Supplementary Report for "Light-in-Flight for a World-in-Motion"

Jongho Lee[1], Ryan J. Suess[2], and Mohit Gupta[1]

[1] University of Wisconsin–Madison, Madison WI 53706, USA
{jlee567, mgupta37}@wisc.edu
[2] Independent Researcher, USA
rjsuess@uwalumni.com

Overview. This document provides derivations, explanations, implementation details, algorithms, and more results supporting the content of the paper titled "Light-in-Flight for a World-in-Motion".

S1 I-ToF Image Formation

In this section, we provide detailed image formation models for indirect time-of-flight (I-ToF) cameras. An I-ToF camera consists of a light source and a sensor. The intensity of the light source is temporally modulated by a periodic modulation function $M(t)$ ($M(t) \geq 0$) with a period T_0. We assume that the modulation function is normalized such that $\frac{1}{T_0}\int_{T_0} M(t)\,dt = 1$. The light emitted by the light source travels to the scene of interest and is reflected back toward the sensor. The radiance of the light incident on a sensor pixel $\mathbf{p}$ is given as

$$R(\mathbf{p};t) = \alpha P_s M\left(t - \frac{2Z}{c}\right) + P_a, \tag{S1}$$

where Z is the scene depth between the camera and the scene point imaged at $\mathbf{p}$, c is the speed of light, and P_s is the average power of the light source. α is a scale factor encapsulating light fall-off, scene albedo, and reflectance properties. P_a is the average power of ambient light (e.g., sunlight) incident on $\mathbf{p}$. Each sensor pixel $\mathbf{p}$ computes the correlation $C(\mathbf{p})$ between $R(\mathbf{p};t)$ and a periodic demodulation function $D(t)$ which has the same period as $M(t)$:

$$C(\mathbf{p}) = \beta \int_T R(\mathbf{p};t)\,D(t)\,dt, \tag{S2}$$

where T is the integration time, and β is a sensor-dependent scale factor encapsulating sensor gain and sensitivity. In order to estimate the unknowns (e.g., scene depth, source strength, and ambient strength), several different $C(\mathbf{p})$ values should be measured by using different pairs of modulation $M(t)$ and demodulation functions $D(t)$. For simplicity and ease of analysis, we focus on sinusoids [7,8,12] for $M(t)$ and $D(t)$. We can define two types of sinusoids for $D(t)$: unipolar ($0 \leq D(t) \leq 1$) and bipolar ($-1 \leq D(t) \leq 1$) demodulation functions.

S1.1 Correlation with Unipolar Demodulation

Let us consider the sinusoidal modulation $M(t)$ and *unipolar* sinusoidal demodulation functions $D(t)$:

$$\underbrace{M(t) = 1 + \cos\left(2\pi f_0 t\right), \qquad D(t) = \frac{1}{2} + \frac{1}{2}\cos\left(2\pi f_0 t\right),}_{\text{Eq. 1 of the main manuscript}} \tag{S3}$$

where the modulation frequency $f_0 = 1/T_0$. Note that $M(t) \geq 0$, $\frac{1}{T_0}\int_{T_0} M(t)\,dt = 1$, and $0 \leq D(t) \leq 1$. In this case, $C\left(\mathbf{p}\right)$ (Eq. S2) is given as

$$C\left(\mathbf{p}\right) = \frac{T}{2}\left(e_s + e_a + \frac{e_s}{2}\cos\left(\frac{4\pi f_0 Z}{c}\right)\right), \tag{S4}$$

where $e_s = \alpha\beta P_s$ and $e_a = \beta P_a$. e_s and e_a are the average number of photoelectrons generated at the sensor per unit time by the light source and the ambient light, respectively. Since Eq. S4 includes three unknowns e_s, e_a, and Z, $N\left(N \geq 3\right)$ number of different $C\left(\mathbf{p}\right)$ values should be measured to decode the unknowns for each pixel $\mathbf{p}$. One way to achieve this is to shift the phase of $D(t)$ N times by different amounts $\psi_n = 2\pi\left(n-1\right)/N$, $n \in \{1, ..., N\}$:

$$\underbrace{C_n\left(\mathbf{p}\right) = \frac{T}{2}\left(e_s + e_a + \frac{e_s}{2}\cos\left(\frac{4\pi f_0 Z}{c} - \psi_n\right)\right)}_{\text{Eq. 2 of the main manuscript}}. \tag{S5}$$

S1.2 Correlation with Bipolar Demodulation

Let us consider the sinusoidal modulation $M(t)$ and *bipolar* sinusoidal demodulation functions $D(t)$ $(-1 \leq D(t) \leq 1)$:

$$M(t) = 1 + \cos\left(2\pi f_0 t\right), \qquad D(t) = \cos\left(2\pi f_0 t\right). \tag{S6}$$

In this case, $C\left(\mathbf{p}\right)$ is

$$C\left(\mathbf{p}\right) = \frac{Te_s}{2}\cos\left(\frac{4\pi f_0 Z}{c}\right). \tag{S7}$$

Note that the unknown e_a is removed as compared to Eq. S4. In order to decode two unknowns e_s and Z for each pixel $\mathbf{p}$, we measure $N\left(N \geq 2\right)$ number of different $C\left(\mathbf{p}\right)$ values by shifting the phase of $D(t)$ N times by different amounts $\psi_n = 2\pi\left(n-1\right)/N$, $n \in \{1, ..., N\}$[3]:

$$C_n\left(\mathbf{p}\right) = \frac{Te_s}{2}\cos\left(\frac{4\pi f_0 Z}{c} - \psi_n\right). \tag{S8}$$

[3] When $N = 2$, however, ψ_n should be changed to $\psi_n = \pi\left(n-1\right)/2$, $n \in \{1, 2\}$.

S1.3 Depth Estimates

When $C_n(\mathbf{p})$ is given by Eq. S5 or S8, the scene depth Z for each pixel $\mathbf{p}$ can be estimated by:

$$\underbrace{\hat{Z}(\mathbf{p}) = \frac{c}{4\pi f_0} \tan^{-1}\left(\frac{\sum_{n=1}^{N} C_n \sin\psi_n}{\sum_{n=1}^{N} C_n \cos\psi_n}\right)}_{\text{Eq. (3) of the main manuscript}}. \tag{S9}$$

We drop $\mathbf{p}$ in $C_n(\mathbf{p})$ for brevity. Eq. S9 holds regardless of whether we use unipolar or bipolar sinusoidal demodulation functions. By computing Eq. S9 for all pixels, we can get a depth map.

