcting for unwanted motion artifacts.

3.2. Implementation

3.2.1. Stage 1: Motion Compensation

We began by synchronizing sensor data with video footage through communication between the IMU sensor and a Raspberry Pi. We utilized the ROBOR RB-SDA-V1 IMU sensor for data collection, setting the initial state by resetting angle and position, and collecting timestamp, gyroscope, accelerometer, and magnetometer data. Given the original video data's specifications of a 1920 $\times$ 1080 resolution at 25 fps, we fixed the transmission period for sensor data collection at 40 ms to ensure consistency in data points. The sensor data utilized for motion compensation underwent noise reduction through a VQF filter. Gyroflow [21], utilizing gyroscope data for stabilizing a video along with optional accelerometer and magnetometer data, was chosen for motion compensation. Motion compensation at the image level was performed using the AKAZE optical flow [7] with sensor data from which noise had been removed. Subsequently, to maintain execution speed while applying neural rendering [5], we used $ffmpeg$ [22] to resize the video data from 1920 $\times$ 1080 (1080p) to 854 $\times$ 480 (480p) as part of the preprocessing steps.

3.2.2. Stage 2: Video Stabilization

Previous methods have attempted to implicitly learn frame motion in color videos through learning-based video stabilization. The learning video stabilization using optical flow [8] method proposes a novel neural network that infers per-pixel warp fields for video stabilization from the optical flow fields of input videos. This technique relies on the optical flow for motion analysis and directly learns stabilization, proposing a pipeline for motion inpainting and warp-field smoothing, offering resilience to moving objects, occlusion, and optical flow inaccuracies. The first step involves preprocessing through computing an affine transformation to eliminate large motions. It then generates masks indicating accurate areas of optical flow for each frame and employs the first five principal components suggested by the PCA flow for inpainting inaccuracies. The absence of masking and inpainting would lead to motion discontinuity and distortion. The stabilization network, as the second step, takes the inpainted optical flow field as network input, generating per-pixel warp fields for frame motion compensation. The third step, postprocessing, considers the potential issues of local discontinuities at valid/invalid boundaries of the optical flow, ensuring warp field continuity and replacing raw warp fields with resulting low-frequency fits.

The output, utilizing a low-frequency warp field for warping the input video, results in a rectangularly cropped image. While this technique achieves full-frame image generation and high stability, it has the drawback of a significant distortion. We prioritized video completion using this technique as the initial step.

3.2.3. Stage 3: Neural Rendering

Applying a motion inpainting algorithm for full-frame image generation introduces distortion. To address this visually uncomfortable effect, a hybrid neural fusion for full-frame video stabilization [5] was developed by applying neural rendering. Traditional methods for improving distortion encode CNN features and perform fusion in the feature space before transforming the fused features into output frames through a decoder, often producing overly blurred images. This method extracts abstract image features, fuses warped features from multiple frames, and decodes them along with fused feature maps for each source frame into output frames and related maps, employing a weighted averaging of generated images for rendering the final output frame. This enhances the output frame's sharpness while avoiding ghosting and glitch artifacts. Additionally, it does not require retraining and can be applied even when accurate camera poses are difficult to obtain. In other words, it applies PCA-flow-based video stabilization [8] to improve cropping results in the loss of high-frequency components and the occurrence of distortion. To recover this, the residual details of the frame are extracted and added to the fused image. The structure used for the image fusion is an Encoder–Decoder network utilizing ResNet blocks and an average pooling layer. The description of the three stages is visually presented in Figure 4.

