OISSR: Optical Image Stabilization Based Super Resolution on Smartphone Cameras

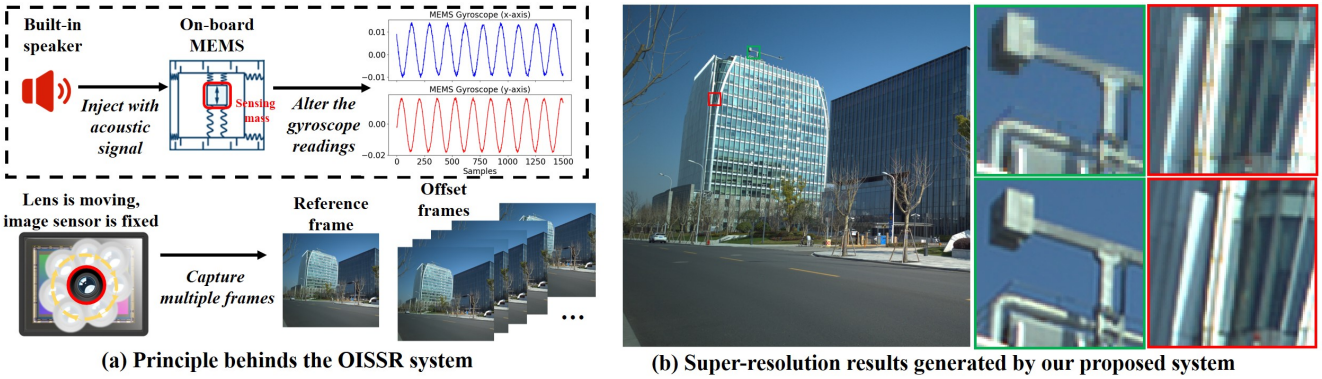
Hao Pan
Shanghai Jiao Tong University
Shanghai, China
panh09@sjtu.edu.cn

Feitong Tan
Simon Fraser University
Burnaby, Canada
feitongt@sfu.ca

Wenhao Li
Shanghai Jiao Tong University
Shanghai, China
fire1997ice@sjtu.edu.cn

Yi-Chao Chen*
Shanghai Jiao Tong University
Shanghai, China
yichao@sjtu.edu.cn

Guangtao Xue**
Shanghai Jiao Tong University
Shanghai, China
gt_xue@sjtu.edu.cn



(a) Principle behinds the OISSR system

(b) Super-resolution results generated by our proposed system

Figure 1: (a) OISSR applies acoustic injection to alter the built-in MEMS gyroscope readings to control the lens motion in OIS-supported cameras and further enables the sub-pixel alignments of multiple frames to facilitate merging into a super-resolution image. (b) The top sub-figures show the original image, whereas the bottom sub-figures show the super-resolution results obtained using the OISSR system

ABSTRACT

Multi-frame super-resolution methods can generate high resolution images by combining multiple captures of the same scene; however, the performance of merged results are susceptible to degradation due to a lack of precision in image registration. In this study, we sought to develop a robust multi-frame super resolution method (called OISSR) for use on smartphone cameras with a optical image stabilizer (OIS). Acoustic injection is used to alter the readings from the built-in MEMS gyroscope to control the lens motion in the OIS module (note that the image sensor is fixed). We employ a priori knowledge of the induced lens motion to facilitate optical flow estimation with sub-pixel accuracy, and the output high-precision pixel alignment vectors are utilized to merge the multiple frames to reconstruct the final super resolution image. Extensive experiments on a OISSR prototype implemented on a Xiaomi 10Ultra demonstrate the high performance and effectiveness of the proposed system in obtaining the quadruple enhanced resolution imaging.

*Guangtao Xue and Yi-Chao Chen are corresponding authors

MM '22, October 10–14, 2022, Lisbon, Portugal.

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547964

CCS CONCEPTS

• **Computing methodologies** → **Image processing**.

KEYWORDS

optical image stabilization; super-resolution; image registration

ACM Reference Format:
Hao Pan, Feitong Tan, Wenhao Li, Yi-Chao Chen, and Guangtao Xue. 2022. OISSR: Optical Image Stabilization Based Super Resolution on Smartphone Cameras. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisbon, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3547964

1 INTRODUCTION

Background: Software-based super-resolution is widely used in smartphone cameras to overcome physical limitations on raw spatial resolution. Single image super-resolution (SISR) algorithms [16, 29, 38] extract structure and texture components to enable the synthesis of an interpolated image with matching weights to increase resolution. However, SISR itself is an ill-posed problem, as evidenced by the fact that it utilizes only the relationship between neighboring pixels, which means that the augmented pixels are a fiction learned from "experience" [12]. By merging multiple low-resolution observations of a given scene that were taken when handholding the camera, multi-frame super-resolution (MFSR) methods generate results that are closer to the ground truth (*i.e.*, greater realism) [12, 50]. Note that the performance of MFSR methods is mainly determined by the precision of image registration and alignment accuracy at the sub-pixel scale [2, 34, 35]. Unfortunately, image registration

methods based solely on visual information are unable to achieve robust sub-pixel accuracy in real-world scenarios [44], which makes super-resolution results more like "denoising" with an accompanying blurring of high-frequency details. Deep learning-based SISR and MFSR methods [4, 12] synthesize high frequency details based on knowledge learned from a training dataset. However, the performance of these supervised learning approaches is unstable due to their strong dependence on the distribution of the training datasets.

Our system: In the current study, we develop a robust multi-frame based approach to super-resolution imaging capable of outstanding accuracy in aligning multiple captures from smartphone cameras. Specifically, we employ the lens-shift optical image stabilization (OIS) module commonly found in smartphone cameras to enable accurate control over lens motion and further employ a priori knowledge of the induced lens motion to facilitate sub-pixel alignments of the multiple captures.

Challenges: Expending the applicability of the lens-shift OIS modules to enhance the imaging resolution imposes two mainly daunting challenges: **(1)** The existing OIS modules in smartphone cameras cannot be controlled via programming (*e.g.*, Android APIs). One challenge is to devise the means to exercise control over the OIS module without modifying the hardware. **(2)** After capturing image sequences with lens motions controlled by the OIS module, like existing MFSR methods, we need to develop a high-precision image registration algorithm to facilitate the merge process for the final super-resolution image. Thus, the other challenge involves developing the means to incorporate prior lens motion information controlled by the OIS module in the image registration process to generate highly accurate sub-pixel alignment vectors.