S1.4 Intensity Estimates

The intensity I for each pixel $\mathbf{p}$ can be estimated by:

$$\underbrace{\hat{I}(\mathbf{p}) = \frac{1}{N}\sqrt{\left(\sum_{n=1}^{N} C_n \cos\psi_n\right)^2 + \left(\sum_{n=1}^{N} C_n \sin\psi_n\right)^2}}_{\text{Eq. (4) of the main manuscript}} \propto Te_s\,. \tag{S10}$$

The intensity I is proportional to the amount of incident signal photons ($= Te_s$), which is proportional to the scene albedo and inversely proportional to the squared depth. By computing Eq. S10 for all pixels, we can get an intensity image.

S2 SNR of Depth and Intensity Estimates

In this section, we derive the standard deviations and signal-to-noise ratio (SNR) for both depth and intensity estimates. For ease of analysis, we assume the use of the unipolar demodulation function with $N = 4$. The same analysis can be extended to the bipolar demodulation function or other values of N.

S2.1 Depth Standard Deviation

The scene depth when $N = 4$ can be estimated from Eq. S9:

$$\hat{Z} = \frac{c}{4\pi f_0} \tan^{-1}\left(\frac{C_2 - C_4}{C_1 - C_3}\right), \tag{S11}$$

where C_n, $n \in \{1, \ldots, 4\}$ is defined as Eq. S5. The depth standard deviation σ_Z can be obtained using the error propagation rule:

$$\sigma_Z = \sqrt{\sum_{n=1}^{4} \left(\frac{\partial Z}{\partial C_n}\right)^2 \mathrm{Var}\left(C_n\right)} = \sqrt{\sum_{n=1}^{4} \left(\frac{\partial Z}{\partial C_n}\right)^2 C_n}, \tag{S12}$$

where $\mathrm{Var}\left(\cdot\right)$ is a variance operator, and $\mathrm{Var}\left(C_n\right) = C_n$ under a Poisson distribution (photo-electron counts follow a Poisson distribution). From Eq. S11,

$$\left(\frac{\partial Z}{\partial C_1}\right)^2 C_1 = \left(\frac{c}{4\pi f_0}\right)^2 \frac{C_1 \left(C_2 - C_4\right)^2}{\left(\left(C_1 - C_3\right)^2 + \left(C_2 - C_4\right)^2\right)^2}, \tag{S13}$$

$$\left(\frac{\partial Z}{\partial C_2}\right)^2 C_2 = \left(\frac{c}{4\pi f_0}\right)^2 \frac{C_2 \left(C_1 - C_3\right)^2}{\left(\left(C_1 - C_3\right)^2 + \left(C_2 - C_4\right)^2\right)^2}, \tag{S14}$$

$$\left(\frac{\partial Z}{\partial C_3}\right)^2 C_3 = \left(\frac{c}{4\pi f_0}\right)^2 \frac{C_3 \left(C_2 - C_4\right)^2}{\left(\left(C_1 - C_3\right)^2 + \left(C_2 - C_4\right)^2\right)^2}, \tag{S15}$$

and

$$\left(\frac{\partial Z}{\partial C_4}\right)^2 C_4 = \left(\frac{c}{4\pi f_0}\right)^2 \frac{C_4 \left(C_1 - C_3\right)^2}{\left(\left(C_1 - C_3\right)^2 + \left(C_2 - C_4\right)^2\right)^2}. \tag{S16}$$

By applying Eqs. S13 - S16 to Eq. S12, we get

$$\begin{aligned} \sigma_Z &= \sqrt{\sum_{n=1}^{4} \left(\frac{\partial Z}{\partial C_n}\right)^2 C_n} \\ &= \frac{c}{4\pi f_0} \frac{\sqrt{\left(C_1 + C_3\right)\left(C_2 - C_4\right)^2 + \left(C_2 + C_4\right)\left(C_1 - C_3\right)^2}}{\left(C_1 - C_3\right)^2 + \left(C_2 - C_4\right)^2}. \end{aligned} \tag{S17}$$

Since

$$C_1 - C_3 = \frac{Te_s}{2} \cos\left(\frac{4\pi f_0 Z}{c}\right), \tag{S18}$$

$$C_1 + C_3 = T\left(e_s + e_a\right), \tag{S19}$$

$$C_2 - C_4 = \frac{Te_s}{2} \sin\left(\frac{4\pi f_0 Z}{c}\right), \tag{S20}$$

and

$$C_2 + C_4 = T\left(e_s + e_a\right), \tag{S21}$$

the depth standard deviation σ_Z is given by

$$\sigma_Z = \frac{c}{2\pi f_0 \sqrt{T}} \frac{\sqrt{e_s + e_a}}{e_s}. \tag{S22}$$

S2.2 Intensity Standard Deviation

From Eq. S10 and $N = 4$,

$$\hat{I} = \frac{1}{4}\sqrt{\left(C_1 - C_3\right)^2 + \left(C_2 - C_4\right)^2}. \tag{S23}$$

where C_n, $n \in \{1, \ldots, 4\}$ is defined as Eq. S5. The intensity standard deviation σ_I can be obtained using the error propagation rule:

$$\sigma_I = \sqrt{\sum_{n=1}^{4} \left(\frac{\partial I}{\partial C_n}\right)^2 \mathrm{Var}\left(C_n\right)} = \sqrt{\sum_{n=1}^{4} \left(\frac{\partial I}{\partial C_n}\right)^2 C_n}. \tag{S24}$$

From Eq. S23,

$$\left(\frac{\partial I}{\partial C_1}\right)^2 C_1 = \frac{1}{16} \frac{C_1 \left(C_1 - C_3\right)^2}{\left(C_1 - C_3\right)^2 + \left(C_2 - C_4\right)^2}, \tag{S25}$$

$$\left(\frac{\partial I}{\partial C_2}\right)^2 C_2 = \frac{1}{16} \frac{C_2 \left(C_2 - C_4\right)^2}{\left(C_1 - C_3\right)^2 + \left(C_2 - C_4\right)^2}, \tag{S26}$$

$$\left(\frac{\partial I}{\partial C_3}\right)^2 C_3 = \frac{1}{16} \frac{C_3 \left(C_1 - C_3\right)^2}{\left(C_1 - C_3\right)^2 + \left(C_2 - C_4\right)^2}, \tag{S27}$$

and

$$\left(\frac{\partial I}{\partial C_4}\right)^2 C_4 = \frac{1}{16} \frac{C_4 \left(C_2 - C_4\right)^2}{\left(C_1 - C_3\right)^2 + \left(C_2 - C_4\right)^2}. \tag{S28}$$