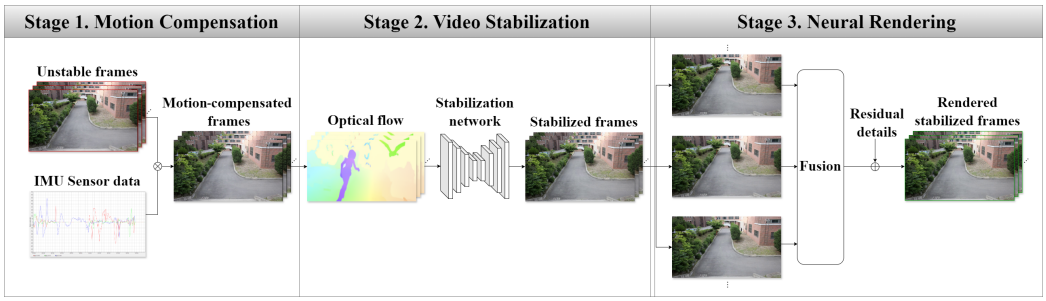


Figure 4. The overall pipeline of the proposed cross-interaction-based hybrid full-frame video stabilization framework. (**Stage 1**) We first perform motion compensation using IMU sensor data. (**Stage 2**) Then, the motion compensated frames are estimated using the optical flow and these are input into a stabilization network to create stabilized frames. (**Stage 3**) Finally, multiple frames are fused and residual details are combined to generate the final rendered stabilized frames.

4. Experiment

4.1. Environment

In this study, we employed a hardware and software setup consisting of a Raspberry Pi 4B (Raspberry Pi Foundation, Cambridge in England) running Ubuntu 20.04, a RB-SDA-V1 IMU sensor (ROBOR, Ansan in Republic of Korea), an GeForce RTX 3080 GPU (NVIDIA, Santa Clara in United States) with CUDA support, and a Canon EOS 60D DSLR camera equipped with an EF-S 18–55 mm IS STM lens (Canon, Tokyo in Japan). The hardware configuration was designed for the simultaneous collection of IMU sensor data and video data, attaching the Raspberry Pi and the IMU sensor to the DSLR. The Raspberry Pi utilized Ubuntu 20.04 as its operating system and communicated with the RB-SDA-V1 IMU sensor via a USB port (see Figure 5).

Figure 5. Hardware configuration.

The collected CSV sensor data were converted into GCSV format, which is recognizable by Gyroflow [21], to apply motion compensation to the original video data. As part of the preprocessing steps, the video resolution was resized from 1080p to 480p to accelerate the processing. Subsequently, we utilized the computing power of the NVIDIA RTX 3080 GPU to perform full-frame video stabilization, leveraging the PCA-flow-based video stabilization [8] and neural rendering [5] algorithms.

4.2. Evaluation Metrics

This setup was aimed at testing the efficacy of our proposed method in improving video stability while minimizing the visual artifacts typically associated with rolling-shutter effects and camera movements. By integrating IMU sensor data for motion compensation and applying advanced video processing techniques, we sought to achieve a high level of stabilization without the loss of video quality due to cropping or distortion. The experiments were designed to validate the performance improvements in video stabilization, as facilitated by our hardware and software integration, and to demonstrate the potential of our approach for real-world applications. We sought to quantitatively assess the improvements made to address the key issues identified by applying three major metrics used in the literature [5,23,24]: the Stability (S) score, the Distortion value (D), and the Cropping ratio (C).

Firstly, the Stability score (S) serves as an indicator of a video's steadiness. The computation process involves calculating the accumulated optical flow, then transforming it into the frequency domain using a Fast Fourier Transform (FFT), and finally, calculating the ratio of selected energy over total energy for quantitative assessment. Here, the accumulated optical flow represents the cumulative path of pixel movement from previous to current frames, based on optical flow vectors, with the initial frame as a reference point. Secondly, the Distortion value (D) indicates the degree of deformation within a video. This metric requires aligning the input with the stabilized output to estimate the extent of transformation, utilizing transformation matrices for this estimation. The calculation of the Distortion value is defined as the worst ratio of the two largest eigenvalues of the affine component across all frames of the video. Thirdly, the Cropping ratio (C) quantifies the proportion of the area lost during the warping process between the input and the stabilized output.

