Bilateral Video Magnification Filter

Shoichiro Takeda[1] Kenta Niwa[1] Mariko Isogawa[1] Shinya Shimizu[1]
Kazuki Okami[2] Yushi Aono[1]
[1]NTT Corporation [2]NTT TechnoCross Corporation
{shoichiro.takeda.us, kenta.niwa.bk, shinya.shimizu.te, yushi.aono.dy}@hco.ntt.co.jp
mariko.isogawa@ieee.org, kazuki.okami.ug@ntt-tx.co.jp

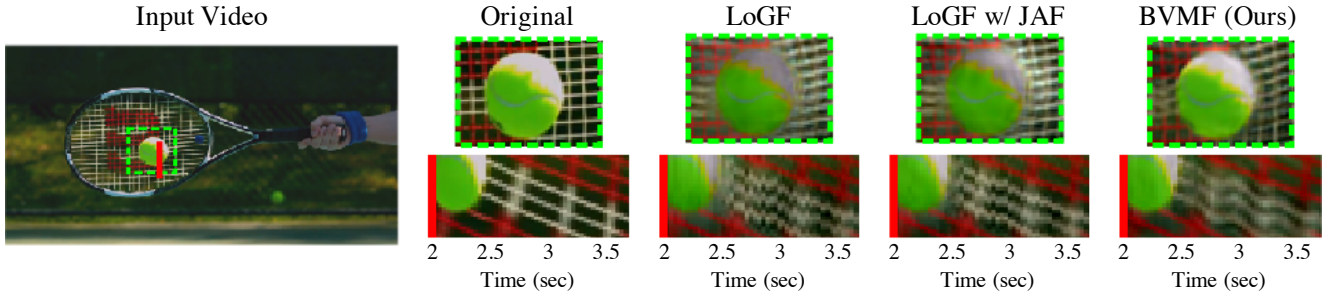

Figure 1. Motion magnification results for subtle string vibrations when a tennis racket hits a ball. The top panels show enlarged image frames in the green dot square in the input video at impact (the ball suddenly stops). The bottom panels show spatiotemporal slices along the red line in the input video. Our proposed BVMF magnified subtle string vibrations with a target frequency of 4 Hz (see the bottom panel) while maintaining the ball shape at impact (see the top panel). In contrast, existing methods, LoGF [30] and LoGF with JAF [22], mis-magnified non-target higher frequency string vibrations and collapsed the ball shape. See supplementary material for the video results.

Abstract

Eulerian video magnification (EVM) has progressed to magnify subtle motions with a target frequency even under the presence of large motions of objects. However, existing EVM methods often fail to produce desirable results in real videos due to (1) mis-extracting subtle motions with a non-target frequency and (2) collapsing results when large de/acceleration motions occur (e.g., objects suddenly start, stop, or change direction). To enhance EVM performance on real videos, this paper proposes a bilateral video magnification filter (BVMF) that offers simple yet robust temporal filtering. BVMF has two kernels; (I) one kernel performs temporal bandpass filtering via a Laplacian of Gaussian whose passband peaks at the target frequency with unity gain and (II) the other kernel excludes large motions outside the magnitude of interest by Gaussian filtering on the intensity of the input signal via the Fourier shift theorem. Thus, BVMF extracts only subtle motions with the target frequency while excluding large motions outside the magnitude of interest, regardless of motion dynamics. In addition, BVMF runs the two kernels in the temporal and intensity domains simultaneously like the bilateral filter does in the spatial and intensity domains. This simplifies implementation and, as a secondary effect, keeps the memory usage low. Experiments conducted on synthetic and real videos show that BVMF outperforms state-of-the-art methods.

1. Introduction

We humans often fail to visually perceive subtle motions in our world: subtle head motions with blood circulation, slight deformation of materials absorbing external forces, or subtle autonomous fluctuations of a flying drone. Such variations are quite useful for helping us deeply understand scene context [1, 3, 4, 25, 28] or anomalous behavior [2, 10], but they are difficult to see with the naked eye.

To magnify such subtle yet important motions in a video, Eulerian video magnification (EVM) methods have been widely researched [6, 10, 13, 17, 21, 22, 25, 26, 28, 30]. EVM methods generally measure local motions in a video as a phase signal within each spatial subband along each orientation at each pixel position [7, 25]. They then perform temporal filtering on the oriented-subband phase signal to extract only subtle motions with a target frequency (e.g., respiratory cycle). However, since subtle motions are easily overwhelmed by large motions of objects, the standard temporal filtering in the early EVM methods [25, 26, 28] often fail to extract subtle motions when objects move largely.

To overcome this issue, temporal filtering in EVM has been continually improved [22, 30]. Specifically, Zhang *et al*. designed the Laplacian of Gaussian filter (LoGF) to perform temporal bandpass filtering while excluding slow large motions, which approximate linearly at short time scales, via its Laplacian property [30]. Furthermore, to exclude

Table 1. Comparisons of temporal filtering results of an EVM with the existing temporal filters [22, 30] and BVMF

Method	Frequency response	Large motions exclusion			Memory usage
		slow	← de/acceleration →	quick	
LoGF [30]	shifted	✓	✗	✗	low
LoGF w/ JAF [22]	shifted	✓	✗	✓	high
BVMF (ours)	non-shifted	✓	✓	✓	low

quick large motions, a combination of LoGF with the jerk-aware filter (JAF) was proposed [22]; JAF excludes only them effectively by assessing jerk-based motion steepness, which represents quick large motions. Thus, this combination extracts subtle motions with a target frequency while excluding slow and quick large motions.

However, LoGF with JAF [22] often fails to produce desirable results in real videos due to the following problems: (1) The passband of LoGF is shifted against the target frequency (see Fig. 2). Thus, LoGF mis-extracts subtle motions with a non-target frequency, or it extracts ones with the target frequency but their magnitude is lower than the original (i.e., the passband gain of LoGF is not unity at the target frequency). (2) LoGF or JAF is specifically designed for excluding only slow or quick large motions, but besides those, there often exist large deceleration or acceleration motions that need to reach the slow or quick large motions in real videos (e.g., objects suddenly start, stop, or change direction). Thus, LoGF with JAF mis-extracts such large de/acceleration motions and collapses results (see Fig. 1). Due to the above problems, EVM in real videos remains a challenging task.