Solutions: The aforementioned challenges are addressed as follows: **(1)** The OIS module relies on the onboard IMU (particularly, the gyroscope) to sense camera shake. Inspired by acoustic injection attacks on micro-electromechanical system (MEMS) sensors [43], we alter the gyroscope readings via sinusoidal acoustic signals with a resonance frequency close to that of the moving mass in the MEMS gyroscope (see Fig. 1(b)). This makes it possible to control the lens position and thereby sample multiple pixel patterns from a fixed image sensor. **(2)** Obtaining the detailed OIS-controlled lens shifting information can benefit the image registration process on the multiple captured frames when the lens moves. However, lens shifting is unavailable on smartphones except for the Google Pixel series [9]; therefore, we first delved into the feedback control mechanism in the OIS to model the process of converting acoustic signals into lens shifts. To ensure accurate image registration, we then design an optimization framework combining known lens shifting information with coarse pixel shift information (obtained from the optical flow method) to output high-precision pixel alignment vectors in a sub-pixel space. Note that our proposed image registration method can also correct the skew introduced by rolling shutter on the offset frames, thus, even the scene is photographed when lens continuously moving, the final super-resolution images generated by our system are not affected by the rolling shutter.

We implement the OISSR prototype on a Xiaomi 10Ultra[1], whose main camera supports OIS. Extensive experiments are conducted to assess the effectiveness of our proposed OISSR. One example of our quadruple enhanced resolution results is shown in Fig. 1(c). Comparing the top and bottom subfigures, our system can sharpen the edges with more high-frequency details and eliminate the photo noise effectively.

The main contributions of this work are as follows:

- We cleverly exploit the potential of the OIS techniques to facilitate super-resolution imaging. To the best of our knowledge, our proposed OISSR system is the first to use lens motion in the OIS module to achieve a robust MFSR technology.
- We verify the linear relationship between lens shifting and the signals used to control the OIS during the capture of multiple images. The lens shifting constraints can be leveraged in the optical flow solution and implement the sub-pixel image registration.
- We develop an optimization framework that combines lens shifting information and the coarse pixel shift to output high-precision sub-pixel alignment vectors for image registration, and obtain the high-performance super resolution images after merging multiple registered frames.

2 RELATED WORKS

2.1 Multi-frame-based Super-resolution

Compared to SISR methods which solely rely on one image prior, MFSR was prevented with the aim to merge multiple low resolution images of the same scene to reconstruct a higher resolution output [42]. In recent works [32, 50], researchers have demonstrated that MFSR is also applicable to the cameras by harnessing natural hand tremors to introduce small offsets among multiple frames. Note however that the performance of handheld MFSR methods is determined primarily by the accuracy of image registration. High frequency details beyond the limitations of the hardware can only be extracted if the low-resolution images are aligned at the sub-pixel level [2, 34, 35]. A number of image registration methods have been applied to super-resolution reconstruction algorithms. The Harris corner detector and SIFT descriptor schemes rely on image features to compensate for an inability to obtain alignment vectors on whole pixels [6]. Methods based on block matching [13] and optical flow [5, 25] depend heavily on the quality of the raw images and are unable to achieve sub-pixel alignment accuracy. Algorithms using automatically computed segmentation maps [8] and tracking algorithms [3, 10] are slow and prone to localization errors. Deep learning-based image registration has also been implemented using artificially synthesized training datasets [17, 37, 39, 45]; however, the resulting trained models perform poorly in situations involving (even slight) camera motion.

Hardware-based solutions have been developed to achieve sub-pixels displacement in handheld-based MFSR. Related works [14, 18, 31] inspired mainstream camera manufacturers to apply pixel-shift technology to commercial cameras (*e.g.*, Sony A7R III, Fujifilm GFX 100, Olympus OM-D E-M5 Mark III, and Panasonic Lumix G9) [19, 33] to produce ultra-high-resolution images from a series of captures, during which the sensor is physically shifted by a fraction of a pixel-width. Note however that the highly specialized hardware used to control the movement of the image sensor is limited only to professional-grade cameras, which are not constrained by the cost and space limitations of smartphone cameras. In the current study, we develop a novel MFSR scheme for smartphone cameras,

[1]The source code is available at: https://github.com/SolskyPan/OISSR

in which the OIS adjusts the lens positions based on the built-in MEMS gyroscope readings.

2.2 Acoustic Injection on MEMS Sensors

MEMS gyroscopes measure Coriolis acceleration by tethering the frame containing the resonating mass to a substrate using springs mounted at 90° relative to the resonating motion. Note however that these tiny mechanical structures are highly susceptible to acoustic interference at their resonance frequencies ($18KHz \sim 30KHz$), such that the sensing mass vibrates at the same frequency as the external sinusoidal sound pressure waves [40]. Researchers have previously exploited this phenomenon to attack MEMS gyroscopes. In [36], researchers demonstrated a denial of service (DoS) attack using resonant acoustic signals to facilitate the intentional crashing of drones. In [41], researchers proposed output biasing and output control attacks to compromise the integrity of MEMS accelerometer readings. In [43], researchers achieved implicit control over a variety of real-world systems via non-invasive attacks targeting embedded inertial sensors. In [27] and [1], researchers demonstrated the feasibility of using inertial sensors in smartphones to eavesdrop on speech signals. Unlike the methods described above, the system developed in this paper employs acoustic injection to alter the readings from the built-in gyroscope in order to manipulate the position of the lens in OIS-supported cameras.

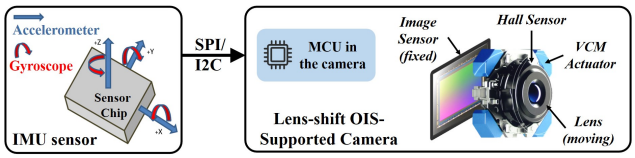
3 PRELIMINARY ANALYSIS

In this section, we first describe the architecture of the lens-shift OIS modules that are commonly found in smartphone cameras. We then demonstrate the feasibility of utilizing acoustic injection to control the motion of the lens, and delved into the detailed working principle of OIS and examine the conversion function from acoustic signals to lens motion information.

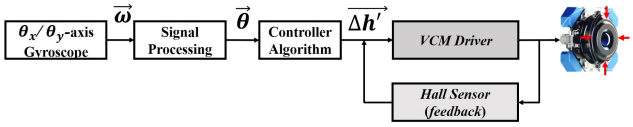
3.1 OIS Architecture Description

In such lens-shift OIS systems, as shown in Fig. 2(a), the onboard IMU sensor (*i.e.*, accelerometer and gyroscope) senses the camera shake during image acquisition, and the voice coil motor (VCM) actuator adjusts the lens position to compensate the camera shake while the image sensor is fixed to the bottom of the camera module. Note that camera shake induces both translational and rotational movements with six DOFs (*e.g.*, on the $x-/y-/z-$axis, and *Roll-/Pitch-/Yaw*-axis respectively). Of all movements, translational and rotational movements (along the x and y axes) induce more image blurring that do motions along z axis [21]. This can be attributed to the fact that the inertial forces associated with snapping a photo are generally not along the z axis, such that translational and rotational movement about z axis does not alter the imaging point on the sensor plane. Thus, the existing OIS technologies on the smartphone camera usually only consider the 4-DOF disturbances (along translational $x-/y-$axis and rotational *Roll-/Pitch*-axis) to derive the control objectives of the lens holder.