To quantitatively evaluate how much the visual quality had improved due to mitigation of visual artifacts, we utilized metrics such as the LPIPS (Learned Perceptual Image Patch Similarity), SSIM (Structural Similarity Index Measure) [25], and PSNR (Peak Signal-to-Noise Ratio). The LPIPS, introduced by Zhang et al. [26], measures the perceptual similarity between images. This metric was proposed to address the discrepancy between human perception and the results obtained from high-resolution images by traditional methods like the PSNR and SSIM. By extracting image features using a pretrained deep neural network and calculating the distance between these features, the LPIPS assesses the perceptual differences between two images. Importantly, using the intermediate layers of a

pretrained neural network allows for the capture of high-level image features, enabling a more accurate measurement of perceptual similarity. The SSIM is a method used in digital image processing to compare two images by measuring their structural similarity. Unlike the PSNR, the SSIM takes into account the characteristics of the human visual system to evaluate image quality. The PSNR measures the objective difference based on the square of errors, whereas the SSIM evaluates how well the image quality matches the subjective human visual experience, addressing issues where images may appear similar but are distinguished significantly by the PSNR. The PSNR represents the ratio of the maximum possible power of a signal to the power of corrupting noise. It is widely used in image and video compression to quantify the difference between two images, measuring the quality of compressed images with higher PSNR values indicating better image quality. PSNR calculations are based on the Mean Squared Error (MSE) between the original and compressed images and are typically expressed in decibels (dB).

The Learned Perceptual Image Patch Similarity (LPIPS) metric is often used to assess the perceptual similarity between two images. It is a sophisticated metric capturing more nuanced differences that are perceptible to the human eye.

$$\text{LPIPS}(x, y) = \sum_{l} \frac{1}{H_l W_l} \sum_{h,w} \|\phi_l(x)_{h,w} - \phi_l(y)_{h,w}\|_2^2 \tag{1}$$

where x and y are the images being compared, l indexes layers in a pretrained network, $\phi_l(x)$ and $\phi_l(y)$ are the feature maps at layer l for each image, H_l and W_l are the dimensions of the feature maps at layer l, and $\|\cdot\|_2$ denotes the Euclidean (L2) norm.

The Structural Similarity Index (SSIM) is used for measuring the similarity between two images. The SSIM index is a full reference metric; in other words, it measures the image quality based on an initial uncompressed or distortion-free image as a reference. The SSIM considers changes in structural information, luminance, and contrast, rather than aggregating errors over pixels like traditional methods such as the MSE.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2}$$

where μ_x is the average of x, μ_y is the average of y, σ_x^2 is the variance of x, σ_y^2 is the variance of y, σ_{xy} is the covariance of x and y, $C_1 = (k_1 L)^2$ and $C_2 = (k_2 L)^2$ are two variables to stabilize the division with a weak denominator, and L, k_1, and k_2 define the dynamic range of the pixel-values.

The Peak Signal-to-Noise Ratio (PSNR) is used as a measure of quality for images and videos. the PSNR represents a ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. It is often expressed on the logarithmic decibel scale.

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{3}$$

where MAX_I is the maximum possible pixel value of the image (e.g., for an 8-bit-per-channel image, $MAX_I = 255$), MSE is the Mean Squared Error between the original and compressed image, defined as:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - K(i, j)]^2 \tag{4}$$