In this paper, we propose a bilateral video magnification filter (BVMF), it is simple yet robust temporal filtering that enhances EVM performance on real videos. Table 1 summarizes comparisons of the existing temporal filters [22, 30] and BVMF. Inspired by the bilateral filter in which two kernels achieve the simple yet robust spatial smoothing process [18, 23], we designed BVMF with two kernels: (I) one kernel performs temporal bandpass filtering via a LoG whose passband peaks at the target frequency with unity gain thanks to new formulations and (II) the other kernel excludes large motions outside the magnitude of interest by Gaussian filtering on the intensity of the phase signal via the Fourier shift theorem (this theorem enables us to measure the magnitude of motions precisely as the intensity of the phase signal). Thus, BVMF extracts only subtle motions with the target frequency while excluding various large motions outside the magnitude of interest regardless of motion dynamics, namely slow, quick, or de/acceleration. In addition, BVMF runs the two kernels in the temporal and intensity domains of the input phase signal simultaneously like the bilateral filter does in the spatial and intensity domains of an image [18, 23]. This simplifies implementation compared to LoGF with JAF that requires multiple input

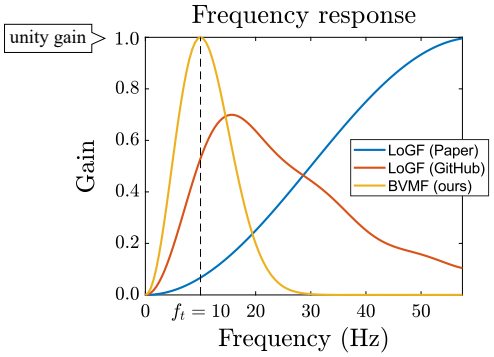

Figure 2. Bandpass frequency response of LoGF [30] (Paper, GitHub)[1] and our BVMF. In this comparison, the target frequency $f_t = 10$ Hz with a sampling rate $f_s = 120$ Hz. The standard deviation σ_{f_t} of LoGF (Paper) is set such that its filter width matches the wavelength of f_t with f_s, referring to the scale selection in blob detection [11, 16]. This is not specifically designed for temporal filtering and thus shifts the passband of LoGF against f_t (see the blue plot). In contrast, BVMF has its passband peak at f_t with unity gain, namely 1.0, thanks to new formulations of Eqs. (5) and (7) (see the yellow plot).

phase signals across spatial subbands to exclude large motions effectively [22]. As a secondary effect, this simple and bilateral implementation keeps the memory usage as low as using just LoGF [30] alone.

The contributions of this paper are as follows. (i) We link the LoG parameters to the passband characteristics by new formulations that strictly set the peak gain of the passband to unity at the target frequency. (ii) We exclude large motions outside the magnitude of interest by Gaussian filtering on the intensity of the phase signal via the Fourier shift theorem. This is simpler than existing approaches [22, 30] while being more robust because it makes no assumptions as to motion dynamics (namely, slow, quick, or de/acceleration). (iii) We, for the first time, introduce the bilateral principle into temporal filtering in EVM, which leads to simpler implementation and, as a secondary effect, lower memory usage than LoGF with JAF [22]. (iv) We conduct extensive experiments and show that our method outperforms the baseline methods, including color magnification results.

[1]The implementation of LoGF differs between the original paper and the official GitHub [30]. The standard deviation σ_{f_t} of LoGF is set as $\sigma_{f_t} = f_s/(4\sqrt{2}f_t)$ (Paper) and $\sigma_{f_t} = f_s/8f_t$ (GitHub). The filter normalization coefficient, Z, is set so that the sum of the absolute value of LoGF coefficients is 1.0, referring to the official GitHub version [30].

2. Related Works

Pioneered by the first Lagrange-based video magnification method [12], many video magnification methods have been researched. Among them, Eulerian video magnification (EVM) methods are the current mainstream [6, 10, 13, 17, 21, 22, 25, 26, 28, 30]. EVM has four stages: (1) constructing a signal representation from each image frame to measure temporal variations in a video, (2) temporal filtering on the signal to extract only subtle variations with a target frequency, (3) magnifying the filtered signal, and (4) collapsing the magnified signal representation to output a magnified image frame. This section summarizes the main research subjects (and also our main focus) in EVM: (1) signal representation and (2) temporal filtering.

2.1. Signal Representation

In motion magnification, several signal representations have been proposed to measure motion variations in a video [10, 15, 17, 25, 26, 28]. For example, the learned representations in a deep convolutional neural network trained by the synthetic motion magnification dataset has attracted much attention [17]. This representation learns a motion-related signal from the synthetic dataset and shows better noise-less results than the existing hand-crafted representations [25, 28]. However, this learned representation often corrupts results due to its strong dataset dependency [21], and how the learned motion-related signal is strictly related to motions in a video is unknown. Therefore, the earlier proposal, the complex steerable pyramid, remains the de facto standard [21, 22, 25, 30]. The complex steerable pyramid consists of a set of oriented-subband analytic signals that enable the phase signal within each spatial subband to be calculated along each orientation at each pixel position. The characteristic of the complex steerable pyramid is that the oriented-subband phase signal is strictly related to local motions within the spatial subband along the orientation via the Fourier shift theorem [7, 19, 25].

Our work uses the complex steerable pyramid for motion magnification because it has a strict relationship between the phase signal and local motions. This enables us to measure the magnitude of motions precisely as the intensity of the phase signal, and thus leads to our robust kernel that excludes large motions effectively (for details, see Section 4).

Note that, in color magnification, most existing works [21, 22, 28, 30] construct the Gaussian pyramid of a color signal as a signal representation to measure color variations in a video. Our work follows this approach and uses the Gaussian pyramid for color magnification.

2.2. Temporal Filtering

The early temporal filtering for EVM used the standard bandpass filters such as the ideal bandpass filter [25, 28] and the differential of Butterworth filter [26]. These filters can produce EVM results when the objects in a video remain static but not when they move largely because they cannot extract only subtle color/motion variations overwhelmed by large motions of objects. To overcome this issue, temporal filtering with image segmentation [6] and depth information [10] have been proposed; both can separate a target image region from large motions. However, this approach requires extra human manipulations [6] or a depth camera [10] to decide the target image region. Thus, Zhang *et al*. [30] and Takeda *et al*. [22] addressed this issue directly by proposing new temporal filters, LoGF and JAF, respectively, as explained in Section 1.

Our work also addresses this issue directly by proposing a new temporal filter called BVMF. Compared to the above temporal filtering [22, 25, 26, 28, 30], BVMF is a simpler yet more robust filtering that enhances EVM performance on real videos in terms of (I) improving frequency selectivity and (II) excluding large motions regardless of motion dynamics (namely, slow, quick, or de/acceleration).