From Eqs. S24 - S28,

$$\begin{aligned}\sigma_I &= \sqrt{\sum_{n=1}^{4}\left(\frac{\partial I}{\partial C_n}\right)^2 C_n} \\ &= \frac{1}{4}\sqrt{\frac{(C_1+C_3)(C_1-C_3)^2+(C_2+C_4)(C_2-C_4)^2}{(C_1-C_3)^2+(C_2-C_4)^2}}.\end{aligned} \tag{S29}$$

By applying Eqs. S18 - S21 to S29, we get

$$\sigma_I = \frac{\sqrt{T(e_s+e_a)}}{4}. \tag{S30}$$

S2.3 SNR of Depth Estimate

We define the SNR of the depth estimate as the ratio between the true Z and depth standard deviation σ_Z (Eq. S22):

$$\underbrace{\mathrm{SNR}_Z = \frac{Z}{\sigma_Z} = \frac{2\pi f_0\sqrt{T}}{c}\frac{e_s Z}{\sqrt{e_s+e_a}}}_{\text{Eq. 6 of the main manuscript}}. \tag{S31}$$

S2.4 SNR of Intensity Estimate

We define the SNR of the intensity estimate as the ratio between the true I and intensity standard deviation σ_I (Eq. S30). Since $I = \frac{Te_s}{8}$ from Eqs. S18, S20, and S23,

$$\underbrace{\mathrm{SNR}_I = \frac{I}{\sigma_I} = \frac{\sqrt{T}e_s}{2\sqrt{e_s+e_a}}}_{\text{Eq. 7 of the main manuscript}}. \tag{S32}$$

S3 Proof of Observation 1

In this section, we prove Observation 1 using the unipolar demodulation function. This proof can also be extended to the bipolar demodulation function.

Observation 1 Consider two correlation image sets captured successively in time, as shown in Fig. 3 of the main manuscript. If the scene motion is *small* and *linear* over the two correlation image sets, the spatial gradient of the intensity image obtained from each correlation image set (although misaligned due to motion) *preserves* its pixel brightness along the true XY-motion over the two intensity images.

Proof. Consider two neighboring sets (set 1 and set 2) of the correlation images as shown in Fig. S1. Each set consists of N correlation images. Set 1 and 2 are captured with the modulation frequencies f_1 and f_2, respectively.

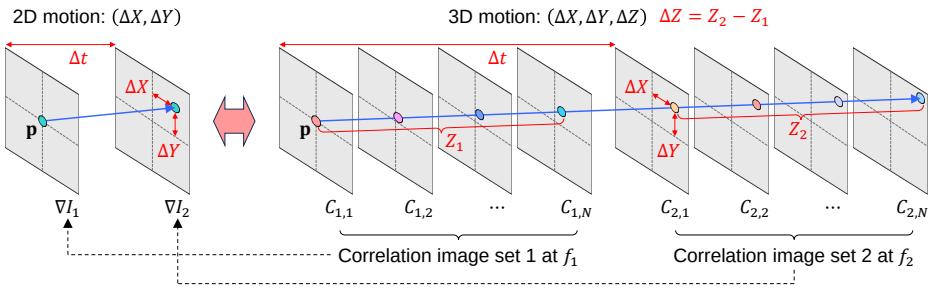

Fig. S1: XY- and Z-motion estimation with brightness-varying correlation images. All correlation images in each set have different pixel values (depicted as distinct colors) along the true XY-motion $(\Delta X, \Delta Y)$, posing a challenge for motion estimation. However, under the small and linear motion, the spatial gradient of the intensity image obtained from each correlation image set maintains its pixel values (represented by the same color) along the motion, facilitating XY-motion estimation. Two depth values (one from each set) can be obtained along the estimated XY-motion, and the Z-motion (ΔZ) can be simply derived from their difference.

Under the small and linear motion, the pixel value of the scene point on the n-th correlation image within the set 1 is given by:

$$\begin{aligned} &C_{1,n}\left(X+(n-1)\,k\Delta X, Y+(n-1)\,k\Delta Y, t+(n-1)\,k\Delta t\right) \\ &= \frac{T\,(e_s+e_a)}{2} + \frac{T e_s}{4}\cos\left(\frac{4\pi f_1 Z}{c} - \psi_n\right)\,(1\le n\le N)\,, \end{aligned} \tag{S33}$$

where (X, Y, t) is the spatio-temporal location of the scene point on the first correlation image within the set 1. $(\Delta X, \Delta Y)$ is the XY-motion between the correlation image sets. k is the fractional coefficient to describe the finer-grained

XY-motion between successive correlation images in each set. We ignore the Z-motion (thus, the variance of e_s along Z) within each set for ease of analysis.

Using the Taylor approximation, the left-hand side of Eq. S33 is approximated as:

$$\begin{aligned} & C_{1,n}\left(X+(n-1)\,k\Delta X, Y+(n-1)\,k\Delta Y, t+(n-1)\,k\Delta t\right) \\ \approx\, & C_{1,n}\left(X,Y,t+(n-1)\,k\Delta t\right)+\frac{\partial C_{1,n}\left(X,Y,t+(n-1)\,k\Delta t\right)}{\partial X}(n-1)\,k\Delta X \\ & +\frac{\partial C_{1,n}\left(X,Y,t+(n-1)\,k\Delta t\right)}{\partial Y}(n-1)\,k\Delta Y. \end{aligned} \tag{S34}$$

Let $C_{1,n}\left(X,Y,t+(n-1)\,k\Delta t\right)=C_{1,n}$ (the correlation values along the same spatial location within the set 1) for brevity. Then, from Eqs. S33 and S34,

$$\begin{aligned} & C_{1,n}+\frac{\partial C_{1,n}}{\partial X}(n-1)\,k\Delta X+\frac{\partial C_{1,n}}{\partial Y}(n-1)\,k\Delta Y \\ \approx\, & \frac{T\left(e_s+e_a\right)}{2}+\frac{Te_s}{4}\cos\left(\frac{4\pi f_1 Z}{c}-\psi_n\right)\left(1\le n\le N\right). \end{aligned} \tag{S35}$$

Let $\frac{\partial C_{1,n}}{\partial X}(n-1)\,k\Delta X+\frac{\partial C_{1,n}}{\partial Y}(n-1)\,k\Delta Y=\Delta C_{1,n}$ for brevity. Then,

$$C_{1,n}\approx\frac{T\left(e_s+e_a\right)}{2}+\frac{Te_s}{4}\cos\left(\frac{4\pi f_1 Z}{c}-\psi_n\right)-\Delta C_{1,n}. \tag{S36}$$