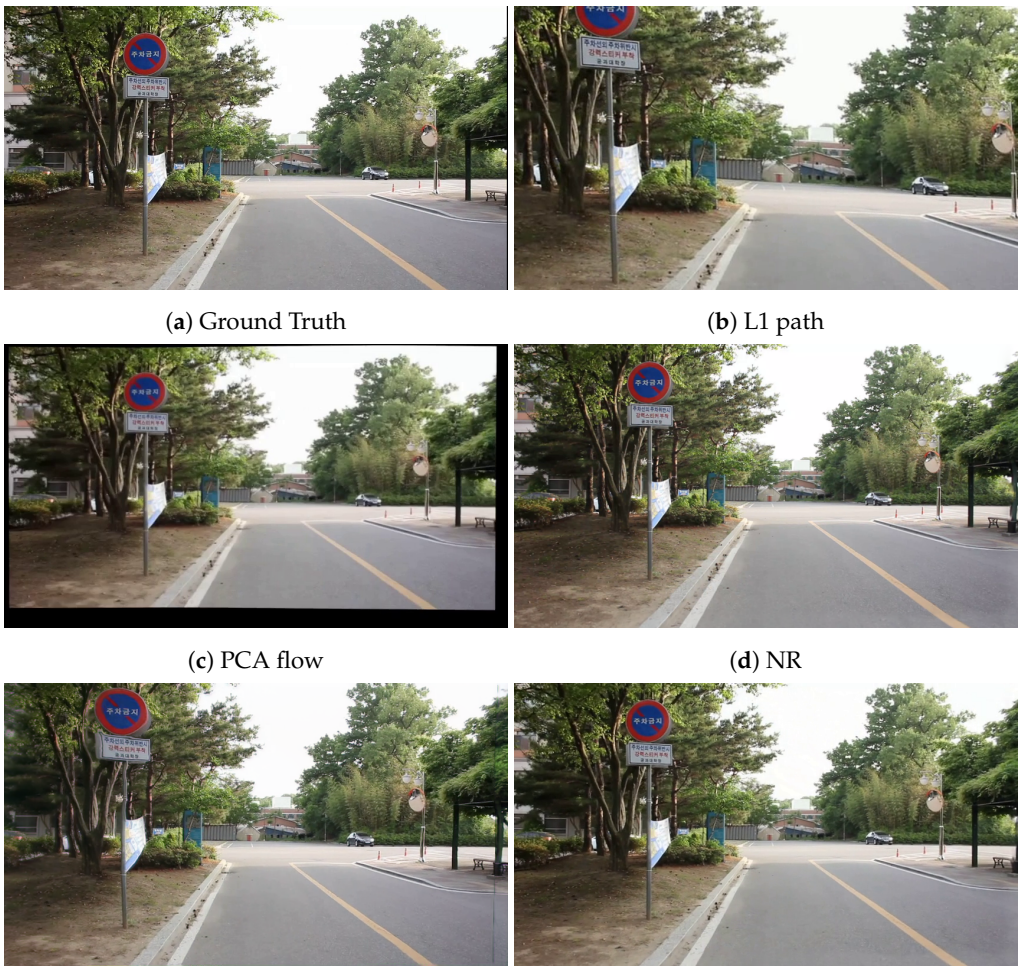
where $I(i, j)$ is the pixel value of the original image at position (i, j), $K(i, j)$ is the pixel value of the compressed image at position (i, j), and m and n are the dimensions of the images.

4.3. Result

4.3.1. Qualitative Results

To rigorously assess enhancements in full-frame video reconstruction, along with mitigation in cropping, distortion, and reductions in visual artifacts, we executed a comprehensive visual analysis and engaged in an in-depth discussion of the findings. The empirical outcomes, as illustrated in Figure 6, exhibited a pronounced divergence from the L1 path methodology [4], which is notably susceptible to cropping anomalies. Conversely, our novel methodology adeptly ensured the generation of complete full-frame visuals, circumventing the loss of pivotal visual information. Furthermore, despite maintaining a consistent ratio between the ground truth and our rendered outcomes, we observed a subtle deviation in the field of view, attributable to the methodological integration of sequential data via the fusion of multiple frames, refined through warping with the RAFT optical flow [10].

A detailed visual inspection within the "Stairs" category revealed that although PCA-flow-based video stabilization [8] significantly addressed the cropping issues inherent in conventional methodologies such as the L1 path [4], it inadvertently amplified the distortion manifested through texture seams and blurring (refer to Figure 7). Notably, our approach manifested superior visual quality by strategically employing both motion compensation and neural rendering techniques.