3. Eulerian Video Magnification

Before introducing our proposed BVMF, we explain the general procedure of EVM. We start by defining notations that will be used throughout this paper.

Given image frames $\{I(\mathbf{x},t) \mid t = 0,\dots,T-1\}$ where $\mathbf{x} = [x,y]^\top$ is a pixel position and t is a time frame, a signal representation is constructed from each $I(\mathbf{x},t)$. As the signal representation, most existing EVM methods [6, 10, 21, 22, 25, 26, 28, 30] and our method construct the Gaussian pyramid of a color signal as $\{I_n(\mathbf{x},t) \mid n = 0,\dots,N-1\}$ for color magnification, where $I_n(\mathbf{x},t)$ is the color signal at a pyramid level n. For motion magnification, the complex steerable pyramid is constructed as $\{A_{\omega_n,\theta}(\mathbf{x},t)e^{i\phi_{\omega_n,\theta}(\mathbf{x},t)} \mid n = 0,\dots,N-1, \theta \in \Theta\}$ where $A_{\omega_n,\theta}(\mathbf{x},t)e^{i\phi_{\omega_n,\theta}(\mathbf{x},t)}$ is an oriented-subband analytic signal. This analytic signal consists of an amplitude signal $A_{\omega_n,\theta}(\mathbf{x},t)$ and a phase signal $\phi_{\omega_n,\theta}(\mathbf{x},t)$ within a spatial subband angular frequency ω_n along an orientation θ. Here, we define a generalized signal notation $S_{\nu_n,\theta}(\mathbf{x},t)$; we have the color signal $I_n(\mathbf{x},t)$ where $S = I$, $\nu_n = n$, and $\theta = \emptyset$, or the phase signal $\phi_{\omega_n,\theta}(\mathbf{x},t)$ where $S = \phi$ and $\nu_n = \omega_n$.

After constructing a signal representation, temporal filtering on $S_{\nu_n,\theta}(\mathbf{x},t)$ is performed to extract only subtle color/motion variations with a target frequency f_t. For this temporal filtering, LoGF [30] $h(t;f_t)$ is used to perform temporal bandpass filtering at f_t while excluding slow large motions, and JAF [22] $W_{\nu_n,\theta}(\mathbf{x},t)$ is used to exclude quick large motions. Thus, we obtain a signal $C_{\nu_n,\theta,f_t}(\mathbf{x},t)$ related to subtle color/motion variations with f_t as

$$C_{\nu_n,\theta,f_t}(\mathbf{x},t) = W_{\nu_n,\theta}(\mathbf{x},t)\left(h(t;f_t) * S_{\nu_n,\theta}(\mathbf{x},t)\right), \quad (1)$$

where $*$ indicates convolution over finite range $t \in \mathcal{T}$.

Next, $C_{\nu_n,\theta,f_t}(\mathbf{x},t)$ is amplified with an amplification factor α and added back to the original signal $S_{\nu_n,\theta}(\mathbf{x},t)$ to obtain a magnified signal $\hat{S}_{\nu_n,\theta,f_t}(\mathbf{x},t)$ as

$$\hat{S}_{\nu_n,\theta,f_t}(\mathbf{x},t) = S_{\nu_n,\theta}(\mathbf{x},t) + \alpha C_{\nu_n,\theta,f_t}(\mathbf{x},t). \quad (2)$$

Finally, the magnified Gaussian or complex steerable pyramid $\{\hat{S}_{\nu_n,\theta,f_t}(\mathbf{x},t) \mid n = 0,\dots,N-1, \theta \in \Theta\}$ is collapsed to output a magnified image frame $\hat{I}(\mathbf{x},t)$; the subtle color/motion variations with f_t in a video are magnified.

However, in fact, the current state-of-the-art temporal filtering of Eq. (1) by LoGF with JAF [22] fails in real videos. Specifically, (1) the passband of LoGF is shifted against f_t due to its design based on blob detection (see Fig. 2), and (2) LoGF with JAF misses large de/acceleration motions that occur between slow and quick large motions because each filter is specifically designed for excluding only slow or quick large motions. In addition, Eq. (1) cannot be implemented simply because JAF $W_{\nu_n,\theta}(\mathbf{x},t)$ requires multiple input signals across spatial subbands to exclude quick large motions effectively based on the coarse-to-fine strategy [9, 14] as $W_{\nu_n,\theta}(\mathbf{x},t) := J\left(\{S_{\nu_n,\theta}(\mathbf{x},t) \mid n \in \mathcal{N}\},\beta\right)$ where $J(\cdot,\cdot)$ is a multivariable function to get $W_{\nu_n,\theta}(\mathbf{x},t)$; $\mathcal{N}$ is the across range and β is a hyper-parameter to adjust JAF strength. In the next section, we propose BVMF instead of Eq. (1) to overcome the above issues and thus expand the applicability of EVM in real videos.

4. Proposed Method

To overcome the difficulty of applying EVM to real videos, we propose BVMF instead of Eq. (1). As explained in Section 1, BVMF $\Gamma(s(t),t)$ performs temporal bandpass filtering while excluding large motions outside the magnitude of interest by two kernels running in the temporal and intensity domains of $S_{\nu_n,\theta}(\mathbf{x},t)$ as follows:

$$\begin{aligned} C_{\nu_n,\theta,f_t}(\mathbf{x},t) &= \Gamma(s(t),t) * S_{\nu_n,\theta}(\mathbf{x},t), \\ \Gamma(s(t),t) &:= G(s(t);\sigma_\varepsilon)\, LoG(t;\sigma_{f_t}), \end{aligned} \quad (3)$$

where $LoG(t;\sigma_{f_t})$ is a LoG kernel that performs temporal bandpass filtering at f_t, and $G(s(t);\sigma_\varepsilon)$ is a Gaussian kernel that excludes large motions outside the magnitude of interest ε by taking a local signal intensity change $s(t) = S_{\nu_n,\theta}(\mathbf{x},t) - \sum_{t\in\mathcal{T}} S_{\nu_n,\theta}(\mathbf{x},t)/|\mathcal{T}|$ within the finite range $t \in \mathcal{T}$ of BVMF as input. We explain the details of each kernel in the following subsections.