Consider the intensity value obtained from the N correlation values which are not aligned due to small and linear scene motion:

$$I\left(X,Y,t\right)=\frac{1}{N}\left(\left(\sum_{n=1}^{N}C_{1,n}\cos\psi_n\right)^2+\left(\sum_{n=1}^{N}C_{1,n}\sin\psi_n\right)^2\right)^{0.5}, \tag{S37}$$

where $C_{1,n}=C_{1,n}\left(X,Y,t+(n-1)\,k\Delta t\right)$ are the correlation values along the same spatial location (not along the true XY-motion) within the set 1. Using Eq. S36,

$$\begin{aligned} \sum_{n=1}^{N}C_{1,n}\cos\psi_n &= \sum_{n=1}^{N}\left(\frac{T\left(e_s+e_a\right)}{2}+\frac{Te_s}{4}\cos\left(\frac{4\pi f_1 Z}{c}-\psi_n\right)-\Delta C_{1,n}\right)\cos\psi_n \\ &= \frac{NTe_s}{8}\cos\left(\frac{4\pi f_1 Z}{c}\right)-\sum_{n=1}^{N}\Delta C_{1,n}\cos\psi_n. \end{aligned} \tag{S38}$$

Similarly,

$$\sum_{n=1}^{N}C_{1,n}\sin\psi_n=\frac{NTe_s}{8}\sin\left(\frac{4\pi f_1 Z}{c}\right)-\sum_{n=1}^{N}\Delta C_{1,n}\sin\psi_n. \tag{S39}$$

By applying Eqs. S38 and S39 to Eq. S37,

$$\begin{aligned} I\left(X,Y,t\right) &= \frac{1}{N}\left(\left(\frac{NTe_s}{8}\cos\left(\frac{4\pi f_1 Z}{c}\right)-\sum_{n=1}^{N}\Delta C_{1,n}\cos\psi_n\right)^2\right. \\ &\quad \left.+\left(\frac{NTe_s}{8}\sin\left(\frac{4\pi f_1 Z}{c}\right)-\sum_{n=1}^{N}\Delta C_{1,n}\sin\psi_n\right)^2\right)^{0.5} \\ &= \frac{1}{N}\left(\left(\frac{NTe_s}{8}\right)^2-\frac{NTe_s}{4}\cos\left(\frac{4\pi f_1 Z}{c}\right)\sum_{n=1}^{N}\Delta C_{1,n}\cos\psi_n\right. \\ &\quad -\frac{NTe_s}{4}\sin\left(\frac{4\pi f_1 Z}{c}\right)\sum_{n=1}^{N}\Delta C_{1,n}\sin\psi_n \\ &\quad \left.+\left(\sum_{n=1}^{N}\Delta C_{1,n}\cos\psi_n\right)^2+\left(\sum_{n=1}^{N}\Delta C_{1,n}\sin\psi_n\right)^2\right)^{0.5}. \end{aligned} \tag{S40}$$

$$\begin{aligned} I\left(X,Y,t\right) &= \frac{Te_s}{8}\left(1-\frac{16}{NTe_s}\cos\left(\frac{4\pi f_1 Z}{c}\right)\sum_{n=1}^{N}\Delta C_{1,n}\cos\psi_n\right. \\ &\quad -\frac{16}{NTe_s}\sin\left(\frac{4\pi f_1 Z}{c}\right)\sum_{n=1}^{N}\Delta C_{1,n}\sin\psi_n \\ &\quad \left.+\left(\frac{8}{NTe_s}\right)^2\left(\sum_{n=1}^{N}\Delta C_{1,n}\cos\psi_n\right)^2+\left(\frac{8}{NTe_s}\right)^2\left(\sum_{n=1}^{N}\Delta C_{1,n}\sin\psi_n\right)^2\right)^{0.5} \\ &\approx \frac{Te_s}{8}\left(1-\frac{8}{NTe_s}\cos\left(\frac{4\pi f_1 Z}{c}\right)\sum_{n=1}^{N}\Delta C_{1,n}\cos\psi_n\right. \\ &\quad -\frac{8}{NTe_s}\sin\left(\frac{4\pi f_1 Z}{c}\right)\sum_{n=1}^{N}\Delta C_{1,n}\sin\psi_n \\ &\quad \left.+\frac{32}{\left(NTe_s\right)^2}\left(\sum_{n=1}^{N}\Delta C_{1,n}\cos\psi_n\right)^2+\frac{32}{\left(NTe_s\right)^2}\left(\sum_{n=1}^{N}\Delta C_{1,n}\sin\psi_n\right)^2\right). \end{aligned} \tag{S41}$$

Eq. S41 is the intensity value from set 1 under the small and linear scene motion, which is different from $I\left(X,Y,t\right)=\frac{Te_s}{8}$ when there is no motion.

Similarly, we derive the intensity value from set 2 under the small and linear scene motion. The pixel value of the scene point on the n-th correlation image

within the set 2 is given by:

$$\begin{aligned} &C_{2,n}\left(X+\Delta X+(n-1)\,k\Delta X, Y+\Delta Y+(n-1)\,k\Delta Y, t+\Delta t+(n-1)\,k\Delta t\right) \\ &=\frac{T\left(e_s+e_a\right)}{2}+\frac{Te_s}{4}\cos\left(\frac{4\pi f_2\left(Z+\Delta Z\right)}{c}-\psi_n\right)\left(1\le n\le N\right), \end{aligned} \tag{S42}$$

where $(X+\Delta X, Y+\Delta Y, t+\Delta t)$ is the spatio-temporal location of the scene point on the first correlation image within the set 2. Note that f_1 and Z in Eq. S33 are changed to f_2 and $Z+\Delta Z$ in Eq. S42. As we derived Eq. S41 from Eq. S33, we can derive the intensity value for set 2 from Eq. S42 under the small and linear scene motion:

$$\begin{aligned} &I\left(X+\Delta X, Y+\Delta Y, t+\Delta t\right) \\ &\approx\frac{Te_s}{8}\left(1-\frac{8}{NTe_s}\cos\left(\frac{4\pi f_2\left(Z+\Delta Z\right)}{c}\right)\sum_{n=1}^{N}\Delta C_{2,n}\cos\psi_n\right. \\ &\quad-\frac{8}{NTe_s}\sin\left(\frac{4\pi f_2\left(Z+\Delta Z\right)}{c}\right)\sum_{n=1}^{N}\Delta C_{2,n}\sin\psi_n \\ &\quad\left.+\frac{32}{\left(NTe_s\right)^2}\left(\sum_{n=1}^{N}\Delta C_{2,n}\cos\psi_n\right)^2+\frac{32}{\left(NTe_s\right)^2}\left(\sum_{n=1}^{N}\Delta C_{2,n}\sin\psi_n\right)^2\right), \end{aligned} \tag{S43}$$

where $C_{2,n}=C_{2,n}\left(X+\Delta X, Y+\Delta Y, t+\Delta t+(n-1)\,k\Delta t\right)$. Again, Eq. S43 is different from $\frac{Te_s}{8}$, which is the value when there is no motion.