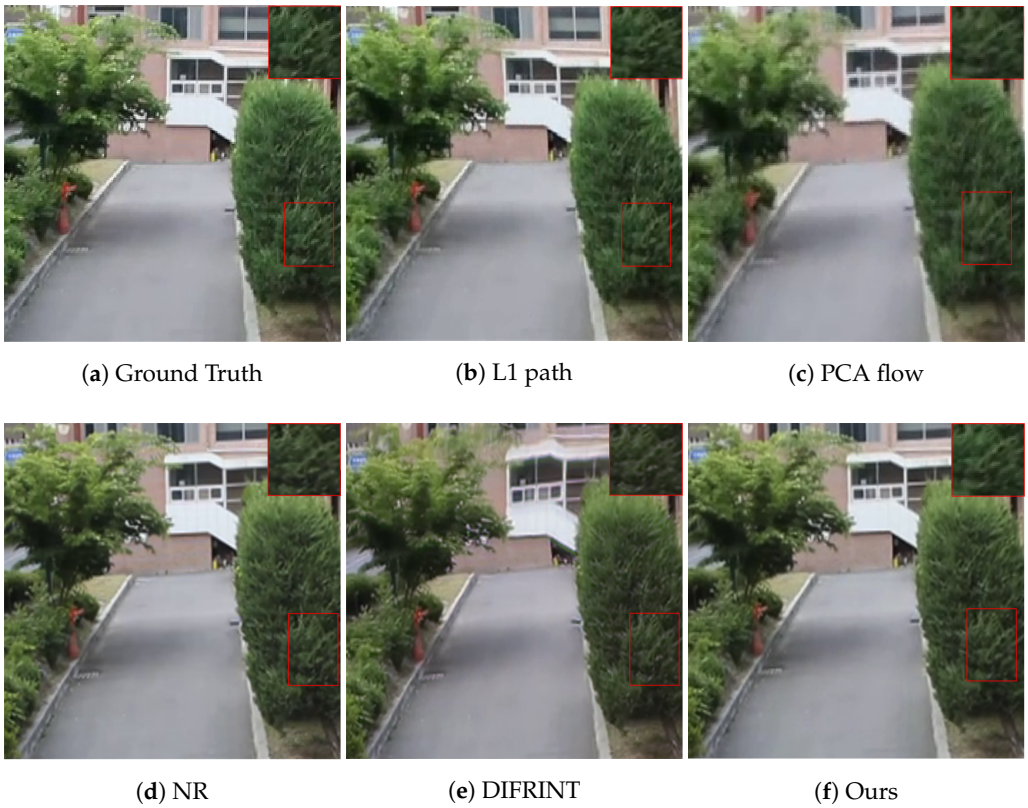


(**a**) Ground Truth (**b**) L1 path

(**c**) PCA flow (**d**) NR

(**e**) DIFRINT (**f**) Ours

Figure 6. Qualitative comparison results (category: Run).



(**a**) Ground Truth (**b**) L1 path (**c**) PCA flow

(**d**) NR (**e**) DIFRINT (**f**) Ours

Figure 7. Qualitative comparison results (category: Stairs).

In our empirical analysis, we particularly concentrated on discrete segments within the test footage to critically evaluate the efficacy of our motion compensation technique. This intensive examination was aimed at ascertaining the capability of our technology in restoring complex details typically obliterated due to camera motion or instability. Implementing our proposed methodology resulted in a discernible enhancement in maintaining fine details, such as the distinct morphology characteristics of foliage, the precise grid patterns on masonry, and the intricate textures of terrestrial surfaces (as depicted in Figure 8). These insights were corroborated by robust experimental validation, signifying that our motion compensation methodology substantially augmented the visual lucidity of these aspects. The revitalization of these intricate details within the visual footage accentuated the proficiency of our methodology in preserving the fidelity of intricate scenes, ensuring that both natural and architectural elements were accurately rendered and readily identifiable, thereby delineating our endeavor to advance the frontiers of video stabilization technologies.

Figure 8. Qualitative analysis of our result on specific regions: (**left**) Tree, (**center**) Brick, (**right**) Sand regions.

4.3.2. Quantitative Results

To quantitatively assess the extent of improvements made against the primary issues arising during video stabilization, we analyzed these improvements using the Stability score (S), Distortion value (D), and Cropping ratio (C) metrics. The Stability scores range

between zero and one, with values closer to one indicating a better quality of stabilization result. Distortion values also range between zero and one, with values closer to one signifying less flutter and more visual comfort. Cropping ratios, similarly, range between zero and one, with values closer to one indicating lesser cropping compared to the original footage. Experimental results are presented in Table 1 and show that our results employing IMU sensor for motion compensation as well as applying neural rendering increased the Cropping ratio by 36.2% per category (Stairs, Walk, Run, and Turn) compared to the L1 path method [4], generating full-frame images (see Figure 9). Secondly, our method produced videos with reduced visual artifacts, with a Distortion value increasing by 0.4–7.5% per category compared to the PCA-flow-based video stabilization method [8]. Our method obtained a better Stability score, which indicated more stable footage. Notably, traditional techniques such as the L1 path method, despite a good Stability score, suffered from a significant lower Distortion value and Cropping ratio due to nonlinear wobble and warping, leading to poor visual quality.