4.1. LoG Kernel for Temporal Bandpass Filtering

Like LoGF [30], to perform temporal bandpass filtering, we defined $LoG(t;\sigma_{f_t})$ in Eq. (3) as

$$LoG(t;\sigma_{f_t}) := -\frac{t^2-\sigma_{f_t}^2}{Z\sigma_{f_t}^4}\exp\left(-\frac{t^2}{2\sigma_{f_t}^2}\right), \quad (4)$$

where σ_{f_t} is the standard deviation and Z is the normalization coefficient. To extract subtle variations with f_t, LoGF [30] designs σ_{f_t} so that its filter width matches the wavelength of f_t with f_s referring to the scale selection in blob detection [11, 16]. However, its passband is shifted higher than f_t (see Fig. 2), so LoGF cannot work well in terms of temporal bandpass filtering.

We first considered that σ_{f_t} should be designed so as to maximize the passband of $LoG(t;\sigma_{f_t})$ at f_t. Thus, we newly formulated an optimization problem for σ_{f_t} in the Fourier domain of $LoG(t;\sigma_{f_t})$ as

$$f_t = \underset{f,\sigma_{f_t}>0}{\operatorname{argmax}}\ \mathcal{F}[LoG(t;\sigma_{f_t})](f), \quad (5)$$

where $\mathcal{F}[\cdot](f)$ is the 1D Fourier spectrum, namely the bandpass frequency response, of the input at a frequency f. Considering $\mathcal{F}[LoG(t;\sigma_{f_t})](f)$ has a single maximum peak at which the gradient $\nabla_f \mathcal{F}[LoG(t;\sigma_{f_t})](f)$ is 0 if $f,\sigma_{f_t} > 0$, we analytically solved Eq. (5) as

$$\nabla_f \mathcal{F}[LoG(t;\sigma_{f_t})](f_t) = 0 \ \Leftrightarrow\ \sigma_{f_t} = \frac{\sqrt{2}}{2\pi f_t}. \quad (6)$$

For details of this derivation, see the supplementary material. This σ_{f_t} ensures that the passband peak of $LoG(t;\sigma_{f_t})$ is set strictly at f_t.

Moreover, to keep the original magnitude of subtle variations with f_t after applying BVMF, we newly formulated and solved an equation for Z so that the peak gain of the passband of $LoG(t;\sigma_{f_t})$ is unity, namely 1.0, at f_t as follows:

$$\begin{aligned} &|\mathcal{F}[LoG(t;\sigma_{f_t})](f_t)| = 1 \\ \Leftrightarrow\ Z &= \left|\mathcal{F}\left[-\frac{t^2-\sigma_{f_t}^2}{\sigma_{f_t}^4}\exp\left(-\frac{t^2}{2\sigma_{f_t}^2}\right)\right](f_t)\right|, \end{aligned} \quad (7)$$

For details of this derivation, see the supplementary material. From Eqs. (6, 7), BVMF (thanks to this LoG kernel) has its passband peak at f_t with unity gain (see the yellow plot in Fig. 2).

Note that we set the length of the finite range $t \in \mathcal{T}$ of BVMF to $|\mathcal{T}| = 2f_s/f_t$, where f_s is a sampling rate of an input video. For details, see the supplementary material.

4.2. Gaussian Kernel for Excluding Large Motions

To exclude large motions, LoGF [30] and JAF [22] exploit features of motion linearity and jerk-based motion steepness, respectively. However, each higher-level motion feature specializes in excluding only slow or quick large motions, and thus misses large de/acceleration motions that occur between slow large motions and quick ones.

We considered that we should exploit the lowest-level motion feature, the magnitude of motions, to exclude large

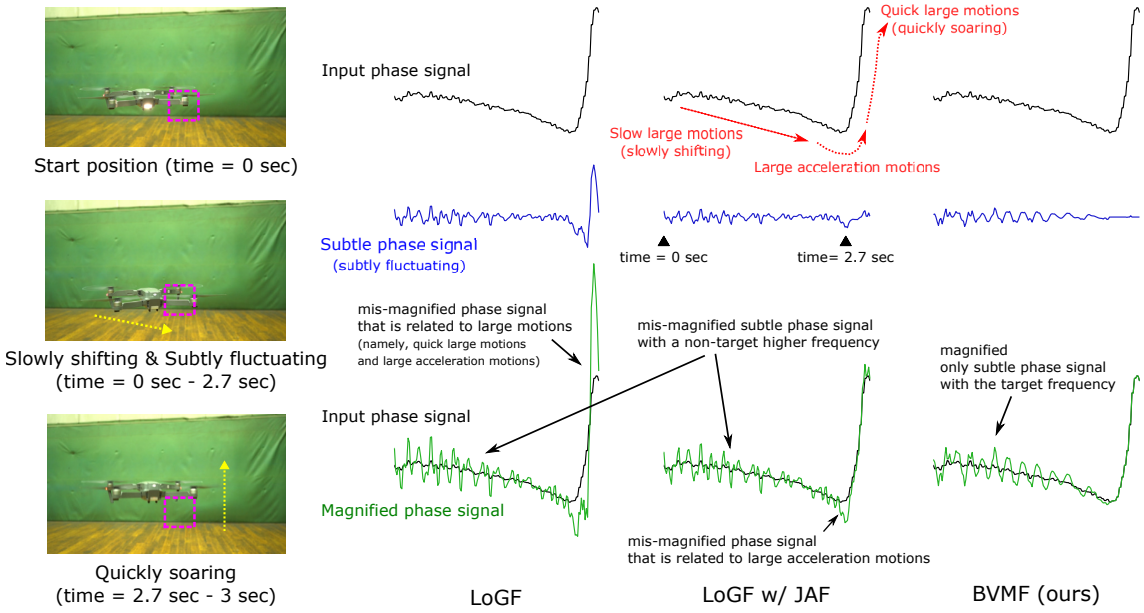

Figure 3. Signal-level comparisons of LoGF [30], LoGF with JAF [22], and our BVMF when magnifying subtle fluctuations of a flying drone. The left panels show drone behaviors along the yellow dot arrows. Given an input phase signal (black) in the purple dot squares in the left panels, each filter extracts a subtle phase signal with the target frequency (blue) that is related to the subtle fluctuations of the drone. The magnified phase signal (green) is the addition of the magnified subtle phase signal to the input. LoGF [30] and LoGF with JAF [22] mis-extract the subtle phase signal with a non-target higher frequency and cannot exclude phase signals related to large (acceleration) motions. In contrast, BVMF extracts and magnifies only the subtle phase signal with the target frequency.