It is no doubt that the OIS module can compensate for camera translational displacement by moving the lens at the appropriate distance in opposite directions; however, the OIS actuator should also move the lens in translation to correct the camera rotational displacement. As shown in Fig. 2(b), after the offset angles $\vec{\theta}$ is



(a) Architecture of a lens-shift OIS camera



(b) Block diagram of control-loop in the OIS module

Figure 2: Working principle of the lens-shift OIS module

calculated from the gyroscope readings $\vec{w}$, the OIS controller algorithm is utilized to calculate the translational compensation vector $\vec{h}$, and VCM actuator then adjusts the lens position to compensate for involuntary jitter. The entire lens control block is implemented within a control-loop where feedback pertaining to lens movement is provided by a Hall sensor, that means the VCM actuator can always adjust the lens to the certain position according to the certain offset angle displacement. [30] presents the linear model linking $\Delta\theta$ and Δh when the camera shake is in the small range ($\Delta\theta \approx 0°$), and the formula is shown as: $\Delta h = Z_c \Delta\theta$, where Z_c is a constant.

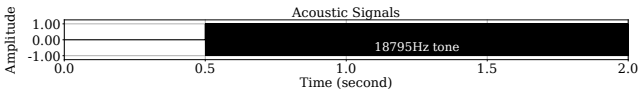
3.2 Controlling Lens via Acoustic Injection

The working principle of OIS in the above section inspires us to use the gyroscope/accelerometer to control lens motion. In Sec. 2.2, we discuss related works that utilize acoustic injection to alter the readings of MEMS sensors. In the current study, we sought to control the MEMS gyroscope readings for the reason that: the acoustic sinusoidal signals that can control gyroscope readings should be close to resonance frequency of the sensing mass, which mainly ranges from $18KHz$ to $30KHz$ [11] and is friendly and inaudible to human ears. By contrast, the acoustic signals required to affect accelerometers would be well within the audible range ($2KHz \sim 10KHz$) [41], which brings acoustic noise to the human ear.

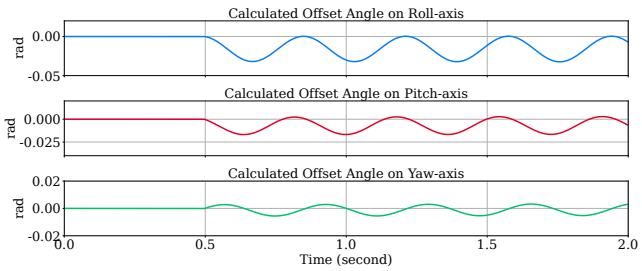
A Xiaomi 10Ultra is used as a test device here. We first identify the resonance frequency (*i.e.*, around $18.79KHz$) of the built-in MEMS gyroscope via frequency sweeping [41]. We then use the built-in speaker to play a *.wav* file of a sinusoidal acoustic signals with the same resonance frequency of the MEMS gyroscope, and the signals are shown in Fig. 3(a). Android APIs are used to collect 6-axis IMU readings at a sampling rate of 200 Hz, and the offset angles calculated from the 3-axis gyroscope readings (actual angular velocity) as follows: $\theta(t+\triangle t) = \theta(t)+\omega[t] \triangle t$, where the $\triangle t = \frac{1}{F_S}$ is the interval between two samples, and $\omega[t]$ is the gyroscope reading in the current timestamp. The resulting calculated offset angles are shown in Fig. 3(b), and the lens position is then controlled by the OIS module with these offset angles.

3.3 Conversation from Lens Shift to Pixel Shift

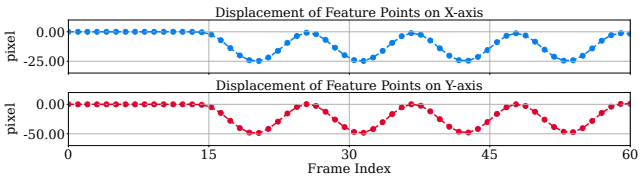
In this study, we impose lens shifts to enable pixel shift sampling from multiple frames of the same scene. Thus, we need to model the relationship between the lens shift and the pixel shift. We simply



(a) Acoustic injection signals: sine tone wave



(b) Calculated offset angles from the built-in 3-axis MEMS gyroscope readings



(c) Displacements of the feature points in the multiple frames

Figure 3: Corresponding results with a stationary Xiaomi 10 Ultra smartphone (fixed on the tripod) under the effects of acoustic signals (start at 0.5 seconds) with frequencies of $18795Hz$ **that played by the built-in speaker**

the mobile phone camera into a pinhole camera model [48] as shown in Fig. 4, and move the lens from *Position 1* to *Position 2* with displacement $\triangle h$ in one axis, such that the imaging of light source A moved from pixel B to pixel B' with displacement $\triangle d$. According to the similar triangles, we can obtain pixel shift $\triangle d$ of light source A as follows:

$$\frac{\triangle h}{\triangle d} = \frac{Z_a}{Z_a + f} \tag{1}$$

where Z_a is the depth of light source A, and f is the focal length of the mobile phone camera. Thus, when capturing multiple frames from the same scene under the effects of lens shift, the depth of the light source remains the same in all images, which means that $\frac{\triangle d}{\triangle h}$ is fixed for each light source in each image pair.

To verify the correctness of the linear model between the lens shift and the pixel shift, we fix the Xiaomi 10 Ultra on the tripod and synchronously capture frames at 30 frames per second of a standard camera calibration checkerboard picture with the same acoustic injection used in Fig. 3(a). We select one corner point in the checkerboard as the feature point, and calculate the pixel information of this feature point in the entire captured frames. The relative pixel displacements of the selected feature point are shown in Fig. 3(c), it is possible to derive the relationship between the offset angles (*i.e.*, lens shift) $\Delta\theta$ and the pixel shift information ΔT as follows:

$$\begin{aligned} \Delta T_x &= a_x \Delta\theta_x, \qquad a_x > 0 \\ \Delta T_y &= a_y \Delta\theta_y, \qquad a_y > 0 \end{aligned} \tag{2}$$

where a_x, a_y are the constant coefficients of the OIS control model.

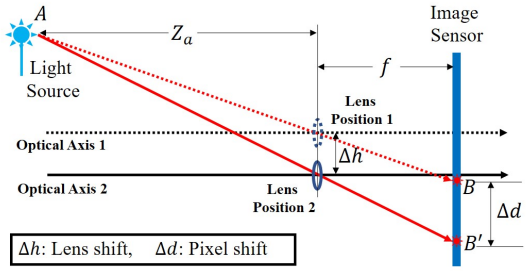


Figure 4: Relationship between lens shift and pixel shift.

To summarize, after we obtain gyroscope readings transmitted to the OIS module from timestamp t_1 to t_2 (take the *Roll*-axis as an example), we can derive lens shift information (along the x-axis) for the OIS module, as follows:

$$\triangle T_x{}_{t_1}^{t_2} = a_x(\triangle \theta_x{}_{t_1}^{t_2}) = a_x(\sum_{t=t_1}^{t_2} \omega_x[t] \triangle t) \tag{3}$$

The same procedure would be followed for lens shifts along the $y-$axis using $\theta_y$axis gyroscope readings. In this manner, we obtain a model by which to convert the gyroscope readings into the pixel shift information of the multiple captures during the lens is moving controlled by the OIS.