Table 1. Quantitative comparison according to the average result of all four categories. The first rank is marked in red color, while the second rank is marked in blue color.

Method	Stability (S) ↑	Distortion (D) ↑	Cropping (C) ↑
L1 path	0.890	0.636	0.734
PCA flow	0.786	0.896	1.000
NR	0.776	0.940	1.000
DIFRINT	0.786	0.911	0.962
Ours	0.815	0.924	1.000

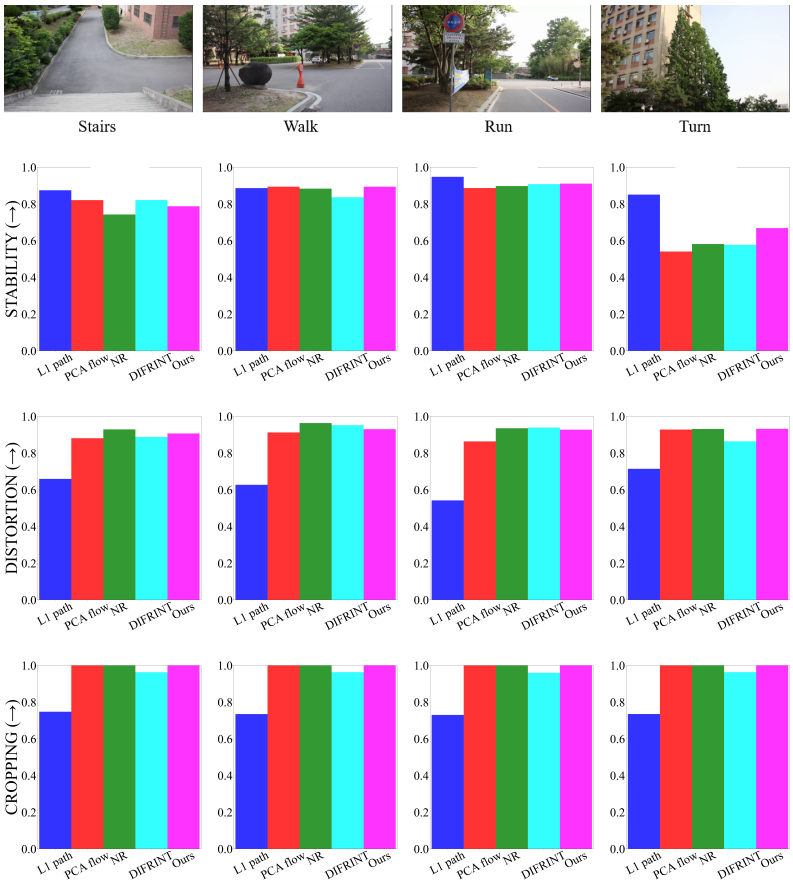


Figure 9. Quantitative comparison using four different categorical data.

Furthermore, we employed metrics such as the LPIPS, SSIM, and PSNR to quantitatively evaluate the extent of improvement in visual quality resulting from the video stabilization. LPIPS values closer to zero reflect a better perceptual similarity and alignment with human perception, especially for images with fine changes or complex textures. SSIM values closer to one suggest a greater similarity between two images, while the PSNR, used to measure the quality between the original and compressed images, indicates better image quality with higher values. The experimental results, utilizing the data from the "Stairs" category and evaluating the methods based on patches such as "Tree", "Brick", and "Sand" (see Figure 8), showed our method outperforming all methods in LPIPS, SSIM, and PSNR metrics for "Tree" (see Table 2). SSIM and PSNR values notably dropped for the L1 path and PCA-flow-based methods when adopting the IMU sensor compensation, possibly due to the preprocessing for motion compensation with the sensor output before applying the algorithm, leading to pixel value discrepancies in the input frames due to distortion.

However, with respect to the computational efficiency, our method prioritized compensation using the IMU sensor and underwent processes involving PCA flow and neural rendering, resulting in a limitation of the extended runtime. As faster methods, the L1 path and DIFRINT methods ranked first and second, respectively (as shown in Table 3). Consequently, research conducted to date has primarily focused on verifying the potential for performance. Future efforts will necessitate improvements in lightweight and speed enhancements.