motions regardless of motion dynamics (namely, slow, quick, or de/acceleration). Thus, we defined the Gaussian kernel $G(s(t);\sigma_\varepsilon)$ in Eq. (3) that excludes large motions outside the magnitude of interest ε by Gaussian filtering on the local phase signal intensity change $s(t)$ as follows:

$$G(s(t);\sigma_\varepsilon) := \exp\left(-\frac{s(t)^2}{2\sigma_\varepsilon^2}\right), \quad \sigma_\varepsilon = \frac{\nu_n \varepsilon}{\sqrt{2\ln 2}}. \tag{8}$$

Specifically, this Gaussian kernel is designed to suppress the output of the LoG kernel of Eq. (4) to be less than half when $s(t)$ exceeds the intensity of the phase signal $\nu_n\varepsilon$ (the numerator of σ_ε) because $\nu_n\varepsilon$ represents the motion magnitude ε via the Fourier shift theorem that holds for the complex steerable pyramid [22, 25, 30] as

$$\mathcal{F}[I(x-\varepsilon)](\nu_n) = \mathcal{F}[I(x)](\nu_n)e^{-i\nu_n\varepsilon}.$$

Therefore, BVMF performs temporal bandpass filtering by the LoG kernel while excluding large motions (or suppressing them to be less than half) outside the magnitude of interest ε regardless of motion dynamics by the Gaussian kernel that assesses ε as $\nu_n\varepsilon$ on the intensity domain of the phase signal. In color magnification, we set $\sigma_\varepsilon = \varepsilon/\sqrt{2\ln 2}$ because color variations ε in a video are represented as the color signal intensity changes ε in the Gaussian pyramid.

Note that some slow large motions that approximate linearly are excluded by the LoG kernel thanks to its Laplacian property. Such linearity is often small and may be missed by the Gaussian kernel. Thus, the LoG kernel is essential in excluding large motions precisely.

Figure 3 shows signal-level comparisons of the existing temporal filters [22, 30] and BVMF when magnifying subtle fluctuations of a flying drone. This figure shows that BVMF extracts and magnifies only the subtle phase signal with the target frequency (for details, see Fig. 3).

Comparing Eq. (3) with Eq. (1), BVMF is simpler to implement than LoGF with JAF [22] because BVMF requires only $S_{\nu_n,\theta}(\mathbf{x},t)$ while JAF, $W_{\nu_n,\theta}(\mathbf{x},t)$, requires $\{S_{\nu_n,\theta}(\mathbf{x},t) \mid n \in \mathcal{N}\}$ as input. Thus, as a secondary effect, BVMF keeps the memory usage as low as using just LoGF [30] alone which also requires only $S_{\nu_n,\theta}(\mathbf{x},t)$.

Generalization. As explained before, BVMF needs the LoG kernel to exclude large motions precisely. Meanwhile, other filters (e.g., a very narrow bandpass filter) may be good alternatives to the LoG kernel depending on the situation. As formulated in Eq. (3), LoGF kernel can be replaced with any FIR filter. Thus, we can design the generalized BVMF with an FIR filter (instead of the LoG kernel) that can be applied to different specific situations. For details, see the supplementary material.

5. Experimental Results

Experimental Setup. To evaluate the effectiveness of BVMF, we conducted experiments on real videos and synthetic ones with ground-truth magnification. We assessed

Video [source]	f_t	f_s	α	ε	β [22]
Synthesis 1 [–]	2–18	60	10	10	–
Synthesis 2 [–]	5	60	10	1.2	0.5
Tennis [20]	4	30	40	0.4	0.2
Light bulb [20]	5	30	15	0.035	4
Drone [22]	5	60	15	0.5	0.5
Gun [30]	4	24	15	0.9	0.2

Table 2. Experimental parameters: target frequency f_t, sampling rate f_s, amplification factor α, magnitude of interest ε for excluding large motions (see Section 4.2), and hyper-parameter β used in JAF [22] (see Section 3).

the effectiveness for the synthetic videos quantitatively and that for the real videos qualitatively. Table 2 shows all parameters used in these experiments. We carefully set ε and β used in JAF [22] to exclude large motions. All the video results are shown in the supplementary materials.

For color magnification, we constructed a Gaussian pyramid of the Y color signal from each image frame and amplified only the Y color signal on the third pyramid level. This approach is similar to existing works [22, 28, 30].

For motion magnification, we constructed a half-octave complex steerable pyramid with eight orientations in the Y color channel from each image frame. We performed the phase unwrapping process in advance of temporal filtering. This approach is similar to existing works [22, 25, 30].

5.1. Synthetic Videos for Quantitative Evaluation

We first quantitatively evaluated the effectiveness of our proposed BVMF in synthetic videos as follows.

Frequency Selectivity (Synthesis 1). In this experiment, we evaluated BVMF in terms of frequency selectivity as to whether only subtle motions with f_t can be magnified in the presence of motions with a different frequency. We compared BVMF with the state-of-the-art temporal bandpass filter, namely LoGF (Paper, GitHub)[1] [22, 30]. We applied each temporal filter to a synthetic video where two balls subtly fluctuate along the horizontal axis with d_1 (left ball) and d_2 (right ball). These subtle fluctuations were defined as $d_i = 0.5 \sin\left(2\pi(f_i/f_s)j\right), i = 1, 2$, where $f_{1,2} \in [2, 14]$ Hz ($f_1 \neq f_2$), $f_s = 60$ Hz, and j is the time frame index. We also created a ground-truth where only the subtle fluctuations of the left ball d_1 were magnified as $10 \cdot d_1$. In this case, we set $f_1 = f_t$ and $\alpha = 10$ for each temporal filter for magnifying *only* d_1 as in the ground-truth; we thus excluded the evaluations where $d_1 = d_2$ ($f_1 = f_2$).