4 SUPER RESOLUTION ALGORITHM

In this paper, we sought to utilize the precise and regular OIS-controlled lens motion in the place of the rough handheld movement, and propose a multi-frame based super resolution system that is illustrated in Fig. 5. We will describe the detailed techniques in the following subsections.

4.1 Multiple Frames Acquisition

We first capture a reference frame using the default parameters (*e.g.*, auto-exposure, auto-focus, auto-white balance) with the lens in the zero-shift position (*i.e.*, unperturbed by acoustic injection). The system then utilize the built-in speaker plays the above-mentioned *.wav* file, while simultaneously capturing multiple RAW frames (*.dng* format) via the moving lens. During the capture of these frames, we also record a timestamp of each frame synchronously with the 3-axis gyroscope readings.

Note that the multiple frames could be captured using long exposure compensation under low-light conditions, and the long exposures would increase the likelihood of blurring when capturing images with the moving lens. Thus, we sought to minimize motion blur by adjusting the speed at which lens is moved in accordance with the exposure parameter (*e.g.*, `SENSOR_EXPOSURE_TIME` in Android) by altering the frequency of the acoustic injection signals (see the supplementary video).

4.2 Optimizing Sub-pixel Alignment

Image registration refers to the process of estimating per-pixel shifts between image pairs. In this study, we model the image registration as an optimization problem, and our optimization objective is the pixel shift information of each the same light source when registering the offset frame to the reference frame. Thus, we sought to optimize pixel alignment in each of the offset frames according

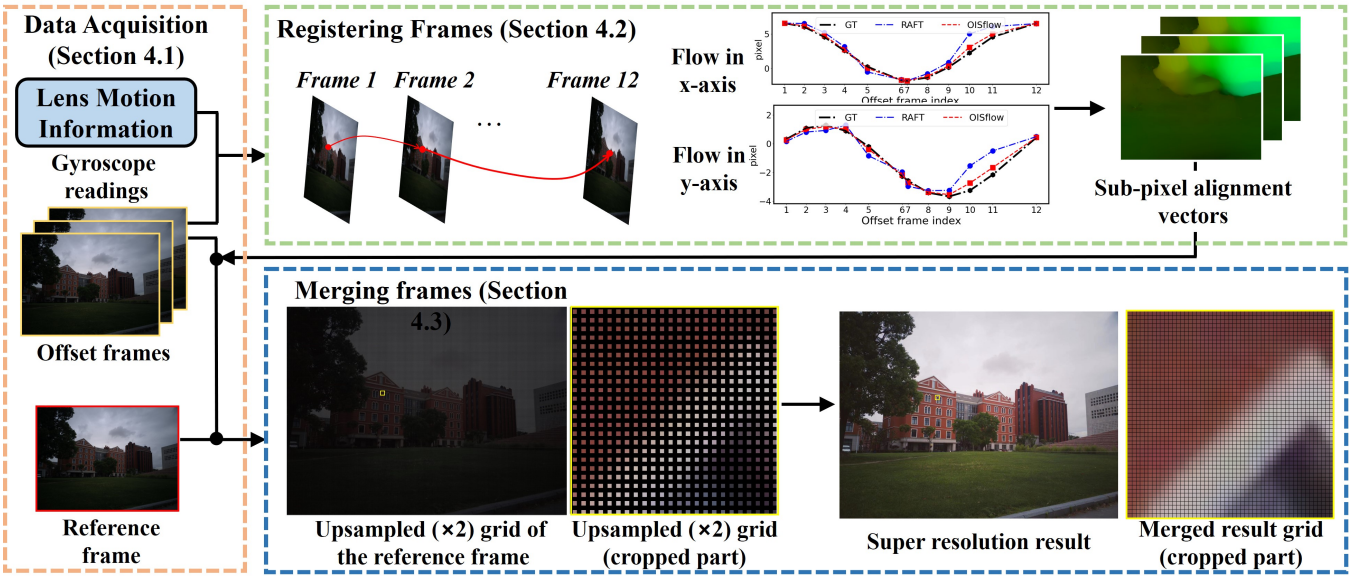


Figure 5: Overview of our proposed OISSR that is a robust super resolution system that leverages the OIS-controlled lens motion

to the reference image using a priori lens shift information in conjunction with a penalty term ($\frac{\Delta d}{\Delta h}$ is fixed) within the optimization function. Generally, optimization involves a defined trade-off between a visual term and a motion term, which imposes priors on the plausibility of lens shift. Note that the visual term is meant to facilitate the alignment of visually similar regions of the image. We also add a smoothing term in the final energy function to enhance the quality of the final alignment results. For each pixel (p) in the image, we minimize the energy function using a visual term, a motion term, and a smoothing term for a given series of captured LR images.

$$E(p) = E_{visual}(p) + \alpha E_{motion}(p) + \beta E_{smooth}(p) \tag{4}$$

Here, the weight α and β balance the relative significance of the three terms, and are set to 1.5(for α) and 3(for β) in all our experiments, and Ω is the image plane.

Visual term. A visual term is defined to facilitate the alignment of visually similar regions in the image, wherein visual alignment is estimated from three iterations of the Lucas-Kanade optical flow method [25]. Here, we denote the offset frame as K_i ($i \in [1, 2, ..k]$) and a reference frame as (K_0). For each pixel p in offset frame K_i, we denote coarse pixel shifting information obtained from the optical flow method as $(u_i(p), v_i(p))$, and our optimized target (the high-precision pixel alignment) is defined as $\hat{u}_i(p), \hat{v}_i(p)$:

$$E_{visual}(p)_i = (\hat{u}_i(p) - u_i(p))^2 + (\hat{v}_i(p) - v_i(p))^2 \tag{5}$$

Thus, the entire visual term can be denoted as follows:

$$E_{visual}(p) = \sum_{i=1}^{k} E_{visual}(p)_i \tag{6}$$

Motion term. As shown in Fig. 4, we obtain the following formula for pixel p in each pair of offset-reference images:

$$\frac{\hat{u}_1(p)}{\triangle\, h_x(1)} = \frac{\hat{u}_2(p)}{\triangle\, h_x(2)} = \cdots = \frac{\hat{u}_k(p)}{\triangle\, h_x(k)}, p \in \Omega \tag{7}$$

$$\frac{\hat{v}_1(p)}{\triangle\, h_y(1)} = \frac{\hat{v}_2(p)}{\triangle\, h_y(2)} = \cdots = \frac{\hat{v}_k(p)}{\triangle\, h_y(k)}, p \in \Omega \tag{8}$$

where $\triangle\, h_x(i)$ and $\triangle\, h_y(i)$ respectively refer to mean lens shift information (*i.e.*, offset angles calculated from the gyroscope readings) along the $x-$ and $y-$axis for the stated time of capturing frame i, where $\hat{u}_i(p)/\hat{v}_i(p)$ indicates $\triangle\, d$ in Fig. 4.