Table 2. Quantitative comparison on a specific region. The first rank is marked in red color, while the second rank is marked in blue color.

Metric	Tree			Brick			Sand		
	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑
L1 path	0.174	0.294	20.740	0.118	0.353	15.924	0.071	0.213	22.820
PCA flow	0.106	0.418	23.496	0.138	0.407	17.296	0.046	0.383	24.783
NR	0.068	0.558	24.271	0.021	0.866	24.853	0.016	0.818	28.711
DIFRINT	0.086	0.579	23.774	0.069	0.591	18.962	0.029	0.629	25.535
Ours	0.040	0.856	28.689	0.031	0.846	22.771	0.024	0.744	27.481

Table 3. Computational time comparison according to the average result of all four categories. The first rank is marked in red color, while the second rank is marked in blue color.

Method	Computational Time Per Frame (Second) ↓
L1 path	0.674
PCA flow	1.512
NR	5.400
DIFRINT	0.677
Ours	5.669

5. Conclusions

The primary objective of our research was to generate visually useful imagery for Augmented Reality applications by preventing cropping through resolution maintenance and enhancing stabilization performance by minimizing the degradation of stability and distortion. In pursuit of this goal, we maintained a focus on improving performance while adequately preserving execution speed, leading to the proposal of a novel hybrid full-frame video stabilization algorithm, based on dual-modality cross-interaction using neural rendering, not previously explored. Our method was evaluated using stability, distortion, and cropping metrics, demonstrating enhanced stabilization, especially when utilizing an IMU sensor to robustly counter flow inaccuracies. Overall, our method uniquely achieved TOP2 in the S/D/C metric and showed the most significant improvement in Turn envi-

ronments. Further, the visual quality induced by the Distortion value was quantitatively compared using the LPIPS, SSIM, and PSNR metrics, providing a detailed analysis. The use of the combination of an IMU sensor and neural rendering technique showed that while maintaining the Distortion value, it resulted in increased Stability score outcomes, effectively reducing visual artifacts caused by shaking. Our technology has proven valuable in correcting video shake, contributing to accurate target detection and tracking.

The application of our technology is contingent upon the availability of IMU sensor values measured concurrently with the original video capture, which presents a limitation in terms of applicability to the vast array of videos available online. Despite this challenge, given that modern AR/VR devices and smartphones are inherently equipped with IMU sensors, leveraging these to acquire videos and applying our technology can result in visually superior, stabilized videos. Additionally, since the warping field depends on optical flow and the IMU sensor values are globally reflected across frames, visual artifacts may still persist at the video edges.

Looking forward, we propose three areas for future work. First, we aim to achieve superior motion compensation by fusing the transformation matrices generated by deep learning-based optical flow calculations, assigning greater weights to the transformation matrices as the IMU sensor values increase. Moreover, there is a need to locally reflect this in the frame to enhance compensation at the video edges. By applying this method, performance in motion estimation is expected to improve since it does not solely rely on optical flow, and consequently, it can reduce visual artifacts caused by large motion and environmental changes during the warping process. Second, we aim to implement superior feature extraction by adding SE (Squeeze-and-Excitation) blocks that perform actions similar to self-attention. This method allows for more accurate calculations of the optical flow across the frame while reducing the parameter size for better efficiency and overcoming accuracy degradation. Third, to ensure the generalization of the proposed method, it is necessary to enrich the dataset with data from daytime/nighttime or sunny day/rainy day/foggy day scenarios and extend the experiments. These future directions underscore our commitment to refining the balance between stabilization, visual quality, and computational efficiency, thereby pushing the boundaries of video stabilization techniques.

Author Contributions: J.J. and Y.B. equally contributed to this work by conceiving the method, analyzing the data, and writing the manuscript; J.J. performed the experiments; Y.B. and K.L. revised and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the research grant of the Chungbuk National University in 2021 and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. 2022R1F1A1073745).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all individual participants included in the study. All participants signed a consent form after the nature and possible consequences of the studies were explained.