Figure 4 shows mean squared error (MSE) over time for each frequency combination of $f_1 = f_t$ and f_2 against the ground-truth. LoGF (Paper) had high MSEs when $f_1 = f_t$ was lower than f_2 due to shifting its passband higher than f_t (see the blue plot in Fig. 2). On the other hand, LoGF (GitHub) had lower MSEs than LoGF (Paper), but its MSEs

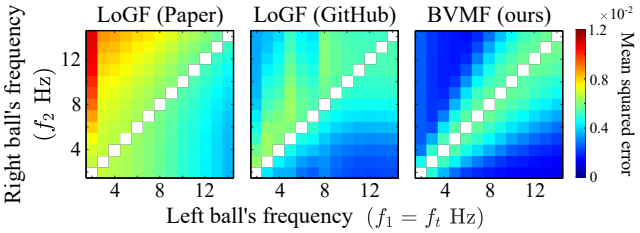

Figure 4. Frequency selectivity of each temporal filter. Mean squared error over time at each frequency combination of $f_1 = f_t$ and f_2 ($f_1 \neq f_2$) against the ground-truth.

were still high when $f_1 = f_t$ was lower than f_2 because its passband was shifted against f_t and the gain at f_t was lower than unity (see the orange plot in Fig. 2). In contrast, BVMF had lower MSEs across almost all frequency combinations than the others. These results suggest that BVMF is superior in terms of frequency selectivity and thus realizes better EVM in real videos.

Note that, to prevent underestimation of LoGF [30] and to perform fair comparisons, we used LoGF (GitHub) as LoGF [30] in the following experiments.

Robustness against Large Motions (Synthesis 2). In this experiment, we evaluated BVMF in terms of robustness against large motions of objects. Specifically, we evaluated whether only subtle motions with f_t can be magnified in the presence of large motions. We compared LoGF [30], LoGF with JAF [22], and BVMF. Moreover, for an ablation study, we additionally compared the LoG kernel of Eq. (4), LoGF with the Gaussian kernel of Eq. (8), and the LoG kernel with JAF. We applied each temporal filter to a synthetic video (Fig. 5 top-left) where top two balls fluctuated subtly with d_1 (left) and d_2 (right) but a bottom ball moved largely with amplitude d_3. The motions of the three balls were defined by $d_i = A_i \sin\left(2\pi(5/f_s)j\right), i = 1, 2, 3$ along the horizontal axis, where $A_1 = 0.5$, $A_2 = 1.0$, $A_3 = 30$ pixels, $f_s = 60$ Hz, and j is the frame index. Due to the larger amplitude A_3 than $A_{1,2}$ with sin function, the bottom ball produces large de/acceleration motions (namely, the bottom ball suddenly changes direction) at the crest or trough of its large sin motions d_3. We also created a ground-truth where only the subtle fluctuations of the top two balls $d_{1,2}$ were amplified as $\hat{d}_{1,2} = 10 \cdot d_{1,2}$. In this case, we set $f_t = 5$ and $\alpha = 10$ for each temporal filter for magnifying only $d_{1,2}$ and not d_3 as in the ground-truth.

Figure 5 shows the magnification results of applying each temporal filter to the synthetic video. Comparing (a) LoGF and (b) the LoG kernel, the latter can magnify subtle fluctuations of the top balls $d_{1,2}$ close to the ground-truth (see the red and green middle panels) because of its strict passband. However, they collapsed the shape of the bottom ball (see the cyan middle and bottom panels). (e) The LoG kernel with JAF [22] mitigates this collapse (see the cyan middle panels) but fails to exclude large de/acceleration

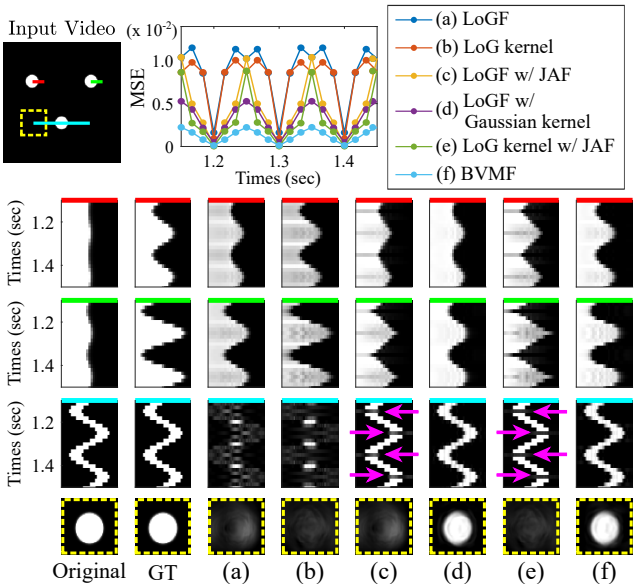

Figure 5. Top-left: a synthetic video where three balls move at different amplitudes along the horizontal red, green, or cyan line. Middle panels: spatiotemporal slices along the above horizontal lines. Bottom panels: enlarged image frames in the yellow dot square on the input video at the suddenly changing motion direction of the bottom ball. Top-Right: MSE against the ground-truth (GT) at each frame. BVMF magnified only subtle fluctuations of the top two balls and shows the lowest MSEs over time.

motions in d_3 at the suddenly changing motion direction (see the purple arrows and the bottom panels). Other temporal filters (c,d) also have either the problem that large de/acceleration motions cannot be excluded (c) or that subtle fluctuations are magnified incorrectly (d). In contrast, (f) BVMF magnifies only $d_{1,2}$ close to the ground-truth while excluding large de/acceleration motions of d_3, which maintains the shape of the bottom ball. As a result, BVMF yields the lowest MSEs over time. Note that we further evaluated BVMF efficacy with different amplitude values, $A_3 \in [5, 100]$, in the supplementary material.

5.2. Real Videos for Qualitative Evaluation

We next qualitatively evaluated the effectiveness of BVMF in real videos. We compared BVMF with two state-of-the-art temporal filters: LoGF [30] and LoGF with JAF [22].

Color Magnification Comparison. Figure 6 shows color magnification results for a blinking filament lamp with a specific frequency. The lamp is suddenly broken by a bullet shot at 2.7 sec. LoGF [30] and LoGF with JAF [22] magnified the subtle blinking but produced luminance saturation (see the purple arrows) and black holes (see the green dot squares) due to the suddenly broken lamp fragments. In contrast, BVMF magnified the subtle blinking more strongly than that by the other filters thanks to its strict passband at f_t with unity gain, and did not produce any noticeable color artifacts.

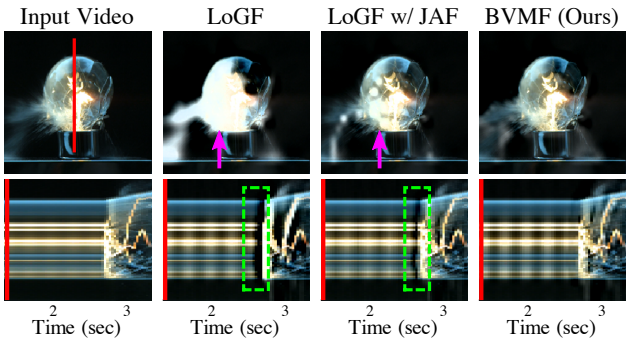

Figure 6. Color magnification results for a blinking filament lamp with a specific frequency. The bottom panels show spatiotemporal slices along the red line in the input video. BVMF clearly magnified the subtle blinking without luminance saturation (see the purple arrows) or black holes (see the green dot squares) due to the suddenly broken lamp fragments by a bullet shot at 2.7 sec.