We define the motion term between two offset-reference image pairs as ((K_m, K_0), and (K_n, K_0)). We record the timestamp indicating the start in capturing the mth/nth offset frame as t_m/t_n, and the reference frame timestamp as t_0. With the OIS-controlled lens motion model in Eq. 3, we can result in the following:

$$\begin{aligned} E_{motion}(p)_{m,n} = &\left(\frac{\hat{u}_m(p)}{\sum_{t=t_0}^{t_m} \omega_x[t]\ \triangle\, t} - \frac{\hat{u}_n}{\sum_{t=t_0}^{t_n} \omega_x[t]\ \triangle\, t}\right)^2 + \\ &\left(\frac{\hat{v}_m}{\sum_{t=t_0}^{t_m} \omega_y[t]\ \triangle\, t} - \frac{\hat{v}_n}{\sum_{t=t_0}^{t_n} \omega_y[t]\ \triangle\, t}\right)^2 \end{aligned} \tag{9}$$

Thus, the entire motion term for each pixel p can be noted as follows:

$$E_{motion}(p) = \sum_{m=1}^{k} \sum_{n=1, n\neq m}^{k} E_{motion}(p)_{m,n} \tag{10}$$

Smoothing term. The smoothing term in offset frame K_i is denoted as follows:

$$E_{smooth}(p)_i = \tau(p)(|\nabla \hat{u}_i(p)|_\varepsilon + |\nabla \hat{v}_i(p)|_\varepsilon) \tag{11}$$

where $|\,.\,|_\varepsilon$ refers to the Huber norm with a threshold of ε; $\nabla \hat{u}_i(p)$ is the gradient adaptive weight of the pixel shift which imposes a strong penalty on the featureless area, and $\nabla \hat{u}(p)$ is defined as $\nabla \hat{u}(p) = e^{-\zeta |\nabla I_x(p)|^\eta}$, where $\nabla I_x(p)$ is the image $x-$axis gradient at pixel p, and $\nabla \hat{v}(p)$ is similarly defined as $\nabla \hat{v}(p) = e^{-\zeta |\nabla I_y(p)|^\eta}$. We fixed the parameters $\zeta = 2.8$, $\eta = 0.6$ in all our experiments.

Thus, the whole smooth term can be denoted as follows:

$$E_{smooth}(p) = \sum_{i=1}^{k} E_{smooth}(p)_i \tag{12}$$

4.3 Merging Multiple Registered Frames

Producing the final output image involves processing all registered frames sequentially for every pixel in the output image by evaluating local contributions to the red, green, and blue color channels from separate input frames. In accordance with the merging process described in [50], we utilized kernel regression to estimate the local contribution of each frame to the super-resolution results. For each color channel, the local contribution can be formulated as follows:

$$C(x,y) = \frac{\sum_{i=1}^{k+1} \sum_j c_{i,j} * w_{i,j}}{\sum_{i=1}^{k+1} \sum_j w_{i,j}} \tag{13}$$

where (x, y) are the pixel 2D coordinates of the 2-scale upsampling image grid, the sum $\sum_{i=1}^{k+1}$ obtained over all contributing frames (k offset frames and one reference frame), $\sum_j$ is a sum over samples within a local neighborhood (in our case 3×3) in the low-resolution frames, *i.e.*, the samples whose coordinates are in the $(\lfloor \frac{x}{2} \rfloor \pm 3, \lfloor \frac{y}{2} \rfloor \pm 3)$ after image registration, $c_{i,j}$ denotes the color value of a pixel at given frame i and sample j. For each sample j whose the original coordinate is (x_j, y_j) before image registration and its alignment vector is (u_j, v_j), we utilize a 2D normalized anisotropic Gaussian RBF to calculate the local sample weight $w_{i,j}$,

$$w_{i,j} = e^{-\frac{1}{2} d_j^T \Psi_j^{-1} d_j} \tag{14}$$

where Ψ is the kernel covariance matrix and d_j is the offset vector of sample j to the output upsampling pixel grid, *i.e.*, $d_j = [2(x_j - u_j) - x, 2(y_j - v_j) - y]^T$. To estimate local information pertaining to the strength and direction of gradients, we apply gradient structure tensor analysis in each frame as a kernel covariance matrix:

$$\Psi_j = \begin{bmatrix} \nabla I_{x_j}^2 & \nabla I_{x_j} I_{y_j} \\ \nabla I_{x_j} I_{y_j} & \nabla I_{y_j}^2 \end{bmatrix} \tag{15}$$

where ∇I_{x_j} and ∇I_{y_j} refer to local gradients in the horizontal and vertical directions in the reference image. The image gradients are computed using the finite forward difference method in the luminance channel within a small 3×3 color window.

5 EVALUATION

5.1 Methodology Evaluation

5.1.1 Pixel Alignment Accuracy. We first sought to verify the effectiveness of the proposed image registration algorithm *OISFlow*. We select a number of existing state-of-the-art optical flow methods for comparison. Each image registration method is applied to each image pair to derive pixel shift information. Note however that we are unable to obtain ground truth values pertaining to pixel shift during image capture. Thus, in the absence of ground truth data, we employ a forward-backward consistency scheme [28] to compare the performance of the various image registration algorithms [15, 17, 25, 37, 39]. Specifically, we utilize multiple frames (*e.g.*, f_0, f_1, and f_2) to create sequences (*e.g.*, $f_0 - f_1 - f_2 - f_0$) and then measured the degree of consistency in estimates of optical scene flow between the same pair of frames with the order reversed (*e.g.*, $f_0 - f_1 - f_2$ and $f_2 - f_0$). Ideally, the pixel alignment vectors should have the same magnitude but opposite orientation. And we can add the whole alignment vectors of each image pair in the image sequences, and calculate the magnitude of the final alignment vectors as the metrics score. The results are shown in the Table 1.

Table 1: Comparison of image registration algorithms, where a lower score indicates better performance

ALG	HS [15]	LK [25]	Flow-Net2 [17]	PWC-Net [37]	RAFT [39]	LK w/ lens	OIS-Flow
Score	4.2	4.1	3.9	3.5	3.1	1.5	**1.3**

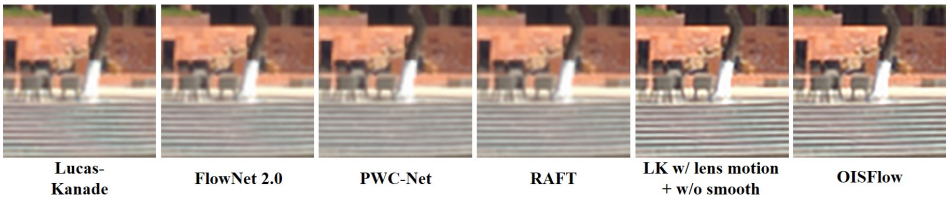
These results indicate that our proposed *OISFlow* achieves the best overall performance. Note that due to the very slight movement of the lens, it is difficult to obtain satisfactory and robust results based solely on visual computation, even when using a SOTA optical flow method (*e.g.*, RAFT) based on deep learning. The proposed image registration algorithm takes into account all of the information pertaining to lens shift in the capture of offset frames. The proposed algorithm also imposes a strong limitation on the optimization framework to ensure high-precision pixel alignment. Also, adding a smoothing term that adaptively weights the neighbors in the extended region can improve the optical flow estimation accuracy and ensure the robustness of our proposed *OISFlow*.