Data Availability Statement: The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lee, J.; Hafeez, J.; Kim, K.; Lee, S.; Kwon, S. A novel real-time match-moving method with HoloLens. *Appl. Sci.* **2019**, *9*, 2889. [CrossRef]
2. Nunes, J.S.; Almeida, F.B.; Silva, L.S.; Santos, V.M.; Santos, A.A.; de Senna, V.; Winkler, I. Three-dimensional coordinate calibration models for augmented reality applications in indoor industrial environments. *Appl. Sci.* **2023**, *13*, 12548. [CrossRef]
3. Shi, J.; Tomasi. Good features to track. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 21–23 June 1994; pp. 593–600.
4. Grundmann, M.; Kwatra, V.; Essa, I. Auto-directed video stabilization with robust l1 optimal camera paths. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 225–232.

5. Liu, Y.L.; Lai, W.S.; Yang, M.H.; Chuang, Y.Y.; Huang, J.B. Hybrid neural fusion for full-frame video stabilization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 2299–2308.
6. Laidig, D.; Seel, T. VQF: Highly accurate IMU orientation estimation with bias estimation and magnetic disturbance rejection. *Inf. Fusion* **2023**, *91*, 187–204. . [CrossRef]
7. Alcantarilla, P.F.; Solutions, T. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Trans. Patt. Anal. Mach. Intell* **2011**, *34*, 1281–1298.
8. Yu, J.; Ramamoorthi, R. Learning video stabilization using optical flow. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8159–8167.
9. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
10. Teed, Z.; Deng, J. Raft: Recurrent all-pairs field transforms for optical flow. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part II 16; pp. 402–419.
11. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
12. Li, H.; Luo, K.; Zeng, B.; Liu, S. Gyroflow+: Gyroscope-guided unsupervised deep homography and optical flow learning. *Int. J. Comput. Vis.* **2024**, 1–19. [CrossRef]
13. Xiao, Y.; Yuan, Q.; He, J.; Zhang, Q.; Sun, J.; Su, X.; Wu, J.; Zhang, L. Space-time super-resolution for satellite video: A joint framework based on multi-scale spatial-temporal transformer. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *108*, 102731. [CrossRef]
14. Xiao, Y.; Yuan, Q.; Jiang, K.; Jin, X.; He, J.; Zhang, L.; Lin, C. Local-Global Temporal Difference Learning for Satellite Video Super-Resolution. *arXiv* **2023**, arXiv:2304.04421.
15. Yan, W.; Sun, Y.; Zhou, W.; Liu, Z.; Cong, R. Deep Video Stabilization via Robust Homography Estimation. *IEEE Signal Process. Lett.* **2023**, *30*, 1602–1606. [CrossRef]
16. Xu, Y.; Zhang, J.; Tao, D. Out-of-boundary view synthesis towards full-frame video stabilization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 4842–4851.
17. Jiang, K.; Wang, Z.; Yi, P.; Chen, C.; Huang, B.; Luo, Y.; Ma, J.; Jiang, J. Multi-scale progressive fusion network for single image deraining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8346–8355.
18. Xiao, Y.; Su, X.; Yuan, Q.; Liu, D.; Shen, H.; Zhang, L. Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection. *IEEE Trans. Geosci. Remote. Sens.* **2021**, *60*, 1–19. [CrossRef]
19. Liu, X.; Yang, Y.; Ma, C.; Li, J.; Zhang, S. Real-time visual tracking of moving targets using a low-cost unmanned aerial vehicle with a 3-axis stabilized gimbal system. *Appl. Sci.* **2020**, *10*, 5064. [CrossRef]
20. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
21. Adrian, E.; Chen, E. Gyroflow Project. Available online: https://gyroflow.xyz (accessed on 22 April 2023).
22. Tomar, S. Converting video formats with FFmpeg. *Linux J.* **2006**, *2006*, 10.
23. Choi, J.; Kweon, I.S. Deep iterative frame interpolation for full-frame video stabilization. *ACM Trans. Graph. (TOG)* **2020**, *39*, 1–9. [CrossRef]
24. Zhang, Z.; Liu, Z.; Tan, P.; Zeng, B.; Liu, S. Minimum latency deep online video stabilization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 23030–23039.
25. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
26. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.