Motion Magnification Comparison. Figure 1 shows motion magnification results for subtle string vibrations when a tennis racket hits a ball. LoGF [30] and LoGF with JAF [22] mis-magnified non-target higher frequency string vibrations (see the jagged string vibrations in the bottom panels) and collapsed the ball shape due to the suddenly stopping ball by hitting (see the top panels). In contrast, BVMF magnified subtle string vibrations with the target frequency while maintaining the ball shape.

Figure 7 shows motion magnification results for a flying drone that moves slowly or soars quickly while stabilizing itself with subtle fluctuations as shown in Fig. 3. LoGF [30] and LoGF with JAF [22] produced blur on the drone body (see the top panels and the yellow arrows) due to the quick soaring, and also mis-magnified non-target higher frequency fluctuations (see the bottom panels). In contrast, BVMF magnified the subtle fluctuations with the target frequency while maintaining the shape of drone body.

Figure 8 shows the motion magnification results for a gun-shooting video to visualize the impact spread throughout an arm. LoGF [30] and LoGF with JAF [22] collapsed the gun shape (see the top panels) due to the sudden recoil motions, and also mis-magnified subtle vibrations with a non-target higher frequency through the arm (their vibrations are sharper, see the bottom panels). In contrast, BVMF magnified the subtle arm vibrations with the target frequency while maintaining the gun shape.

Memory Usage Comparison. Table 3 shows memory usage comparisons for the drone video with 640×360 pixels and 300 time frames in a naive implementation. LoGF [30] uses the memory to perform only temporal bandpass filtering. LoGF with JAF [22] uses extra memory because

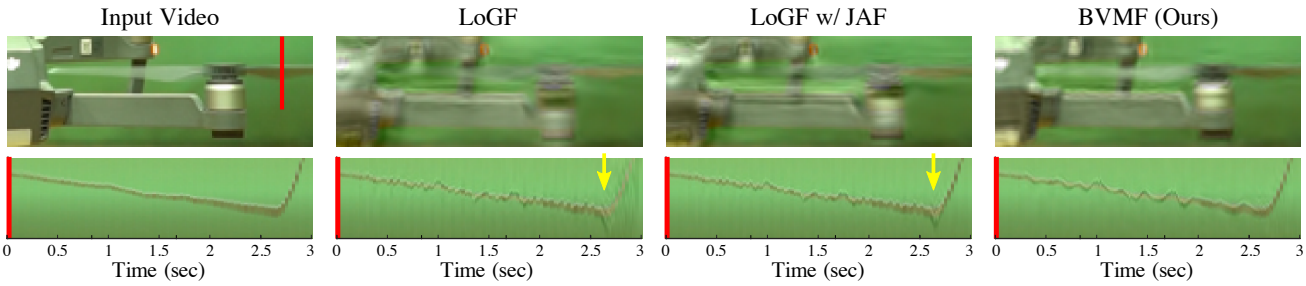

Figure 7. Motion magnification results of Fig. 3. The top panels show enlarged image frames in the purple dot square in Fig. 3 at the quickly soaring. The bottom panels show spatiotemporal slices along the red line in the input video. BVMF magnified subtle fluctuations of the drone with the target frequency (see the bottom panels) while maintaining the drone body (see the top panels).

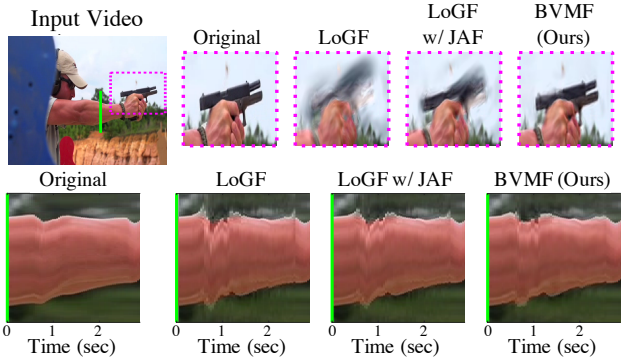

Figure 8. Gun-shooting video: visualizing the impact spread throughout an arm. Top panels show enlarged images in the purple dot square in the input video when a gun fires. The bottom panels show spatiotemporal slices along the green line in the arm. BVMF magnified subtle vibrations through the arm with the target frequency while maintaining the gun shape.

JAF requires multiple input signals across spatial subbands. While BVMF achieves the best EVM result, it keeps the memory usage as low as that of LoGF thanks to its simple and bilateral implementation.

6. Discussions and Limitations

Our proposal, BVMF, will greatly expand the applicability of EVM to real videos by improving frequency selectivity and excluding various large motions. For further progress in EVM, we discuss the following future issues that remain to be solved: (i) In motion magnification, BVMF assumes the use of the complex steerable pyramid, which has a strict relationship between a phase signal and local motions, to design the Gaussian kernel of Eq. (8). If we set σ_ε in the Gaussian kernel heuristically, we would obtain desirable results with other signal representations, e.g., the learned motion representation [17]. However, we should design an advanced BVMF that can be applied to any signal representations in future work. (ii) Similar to JAF [22], the Gaussian kernel of BVMF strongly responds to quick large motions and suppresses all of them even if they hide subtle variations. Thus, magnifying subtle variations hidden in quick large motions still remains a challenging task in EVM research community. (iii) The computational time of BVMF is slower than that of the existing methods [22, 30] due to its bilateral operation. Many fast techniques to the bilateral filter have been proposed [5, 8, 27], and thus we will seek a way to incorporate those techniques into BVMF. (iv) As a general issue with EVM, mis-magnification of subtle noise in a video has been of recent interest and many approaches have been proposed to solve this issue [21, 24, 29]. BVMF could be combined with the above approaches because of its independence from them, but designing a noise-robust BVMF variant remains an interesting future work.