To verify the importance of image registration in multi-frame super-resolution, we merge the registered images (here we select six frames) using a variety of pixel alignment algorithms and then compare the super-resolution results. The RAW images (*.dng* format) with a resolution of 4080×3072 were taken by the Xiaomi 10Ultra camera. We crop a subfigure with 300×300 resolution at the specific position on the reference frame as a HR ground truth. We also crop the subfigures with the same resolution and specific position and downsize them to 150×150 pixels on the entire frames as the multiple LR frames, *i.e.*, a reference frame and six offset frames. As can be seen in Fig. 6(a), due to the fact that the scene was taken particularly far away, many high frequency details cannot be properly represented in the LR frames, such as the presence of colored noise on the steps. This noise results in ignored errors in the alignment vector for each reference- offset image pair when using traditional optical flow methods that only rely on the RGB information. The super-resolution images merged by the multiple registered LR frames with different image registration methods are shown in Fig. 6(c). We find that our proposed *OISFlow* performs better than other optical flow methods with the less misalignment, and the merged result based on our proposed pixel alignment method *OISFlow* makes the high-frequency details become visible. Overall, we determine that the super-resolution images obtained by merging frames with sub-pixel alignment are superior in terms of visual quality.

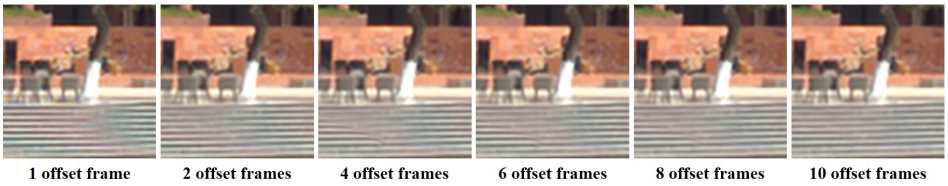
5.1.2 Image Quality and the Number of Merged Frames. From a theoretical perspective, increasing the number of frames should increase the amount of information in an image; however, the proposed image registration scheme is based on an optimization framework that calculates pixel alignment vectors for the entire frames at the same time. Thus, including an excessive number of frames could not lead to any improvement in image registration

(a) Reference image with 4080 × 3072 pixels captured by the Xiaomi 10Ultra's camera

(b) The leftmost is reference LR frame (cropped and downsampled as 150 × 150 pixels). The middle four are the offset frame examples. THe rightmost is HR ground truth frame (cropped as 300 × 300 pixels)



(c) Quality of merged super-resolution images is proportional to pixel alignment accuracy



(d) Influence of the number of merged LR frames on quality of super-resolution images

Figure 6: Methodology evaluation results of our proposed OISSR. We advise the readers to zoom in these images for comparison

performance with a corresponding similar quality of the super-resolution results. Thus, we vary the number of offset frames from 1 to 10. Part of the merged results is shown in Fig. 6(d). We find that the quality of the super-resolution image, in terms of denoising and high-frequency details, hardly improves when more than eight offset frames are used. In the following experiments to compare the overall performance, we use six offset frames and one reference frame to generate the super-resolution image.

5.2 Comparison of Super-resolution Systems

The proposed algorithm is compared with three representative SISR solutions and two representative MFSR solution:

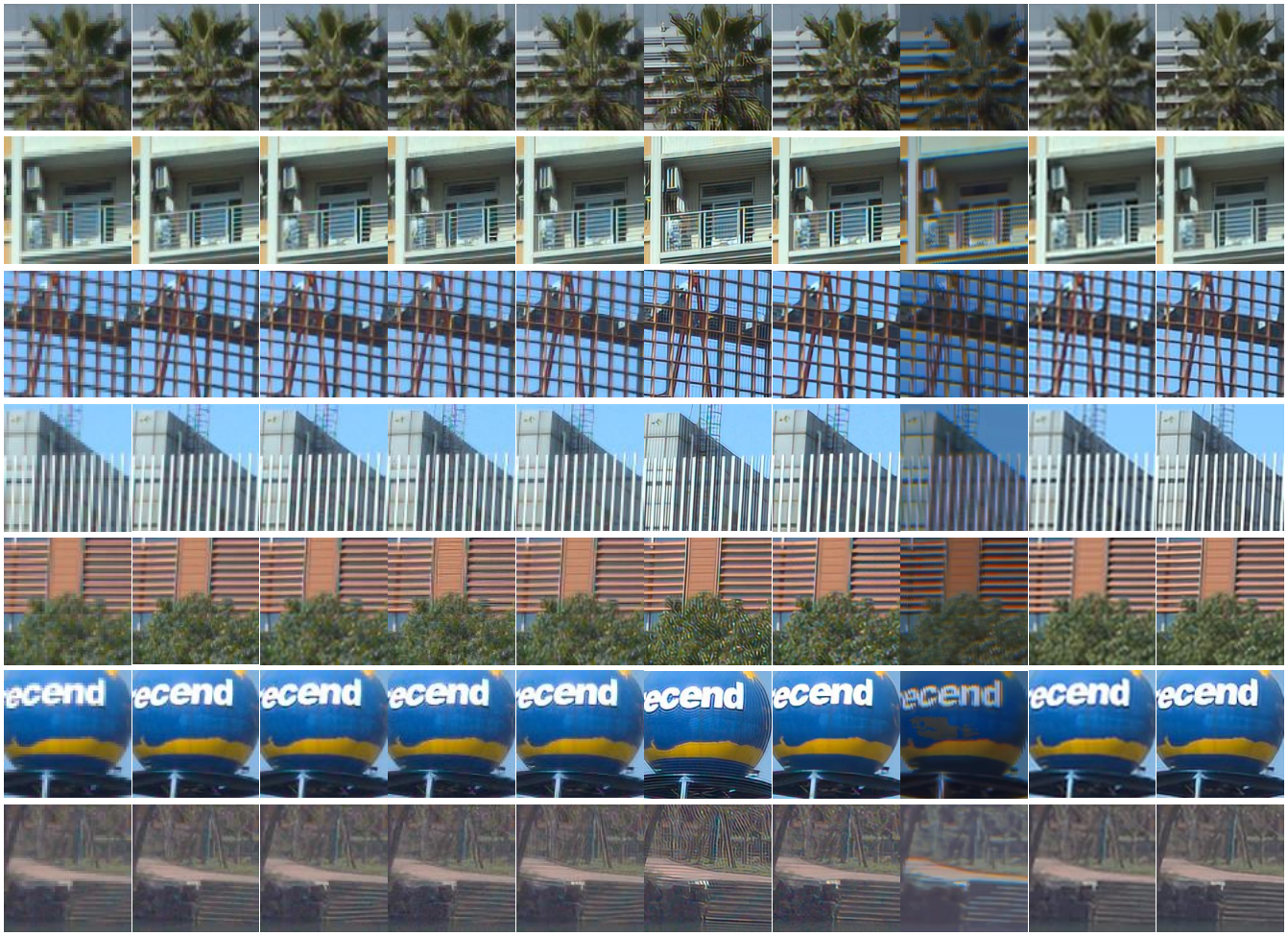
- *Cubic Interpolation* [46], a traditional interpolation method that refers to the bicubic interpolation in the neighborhood of 4 × 4 pixels to enable upsampling.
- *LAPAR* [23] and *SRFlow* [26] are two latest deep learning based SISR technologies used to super-resolution imaging, *i.e.*, image denoising and JPEG image deblocking.
- *MuCAN* [22], BasicVSR [7], and COMISR [24] are deep learning based video super-resolution technology, that utilizes multiple low-resolution frames to generate a high-resolution prediction of the reference frame.
- *DeepRep* is a novel deep-learning based MFSR method for burst images, that takes multiple noisy RAW images as input and generates a denoised and super-resolution RGB image.
- *Handheld multi-frame super-resolution* [50] uses hand tremors to introduce small offsets during the capture of multiple raw frames to facilitate merging as a super-resolution image. We used the implementation of this work in [20].