For the small and linear motion, $\frac{\partial C_{1,n}}{\partial X}=\frac{\partial C_{2,n}}{\partial X}$, $\frac{\partial C_{1,n}}{\partial Y}=\frac{\partial C_{2,n}}{\partial Y}$, and $\Delta C_{1,n}=\Delta C_{2,n}=\Delta C_n$. As a result, we can derive the following from Eqs. S41 and S43:

$$\begin{aligned} &I\left(X+\Delta X, Y+\Delta Y, t+\Delta t\right)-I\left(X,Y,t\right) \\ &\approx\frac{1}{N}\sum_{n=1}^{N}\Delta C_n\cos\psi_n\left(\cos\left(\frac{4\pi f_1 Z}{c}\right)-\cos\left(\frac{4\pi f_2\left(Z+\Delta Z\right)}{c}\right)\right) \\ &\quad+\frac{1}{N}\sum_{n=1}^{N}\Delta C_n\sin\psi_n\left(\sin\left(\frac{4\pi f_1 Z}{c}\right)-\sin\left(\frac{4\pi f_2\left(Z+\Delta Z\right)}{c}\right)\right). \end{aligned} \tag{S44}$$

As indicated in Eq. S44, the intensity value obtained from the N correlation values at the same spatial location does *not* preserve its brightness even along the true XY-motion. However, since Z, ΔZ, and ΔC_n are similar over the local neighborhood, $I\left(X+\Delta X, Y+\Delta Y, t+\Delta t\right)-I\left(X,Y,t\right)$ value is also similar over the local neighborhood. Therefore, we can derive the following equations:

$$\frac{\partial}{\partial X}I\left(X+\Delta X, Y+\Delta Y, t+\Delta t\right)-\frac{\partial}{\partial X}I\left(X,Y,t\right)=0, \tag{S45}$$

and

$$\frac{\partial}{\partial Y}I\left(X+\Delta X, Y+\Delta Y, t+\Delta t\right)-\frac{\partial}{\partial Y}I\left(X,Y,t\right)=0. \tag{S46}$$

The partial derivatives of I along X- and Y-directions do not change along the XY-moton. Eqs. S45 and S46 can be combined to:

$$\underbrace{\frac{\partial|\nabla I|}{\partial X}\Delta X + \frac{\partial|\nabla I|}{\partial Y}\Delta Y + \frac{\partial|\nabla I|}{\partial t}\Delta t = 0}_{\textbf{Eq. 8 of the main manuscript}}, \tag{S47}$$

where $\nabla = \left(\frac{\partial}{\partial X}, \frac{\partial}{\partial Y}\right)^T$ denotes the spatial gradient.

S4 Comparisons with Doppler ToF Imaging

In this section, we compare our approach with Doppler ToF imaging [4] in terms of the standard deviation of the measured axial velocity and the number of measurements. Our approach estimates the axial motion (Z-motion) from the measured depth difference along the estimated XY-motion. In contrast, Doppler ToF imaging estimates the axial motion by measuring the Doppler frequency shift, which is proportional to the axial velocity. In Doppler ToF imaging, two correlation measurements are obtained using two bipolar demodulation functions: one with the same frequency as the modulation frequency of the light source (homodyne measurement) and the other with the orthogonal frequency (heterodyne measurement). The axial velocity is estimated from the ratio between these two measurements. Refer to [4] for more details.

For ease of noise analysis and fair comparisons, we impose several constraints. First, we assume unipolar demodulation functions for both our approach and Doppler ToF imaging. Since the correlation values are always positive with unipolar demodulation functions, their variances are simply their mean values. Second, we assume four correlation measurements (instead of two measurements as in [4]) to estimate axial velocity in Doppler ToF imaging. With unipolar demodulation functions, the correlation values include DC offset, which does not contain useful information for velocity estimation. Thus, we need additional measurements to remove this DC offset. For the same reason, we assume four measurements to estimate a depth value in our approach. Additionally, we assume that all correlation measurements in both cases are perfectly aligned after the lateral motion estimation to focus on the axial motion estimation.

S4.1 New Image Formation for Doppler ToF Imaging

We derive the new image formation model for Doppler ToF imaging under the aforementioned constraints. Refer to [4] for the original image formation model. The received signal at the sensor is given by Eq. S1 (we also include β from Eq. S2):

$$\begin{aligned} R\left(t\right) &= \alpha\beta P_s M\left(t - \frac{2Z}{c}\right) + \beta P_a \\ &= e_s M(t - \frac{2Z}{c}) + e_a, \end{aligned} \tag{S48}$$

where $e_s = \alpha\beta P_s$ and $e_a = \beta P_a$ are the average number of photo-electrons generated at the sensor per unit time by the light source and the ambient light, respectively. With a sinusoidal modulation function $M(t) = 1 + \cos\left(2\pi f_0 t\right)$,

$$R\left(t\right) = e_s + e_s \cos\left(2\pi\left(f_0 + \Delta f\right)t - \phi\right) + e_a, \tag{S49}$$

where $\Delta f = \frac{2v}{c} f_0$ is the Doppler frequency shift, v is the axial velocity, and $\phi = \frac{4\pi f_0 Z}{c}$ is the phase shift due to the propagation distance.