Method	Memory usage (GB)
LoGF [30]	1.833
LoGF w/ JAF [22]	2.110
BVMF (ours)	1.834

Table 3. Memory usage comparisons for the drone video

7. Conclusions

We proposed the bilateral video magnification filter (BVMF) as a simple yet robust temporal filtering that enhances Eulerian video magnification (EVM) performance on real videos. BVMF is superior to existing temporal filtering in terms of (I) its improved frequency selectivity thanks to new formulations that strictly set the peak gain of the passband to unity at the target frequency, (II) its exclusion of various large motions outside the magnitude of interest. In addition, its simple and bilateral implementation keeps the memory usage low. Our experiments on synthetic and real videos demonstrated that BVMF outperformed existing temporal filtering quantitatively and qualitatively. As a result of its effectiveness and simplicity, BVMF will be widely applicable to understand sports scenes (e.g., tennis and shooting), analyze mechanical behavior (e.g., drones), and observe physical phenomena (e.g., blinking lamps).

References

[1] Alborz Amir-Khalili, Jean-Marc Peyrat, Julien Abinahed, Osama Al-Alao, Abdulla Al-Ansari, Ghassan Hamarneh, and Rafeef Abugharbieh. Auto localization and segmentation of occluded vessels in robot-assisted partial nephrectomy. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 407–414, 2014. 1

[2] Nasir A. Aziz and Martijn R. Tannemaat. A microscope for subtle movements in clinical neurology. *Neurology*, 85(10):920–920, 2015. 1

[3] Guha Balakrishnan, Fredo Durand, and John Guttag. Detecting pulse from head motions in video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3430–3437, 2013. 1

[4] Abe Davis*, Katherine L. Bouman*, Justin G. Chen, Michael Rubinstein, Oral Büyüköztürk, Frédo Durand, and William T. Freeman. Visual vibrometry: Estimating material properties from small motions in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(4):732–745, 2017. 1

[5] Frédo Durand and Julie Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. *ACM Transactions on Graphics (TOG)*, 21(3):257–266, July 2002. 8

[6] Mohamed A. Elgharib, Mohamed Hefeeda, Frédo Durand, and William T. Freeman. Video magnification in presence of large motions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4119–4127, 2015. 1, 3

[7] David J Fleet and Allan D Jepson. Computation of component image velocity from local phase information. *International journal of computer vision (IJCV)*, 5(1):77–104, 1990. 1, 3

[8] Ruturaj G. Gavaskar and Kunal N. Chaudhury. Fast adaptive bilateral filtering. *IEEE Transactions on Image Processing*, 28(2):779–790, 2019. 8

[9] Yinlin Hu, Rui Song, and Yunsong Li. Efficient coarse-to-fine patchmatch for large displacement optical flow. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5704–5712, 2016. 4

[10] Julian F.P. Kooij and Jan C. van Gemert. Depth-aware motion magnification. In *European Conference on Computer Vision (ECCV)*, pages 467–482, 2016. 1, 3

[11] Tony Lindeberg. Feature detection with automatic scale selection. *International journal of computer vision (IJCV)*, 30(2):79–116, Nov. 1998. 2, 4

[12] Ce Liu, Antonio Torralba, William T. Freeman, Frédo Durand, and Edward H. Adelson. Motion magnification. *ACM Transactions on Graphics (TOG)*, 24(3):519–526, 2005. 3

[13] L. Liu, L. Lu, J. Luo, J. Zhang, and X. Chen. Enhanced eulerian video magnification. In *International Congress on Image and Signal Processing (ICISP)*, pages 50–54, 2014. 1, 3

[14] Simone Meyer, Oliver Wang, Henning Zimmer, Max Grosse, and Alexander Sorkine-Hornung. Phase-based frame interpolation for video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1410–1418, 2015. 4

[15] Dorkenwald Michael, Buchler Uta, and Ommer Bjorn. Unsupervised magnification of posture deviations across subjects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8256–8266, 2020. 3

[16] Krystian Mikolajczyk and Cordelia Schmid. Indexing based on scale invariant interest points. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 525–531, 2001. 2, 4

[17] Tae-Hyun Oh, Ronnachai Jaroensri, Changil Kim, Mohamed Elgharib, Frédo Durand, William T Freeman, and Wojciech Matusik. Learning-based video motion magnification. In *European Conference on Computer Vision (ECCV)*, pages 663–679, 2018. 1, 3, 8

[18] Sylvain Paris, Pierre Kornprobst, Jack Tumblin, and Frédo Durand. A gentle introduction to bilateral filtering and its applications. In *ACM International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH) Courses*, page 1–es. Association for Computing Machinery, 2007. 2

[19] Eero P Simoncelli and William T Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *IEEE International Conference on Image Processing (ICIP)*, volume 3, pages 444–447, 1995. 3

[20] Storyblocks.com. www.videoblocks.com. 6

[21] Shoichiro Takeda, Yasunori Akagi, Kazuki Okami, Megumi Isogai, and Hideaki Kimata. Video magnification in the wild using fractional anisotropy in temporal distribution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1614–1622, 2019. 1, 3, 8

[22] Shoichiro Takeda, Kazuki Okami, Dan Mikami, Megumi Isogai, and Hideaki Kimata. Jerk-aware video acceleration magnification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1769–1777, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[23] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *IEEE International Conference on Computer Vision (ICCV)*, pages 839–846. IEEE Computer Society, 1998. 2

[24] Manisha Verma and Shanmuganathan Raman. Edge-aware spatial filtering-based motion magnification. In *International Conference on Computer Vision & Image Processing*, pages 117–128, 2018. 8

[25] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T. Freeman. Phase-based video motion processing. *ACM Transactions on Graphics (TOG)*, 32(4):80:1–80:10, 2013. 1, 3, 5, 6

[26] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T. Freeman. Riesz pyramids for fast phase-based video magnification. In *IEEE International Conference on Computational Photography (ICCP)*, pages 1–10, 2014. 1, 3

[27] Ben Weiss. Fast median and bilateral filtering. *ACM Transactions on Graphics (TOG)*, 25(3):519–526, 2006. 8

[28] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics (TOG)*, 31(4), 2012. 1, 3, 6

[29] Xiu Wu, Xuezhi Yang, Jing Jin, and Zhao Yang. PCA-based magnification method for revealing small signals in video. *Signal, Image and Video Processing*, 12:1293–1299, 2018. 8

[30] Yichao Zhang, Silvia L. Pintea, and Jan C. van Gemert. Video acceleration magnification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 502–510, 2017. https://acceleration-magnification.github.io/. 1, 2, 3, 4, 5, 6, 7, 8