Table 2: PSNR [47] and SSIM [49] comparisons with selected four super-resolution systems and our proposed OISSR.

	Cubic	LAPAR	SRFlow	MuCAN	BasicVSR
PSNR	30.168	33.011	32.28	33.112	35.387
SSIM	0.911	0.924	0.931	0.944	0.956

	COMISR	DeepRep	Handheld	Ours
PSNR	34.312	32.112	34.187	**36.49**
SSIM	0.941	0.918	0.936	**0.959**

As shown in Fig. 7, cubic interpolation produced the worst overall results, as indicated by the inability to fill in missing pixel information. The LAPAR and SRFlow methods produce images of higher quality; however, these deep-learning based SISR methods sometimes are unable to add the realistic details and prevent the formation artifacts, especially in the low-frequency areas. After comparing the last row of Fig. 7 in detail, we observe that the LAPAR seriously distorts the steps, and the SRFlow generates multiple artifacts in the trunk. The deep learning based video SR methods, such as MuCAN, BasicVSR, and COMISR, also suffer from the unstable performance and sometimes generate the errors and artifacts in the generated SR results. The SR images generated by DeepRep have multiple grainy artifacts and chromatic aberrations. For chromatic aberrations, one explanation is that DeepRep directly uses RAW images as input for the burst SR task and may have domain shift problems, *e.g.*, our data have a different distribution of RAW



(a) LR (cropped) (b) HR GT (c) LAPAR [23] (d) SRFlow [26] (e) MuCAN [22] (f) BasicVSR [7] (g) COMISR [24] (h) DeepRep [4] (i) Handheld [50] (j) Ours

Figure 7: End-to-end comparison of proposed OISSR system and other super-resolution systems from the capture of low-resolution images to the generation of super-resolution images. Please zoom in these images for comparison

values than their training dataset because the RAW images were captured from different CMOS imaging sensors. Handheld MFSR provided pixel information without any artifacts; however, it still generated blurring in high frequency regions and edges due to a lack of accuracy in pixel alignment. Our proposed OISSR system is more robust against the various shooting scenes, and it can generate super-resolution images with more realistic high-frequency details, but with less CMOS imaging noise.

The corresponding quality analysis of our algorithm on the collected datasets (the original images are regarded as ground-truth, and two-time downsampling are taken as input of the systems) is shown in Table 2. We also observe that our proposed system can obtain the best imaging performance among the related SISR, MFSR and video-based SR methods. Note that in handheld shooting mode, MFSR can use information pertaining to camera pose (inferred from gyroscope readings) to improve image registration performance. Nonetheless, the built-in gyroscope's low precision on sensing the slight and irregular hand shake greatly limits the accuracy of pose-related data.

6 CONCLUSION

In this paper, we present a robust multi-frame super-resolution system for OIS-supported smartphone cameras. The proposed OISSR system controls lens motion via acoustic injection to facilitate high-precision sub-pixel alignment by combining lens shift information with pixel shift information obtained using the optical flow method. In experiments, our proposed OISSR can increase image resolution by 4 times the pixel count, and extensive experiments demonstrate the robust performance of OISSR in a variety of scenes.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive comments. This work is supported in part by NSFC (62072306, 61936015) and Program of Shanghai Academic Research Leader (20XD1402100).

REFERENCES

[1] S Abhishek Anand and Nitesh Saxena. 2018. Speechless: Analyzing the threat to speech privacy from smartphone motion sensors. In *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1000–1017.

[2] Simon Baker and Takeo Kanade. 2002. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 9 (2002), 1167–1183.

[3] Benedicte Bascle, Andrew Blake, and Andrew Zisserman. 1996. Motion deblurring and super-resolution from an image sequence. In *European conference on computer vision*. Springer, 571–582.

[4] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. 2021. Deep burst super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9209–9218.

[5] James C Brailean and Aggelos K Katsaggelos. 1995. Simultaneous recursive displacement estimation and restoration of noisy-blurred image sequences. *IEEE Transactions on Image Processing* 4, 9 (1995), 1236–1251.

[6] David Capel and Andrew Zisserman. 2003. Computer vision applied to super resolution. *IEEE Signal Processing Magazine* 20, 3 (2003), 75–86.

[7] Kelvin C.K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. 2021. BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[8] Michael M Chang, A Murat Tekalp, and M Ibrahim Sezan. 1997. Simultaneous motion estimation and segmentation. *IEEE transactions on image processing* 6, 9 (1997), 1326–1333.

[9] Android Developers. 2022. OisSample. https://developer.android.com/reference/android/hardware/camera2/params/OisSample. (2022).

[10] P Erhan Eren, M Ibrahim Sezan, and A Murat Tekalp. 1997. Robust, object-based high-resolution image reconstruction from low-resolution video. *IEEE Transactions on Image Processing* 6, 10 (1997), 1446–1451.

[11] Ming Gao, Feng Lin, Weiye Xu, Muertikepu Nuermaimaiti, Jinsong Han, Wenyao Xu, and Kui Ren. 2020. Deaf-aid: mobile IoT communication exploiting stealthy speaker-to-gyroscope channel. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–13.

[12] Clément Godard, Kevin Matzen, and Matt Uyttendaele. 2018. Deep burst denoising. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 538–554.

[13] Russell C Hardie, Kenneth J Barnard, and Ernest E Armstrong. 1997. Joint MAP registration and high-resolution image estimation using a sequence of undersampled images. *IEEE transactions on Image Processing* 6, 12 (1997), 1621–1633.