In order to estimate the Doppler frequency shift Δf from which we can estimate the axial velocity v, we compute the correlation values between the received signal and the sensor's demodulation functions with different frequencies and phase shifts, which are defined as:

$$D_{A,n}(t) = \frac{1}{2} + \frac{1}{2}\cos\left(2\pi f_0 t - (n-1)\pi\right), \quad n \in (1,2) \tag{S50}$$

and

$$D_{B,n}(t) = \frac{1}{2} + \frac{1}{2}\cos\left(2\pi f_B t - (n-1)\pi\right), \quad n \in (1,2). \tag{S51}$$

$D_{A,n}$ and $D_{B,n}$ are the demodulation functions with the frequencies f_0 and f_B, respectively. f_0 is the same frequency as the modulation frequency of the light source. $f_B = f_0 + \frac{1}{T}$ as defined in [4]. The correlation values between the received signal and these demodulation functions are given as:

$$\begin{aligned} C_{A,n} &= \int_T R(t)\, D_{A,n}(t)\, dt \\ &= \frac{T}{2}(e_s + e_a) + \frac{e_s}{8\pi\Delta f}\left(\sin\left(2\pi\Delta f T - \phi + (n-1)\pi\right) - \sin\left(-\phi + (n-1)\pi\right)\right) \end{aligned} \tag{S52}$$

and

$$\begin{aligned} C_{B,n} &= \int_T R(t)\, D_{B,n}(t)\, dt \\ &= \frac{T}{2}(e_s + e_a) + \frac{e_s}{8\pi\left(f_0 + \Delta f - f_B\right)}\left(\sin\left(2\pi\left(f_0 + \Delta f - f_B\right)T - \phi + (n-1)\pi\right)\right. \\ &\quad \left. - \sin\left(-\phi + (n-1)\pi\right)\right). \end{aligned} \tag{S53}$$

From Eqs. S52 and S53,

$$C_{A,1} = \frac{T}{2}(e_s + e_a) + \frac{e_s}{8\pi\Delta f}\left(\sin\left(2\pi\Delta f T - \phi\right) + \sin\phi\right), \tag{S54}$$

$$C_{A,2} = \frac{T}{2}(e_s + e_a) + \frac{e_s}{8\pi\Delta f}\left(-\sin\left(2\pi\Delta f T - \phi\right) - \sin\phi\right), \tag{S55}$$

$$C_{B,1} = \frac{T}{2}(e_s + e_a) + \frac{e_s}{8\pi\left(f_0 + \Delta f - f_B\right)}\left(\sin\left(2\pi\left(f_0 + \Delta f - f_B\right)T - \phi\right) + \sin\phi\right), \tag{S56}$$

and

$$C_{B,2} = \frac{T}{2}(e_s + e_a) + \frac{e_s}{8\pi\left(f_0 + \Delta f - f_B\right)}\left(-\sin\left(2\pi\left(f_0 + \Delta f - f_B\right)T - \phi\right) - \sin\phi\right). \tag{S57}$$

In order to extract the velocity information, we take the following ratio from these measurements:

$$\frac{C_{B,1}-C_{B,2}}{C_{A,1}-C_{A,2}}=\frac{\Delta f}{f_0+\Delta f-f_B}. \tag{S58}$$

The right-hand side of Eq. S58 is the same as Eq. (14) in [4]. From Eq. S58 and $\Delta f=\frac{2v}{c}f_0$, the axial velocity $v_{\Delta f}$ measured by Doppler ToF imaging is given as:

$$v_{\Delta f}=\frac{c}{2f_0}\left(f_0-f_B\right)\frac{C_{B,1}-C_{B,2}}{C_{A,1}-C_{A,2}-C_{B,1}+C_{B,2}}. \tag{S59}$$

Using $f_0-f_B=-\frac{1}{T}$,

$$v_{\Delta f}=\frac{c}{2f_0T}\frac{C_{B,1}-C_{B,2}}{C_{B,1}-C_{B,2}-C_{A,1}+C_{A,2}}. \tag{S60}$$

S4.2 Velocity Standard Deviation by Doppler ToF Imaging

When the axial velocity is estimated by Doppler ToF imaging, the standard deviation of the velocity estimates is given by:

$$\underbrace{\sigma_{v_{\Delta f}}=\frac{2\pi c}{f_0\sqrt{T}}\frac{\sqrt{e_s+e_a}}{e_s}\frac{\sqrt{\frac{1}{\left(\Delta f-\frac{1}{T}\right)^2}+\frac{1}{\Delta f^2}}}{\left(\frac{1}{\Delta f-\frac{1}{T}}-\frac{1}{\Delta f}\right)^2}\frac{1}{\left|\sin\left(2\pi\Delta fT-\phi\right)+\sin\phi\right|}}_{\text{Eq. 10 of the main manuscript}}, \tag{S61}$$

where $\phi=\frac{4\pi f_0Z}{c}$, and Δf is the Doppler frequency shift defined as $\Delta f=\frac{2v}{c}f_0$.

Proof. The velocity standard deviation $\sigma_{v_{\Delta f}}$ by Doppler ToF can be obtained using the error propagation rule:

$$\begin{aligned}\sigma_{v_{\Delta f}}&=\sqrt{\sum_{n=1}^{2}\left(\frac{\partial v}{\partial C_{A,n}}\right)^2\mathrm{Var}\left(C_{A,n}\right)+\sum_{n=1}^{2}\left(\frac{\partial v}{\partial C_{B,n}}\right)^2\mathrm{Var}\left(C_{B,n}\right)}\\&=\sqrt{\sum_{n=1}^{2}\left(\frac{\partial v}{\partial C_{A,n}}\right)^2C_{A,n}+\sum_{n=1}^{2}\left(\frac{\partial v}{\partial C_{B,n}}\right)^2C_{B,n}}.\end{aligned} \tag{S62}$$

From Eq. S60,

$$\frac{\partial v}{\partial C_{A,1}}=\frac{c}{2f_0T}\frac{C_{B,1}-C_{B,2}}{\left(C_{B,1}-C_{B,2}-C_{A,1}+C_{A,2}\right)^2}, \tag{S63}$$

$$\frac{\partial v}{\partial C_{A,2}}=\frac{c}{2f_0T}\frac{-\left(C_{B,1}-C_{B,2}\right)}{\left(C_{B,1}-C_{B,2}-C_{A,1}+C_{A,2}\right)^2}, \tag{S64}$$

$$\frac{\partial v}{\partial C_{B,1}} = \frac{c}{2f_0T}\frac{-\left(C_{A,1}-C_{A,2}\right)}{\left(C_{B,1}-C_{B,2}-C_{A,1}+C_{A,2}\right)^2}, \tag{S65}$$

and

$$\frac{\partial v}{\partial C_{B,2}} = \frac{c}{2f_0T}\frac{C_{A,1}-C_{A,2}}{\left(C_{B,1}-C_{B,2}-C_{A,1}+C_{A,2}\right)^2}. \tag{S66}$$

By applying Eqs. S63 - S66 to Eq. S62,

$$\sigma_{v_{\Delta f}} = \frac{c}{2f_0T}\frac{\sqrt{\left(C_{B,1}-C_{B,2}\right)^2\left(C_{A,1}+C_{A,2}\right)+\left(C_{A,1}-C_{A,2}\right)^2\left(C_{B,1}+C_{B,2}\right)}}{\left(C_{B,1}-C_{B,2}-C_{A,1}+C_{A,2}\right)^2}. \tag{S67}$$

Since

$$C_{A,1}-C_{A,2} = \frac{e_s}{4\pi\Delta f}\left(\sin\left(2\pi\Delta fT-\phi\right)+\sin\phi\right), \tag{S68}$$

$$C_{A,1}+C_{A,2} = T\left(e_s+e_a\right), \tag{S69}$$

$$C_{B,1}-C_{B,2} = \frac{e_s}{4\pi\left(f_0+\Delta f-f_B\right)}\left(\sin\left(2\pi\left(f_0+\Delta f-f_B\right)T-\phi\right)+\sin\phi\right), \tag{S70}$$

and

$$C_{B,1}+C_{B,2} = T\left(e_s+e_a\right), \tag{S71}$$

The velocity standard deviation by Doppler ToF is given as:

$$\sigma_{v_{\Delta f}} = \frac{2\pi c}{f_0\sqrt{T}}\frac{\sqrt{e_s+e_a}}{e_s}\frac{\sqrt{\frac{1}{\left(\Delta f-\frac{1}{T}\right)^2}+\frac{1}{\Delta f^2}}}{\left(\frac{1}{\Delta f-\frac{1}{T}}-\frac{1}{\Delta f}\right)^2}\frac{1}{\left|\sin\left(2\pi\Delta fT-\phi\right)+\sin\phi\right|}. \tag{S72}$$

S4.3 Velocity Standard Deviation by Depth Difference

When the axial velocity (or Z-motion) is estimated from depth difference, the standard deviation of the velocity estimates is given by:

$$\underbrace{\sigma_{v_{\Delta Z}} = \frac{c}{\sqrt{2}\pi f_0\sqrt{T}\Delta t}\frac{\sqrt{e_s+e_a}}{e_s}}_{\textbf{Eq. 9 of the main manuscript}}. \tag{S73}$$

Proof. The axial velocity $v_{\Delta Z}$ measured by depth difference is given as:

$$v_{\Delta Z} = \frac{Z_2-Z_1}{\Delta t}, \tag{S74}$$

where Z_1 and Z_2 are two depth values along the scene motion, and Δt is the time while the scene point moves from Z_1 to Z_2.

$$\begin{aligned} \mathrm{Var}\left(v_{\Delta Z}\right) &= \left(\frac{1}{\Delta t}\right)^2 \mathrm{Var}\left(Z_2 - Z_1\right) \\ &= \left(\frac{1}{\Delta t}\right)^2 \left(\mathrm{Var}\left(Z_2\right) + \mathrm{Var}\left(Z_1\right)\right) \end{aligned} \tag{S75}$$

From Eq. S22,

$$\mathrm{Var}\left(Z_2\right) = \mathrm{Var}\left(Z_1\right) = \sigma_Z^2 = \left(\frac{c}{2\pi f_0 \sqrt{T}} \frac{\sqrt{e_s + e_a}}{e_s}\right)^2. \tag{S76}$$

By applying Eq. S76 to Eq. S75,

$$\begin{aligned} \mathrm{Var}\left(v_{\Delta Z}\right) &= \left(\frac{1}{\Delta t}\right)^2 \left(\left(\frac{c}{2\pi f_0 \sqrt{T}} \frac{\sqrt{e_s + e_a}}{e_s}\right)^2 + \left(\frac{c}{2\pi f_0 \sqrt{T}} \frac{\sqrt{e_s + e_a}}{e_s}\right)^2\right) \\ &= \left(\frac{1}{\Delta t}\right)^2 \left(\frac{c}{\sqrt{2}\pi f_0 \sqrt{T}} \frac{\sqrt{e_s + e_a}}{e_s}\right)^2. \end{aligned} \tag{S77}$$

Therefore, the standard deviation of the estimated axial velocity by depth difference $\sigma_{v_{\Delta Z}}$ is given as:

$$\sigma_{v_{\Delta Z}} = \sqrt{\mathrm{Var}\left(v_{\Delta Z}\right)} = \frac{c}{\sqrt{2}\pi f_0 \sqrt{T} \Delta t} \frac{\sqrt{e_s + e_a}}{e_s}. \tag{S78}$$

S4.4 Comparisons of Number of Measurements

Comparing the required number of correlation measurements between our approach and Doppler ToF imaging [4] in a fair manner is challenging. This is because our approach estimates depth, intensity, XY-motion, and Z-motion, while Doppler ToF imaging primarily estimates Z-motion (although theoretically, depth and intensity can also be estimated with additional correlation measurements). With bipolar demodulation functions, as utilized in the original image formation of Doppler ToF imaging [4], the process requires a minimum of three correlation measurements to estimate depth, intensity, and Z-motion. In contrast, our approach necessitates a minimum of four correlation measurements to estimate depth, intensity, XY-motion, and Z-motion.

S5 Depth Estimation with Multi-Frequency Coding

Our approach adopts a multi-frequency scheme [6] to achieve high depth precision with a long depth range. While a combination of low and high frequencies [11] can be used to achieve this goal, we use two relatively high frequencies [6] for two correlation image sets to achieve two high-SNR depth maps where a high-quality Z-motion can be estimated by depth difference. In multi-frequency schemes, two correlation image sets are captured with two different modulation frequencies, and two interim depth maps are obtained from the two correlation image sets. While conventional multi-frequency schemes decode one final depth map from the two interim depth maps, our approach estimates two final depth maps from the two correlation image sets to estimate the Z-motion as well.

For each pixel, consider two correlation sets $C_{1,n}$ ($n \in \{1, \ldots, N\}$) and $C_{2,n}$ ($n \in \{1, \ldots, N\}$) which are captured with the modulation frequencies f_1 and f_2, respectively. The measurable depth ranges for $C_{1,n}$ and $C_{2,n}$ are $\frac{c}{2f_1}$ and $\frac{c}{2f_2}$, respectively. Let us assume that Z_1 ($\leq \frac{c}{2f_1}$) and Z_2 ($\leq \frac{c}{2f_2}$) are two interim depth values obtained from $C_{1,n}$ ($n \in \{1, \ldots, N\}$) and $C_{2,n}$ ($n \in \{1, \ldots, N\}$), respectively. Z_1 and Z_2 are different from the true depth values, and our goal is to recover their true depth values. The true depth for $C_{1,n}$ ($n \in \{1, \ldots, N\}$) will be $N_1 Z_1$ ($N_1 \in \mathbb{N}$), and the true depth for $C_{2,n}$ ($n \in \{1, \ldots, N\}$) will be $N_2 Z_2$ ($N_2 \in \mathbb{N}$). We find N_1 and N_2 by minimizing $|N_1 Z_1 - N_2 Z_2|$ under the small Z-motion constraint.