[14] Russell C Hardie, Kenneth J Barnard, John G Bognar, Ernest E Armstrong, and Edward A Watson. 1998. High-resolution image reconstruction from a sequence of rotated and translated frames and its application to an infrared imaging system. *Optical Engineering* 37, 1 (1998), 247–260.

[15] Berthold KP Horn and Brian G Schunck. 1981. Determining optical flow. *Artificial intelligence* 17, 1-3 (1981), 185–203.

[16] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. 2015. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5197–5206.

[17] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2462–2470.

[18] Michal Irani and Shmuel Peleg. 1991. Improving resolution by image registration. *CVGIP: Graphical models and image processing* 53, 3 (1991), 231–239.

[19] Camera Jabber. 2021. Which cameras have Pixel Shift? https://camerajabber.com/buyersguides/which-cameras-have-pixel-shift/. (2021).

[20] kunzmi github. 2022. ImageStackAlignator: Implementation of Google's Handheld Multi-Frame Super-Resolution algorithm. https://github.com/kunzmi/ImageStackAlignator. (2022).

[21] Fabrizio La Rosa, Maria Celvisia Virzì, Filippo Bonaccorso, and Marco Branciforte. 2015. Optical Image Stabilization (OIS). *STMicroelectronics. Available online: http://www. st. com/resource/en/white_paper/ois_white_paper. pdf* (2015).

[22] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. 2020. Mucan: Multi-correspondence aggregation network for video super-resolution. In *European Conference on Computer Vision*. Springer, 335–351.

[23] Wenbo Li, Kun Zhou, Lu Qi, Nianjuan Jiang, Jiangbo Lu, and Jiaya Jia. 2020. Lapar: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. *Advances in Neural Information Processing Systems* 33 (2020), 20343–20355.

[24] Yinxiao Li, Pengchong Jin, Feng Yang, Ce Liu, Ming-Hsuan Yang, and Peyman Milanfar. 2021. COMISR: Compression-Informed Video Super-Resolution. In *ICCV*.

[25] Bruce D Lucas, Takeo Kanade, et al. 1981. An iterative image registration technique with an application to stereo vision. Vancouver, British Columbia.

[26] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. 2020. Srflow: Learning the super-resolution space with normalizing flow. In *European conference on computer vision*. Springer, 715–732.

[27] Yan Michalevsky, Dan Boneh, and Gabi Nakibly. 2014. Gyrophone: Recognizing speech from gyroscope signals. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*. 1053–1067.

[28] Philippos Mordohai. 2012. On the Evaluation of Scene Flow Estimation. In *Computer Vision – ECCV 2012. Workshops and Demonstrations*. Springer Berlin Heidelberg, 148–157.

[29] Karl S Ni and Truong Q Nguyen. 2007. Image superresolution using support vector regression. *IEEE Transactions on Image Processing* 16, 6 (2007), 1596–1610.

[30] Hao Pan, Feitong Tan, Yi-Chao Chen, Gaoang Huang, Qingyang Li, Wenhao Li, Taoxue Guang, Lili Qiu, and Xiaoyu Ji. 2022. DoCam: Depth Sensing with an Optical Image Stabilization Supported RGB Camera. In *Proceedings of the 28th Annual International Conference on Mobile Computing and Networking*.

[31] Shmuel Peleg, Danny Keren, and Limor Schweitzer. 1987. Improving image resolution using subpixel motion. *Pattern recognition letters* 5, 3 (1987), 223–226.

[32] PetaPixel. 2015. A Practical Guide to Creating Superresolution Photos with Photoshop. https://petapixel.com/2015/02/21/a-practical-guide-to-creating-superresolution-photos-with-photoshop/. (2015).

[33] Cory Rice. 2018. Pixel-Shift Shootout: Olympus vs. Pentax vs. Sony vs. Panasonic. https://www.bhphotovideo.com/explora/photography/tips-and-solutions/pixel-shift-shootout-olympus-vs-pentax-vs-sony-vs-panasonic. (2018).

[34] Dirk Robinson and Peyman Milanfar. 2004. Fundamental performance limits in image registration. *IEEE Transactions on Image Processing* 13, 9 (2004), 1185–1199.

[35] Dirk Robinson and Peyman Milanfar. 2006. Statistical performance analysis of super-resolution. *IEEE Transactions on Image Processing* 15, 6 (2006), 1413–1428.

[36] Yunmok Son, Hocheol Shin, Dongkwan Kim, Youngseok Park, Juhwan Noh, Kibum Choi, Jungwoo Choi, and Yongdae Kim. 2015. Rocking drones with intentional sound noise on gyroscopic sensors. In *24th {USENIX} Security Symposium ({USENIX} Security 15)*. 881–896.

[37] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. 2018. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8934–8943.

[38] Yu-Wing Tai, Shuaicheng Liu, Michael S Brown, and Stephen Lin. 2010. Super resolution using edge prior and single image detail synthesis. In *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2400–2407.

[39] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*. Springer, 402–419.

[40] William T Thomson. 2018. *Theory of vibration with applications*. CrC Press.

[41] Timothy Trippel, Ofir Weisse, Wenyuan Xu, Peter Honeyman, and Kevin Fu. 2017. WALNUT: Waging doubt on the integrity of MEMS accelerometers with acoustic injection attacks. In *2017 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 3–18.

[42] R Tsai. 1984. Multiframe image restoration and registration. *Advance Computer Visual and Image Processing* 1 (1984), 317–339.

[43] Yazhou Tu, Zhiqiang Lin, Insup Lee, and Xiali Hei. 2018. Injected and delivered: Fabricating implicit control over actuation systems by spoofing inertial sensors. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*. 1545–1562.

[44] Michalis Vrigkas, Christophoros Nikou, and Lisimachos P Kondi. 2013. Accurate image registration for MAP image super-resolution. *Signal Processing: Image Communication* 28, 5 (2013), 494–508.

[45] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. 2013. DeepFlow: Large displacement optical flow with deep matching. In *Proceedings of the IEEE international conference on computer vision*. 1385–1392.

[46] Wikipedia. 2022. Wikipedia, Bicubic Interpolation. https://en.wikipedia.org/wiki/Bicubic_interpolation. (2022).

[47] Wikipedia. 2022. Wikipedia, Peak signal-to-noise ratio (PSNR). https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio. (2022).

[48] Wikipedia. 2022. Wikipedia, Pinhole camera. https://en.wikipedia.org/wiki/Pinhole_camera. (2022).

[49] Wikipedia. 2022. Wikipedia, Structural similarity (SSIM). https://en.wikipedia.org/wiki/Structural_similarity. (2022).

[50] Bartlomiej Wronski, Ignacio Garcia-Dorado, Manfred Ernst, Damien Kelly, Michael Krainin, Chia-Kai Liang, Marc Levoy, and Peyman Milanfar. 2019. Handheld multi-frame super-resolution. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–18.