For the multi-frequency scheme using two modulation frequencies f_1 and f_2, the effective modulation frequency f_0 is

$$f_0 = \mathrm{GCD}\left(f_1, f_2\right), \tag{S79}$$

where GCD is the greatest common divisor. The measurable depth range is $\frac{c}{2f_0}$.

S6 Burst Denoising: Increasing Integration Time (and SNR) Computationally

In this section, we describe implementation details for burst denoising with I-ToF correlation images. Burst denoising [1–3, 5, 9, 10] is a popular method to enhance the SNR without introducing motion artifacts or saturation by increasing the capture time *computationally*. It involves capturing a burst of images, each with a short capture time, and aligning and merging them along the motion trajectory to increase the SNR. Burst denoising is computationally efficient enough to be implemented on smartphones [2]. We exploit burst imaging to enhance the SNR of the correlation images and, thus, the resulting depth and intensity estimates. The high-quality depth and intensity estimates obtained through burst denoising also facilitate accurate 3D motion estimation.

S6.1 Constructing a Burst of Correlation Images

Each correlation image within a set is defined as a reference image in turn. For each reference image, we construct a burst of correlation images from the stream of captured frames. Each burst comprises M ($M = 9$ for our simulations and experiments) number of the correlation images (including the reference image) with the same demodulation phase shift (ψ_n in Eq. S5) and the same modulation frequency to ensure consistent brightness for the same scene point.

S6.2 Finding Similar Image Patches

For each pixel of the reference image, we define a reference patch $r\,(x, y)$ ($1 \leq x \leq N_x$, $1 \leq y \leq N_y$) such that the pixel is located at the upper left corner of the reference patch. Next, we define a $S_{\text{intra}} \times S_{\text{intra}} \times S_{\text{inter}}$ search volume such that the reference patch is located at the center of the first slice of the search volume. A target patch $t\,(x, y)$ ($1 \leq x \leq N_x$, $1 \leq y \leq N_y$) with the same size slides over the search volume to find the similar image patches. We define a distance d_{patch} between the reference and target patches as:

$$d_{\text{patch}} = \sum_{x=1}^{N_x} \sum_{y=1}^{N_y} |r(x, y) - t(x, y)|^2. \tag{S80}$$

The values of Eq. S80 over the search volume can be efficiently computed in the frequency domain using 3D FFT. The set of similar image patches is defined as N_{sim} number of image patches with the smallest d_{patch} values. Although only one image patch is found for each slice in the conventional burst denoising, we find multiple similar image patches per slice if there exists abundant spatial correlation in the same slice. If strong Z-motion exists, the source strength attenuates rapidly along the motion, and the correlation value changes even for the same scene point. Therefore, we exploit spatial correlation as much as possible under strong Z-motion. We use $N_x = N_y = 8$, $S_{\text{intra}} = 21$, and $S_{\text{inter}} = 9$ in our simulations and experiments.

S6.3 Wiener Filtering: Merging Image Patches

After finding similar patches for each reference patch, we merge them to get the reconstruction of the reference patch. If we merge them in the original domain (e.g., pixel-wise averaging over similar patches), it is not robust to motion estimation failure [2]. We can achieve more robust reconstruction by merging the similar patches in the frequency domain [2]. If we define the 2D DFT of the similar patches as $T_z\left(f_x, f_y\right)$ $(z \in \{1, \ldots, N_{\text{sim}}\})$ and assume $T_1\left(f_x, f_y\right)$ as the 2D DFT of the reference patch, the reconstruction of T_1 in the frequency domain is given by Wiener filtering:

$$\hat{T}_1\left(f_x, f_y\right) = \frac{1}{N_{\text{sim}}} \sum_{z=1}^{N_{\text{sim}}} T_z\left(f_x, f_y\right) + A_z\left(f_x, f_y\right)\left(T_1\left(f_x, f_y\right) - T_z\left(f_x, f_y\right)\right), \quad \text{(S81)}$$

where

$$A_z\left(f_x, f_y\right) = \frac{\left|D_z\left(f_x, f_y\right)\right|^2}{\left|D_z\left(f_x, f_y\right)\right|^2 + \sigma_N^2}, \quad \text{(S82)}$$

and

$$D_z\left(f_x, f_y\right) = T_1\left(f_x, f_y\right) - T_z\left(f_x, f_y\right). \quad \text{(S83)}$$

σ_N is noise variance. The reconstruction in the original domain can be obtained by inverse DFT of $\hat{T}_1\left(f_x, f_y\right)$. We repeat this for all pixels to reconstruct the high-quality reference image.

S7 Parameter Values for Simulations

Table 1: Parameter values used for simulations.

Figure	$E[e_s]$ (e^-/s)	$E[e_a]$ (e^-/s)	T (ms)	f_0 (MHz)	f_1 (MHz)	f_2 (MHz)
Fig.1	5×10^7	3×10^7	1	10	30	20
Fig.2	3×10^5	10^4	1	20	—	—
Fig.5	10^5	10^4	1	20	60	40
Fig.6	10^5	10^4	1	30	90	60
Fig.7 col1	10^7	10^6	2	10	30	20
Fig.7 col2	10^7	10^6	2	10	30	20
Fig.7 col3	10^7	10^7	2	10	30	20
Fig.7 col4	10^7	10^6	1	10	30	20
Fig.7 col5	10^7	10^6	1	10	30	20
Fig.8	2×10^7	10^6	2	10	30	20

$E[e_s]$: average source strength over all pixels.
$E[e_a]$: average ambient strength over all pixels.
T: integration time for each correlation image.
f_0: effective modulation frequency.
f_1: modulation frequency for correlation image set 1.
f_2: modulation frequency for correlation image set 2.

S8 Algorithm

The overall algorithm of our approach is as follows.

Algorithm 1: High-quality all-in-one I-ToF imaging

Input: A stream of the correlation image sets captured with two modulation frequencies f_1 and f_2 alternately

Output: Streams of high-quality depth maps, intensity images, XY-motion estimates, and Z-motion estimates

for $C_{1,n}$, $n \in \{1, ..., N\}$ *and* $C_{2,n}$, $n \in \{1, ..., N\}$ **do**
  for *n=1 to N* **do**
    Burst denoising;
  end

  $I_1 \leftarrow$ blurred intensity from $C_{1,n}$, $n \in \{1, ..., N\}$;
  $I_2 \leftarrow$ blurred intensity from $C_{2,n}$, $n \in \{1, ..., N\}$;
  $\nabla I_1 \leftarrow$ spatial gradient of I_1;
  $\nabla I_2 \leftarrow$ spatial gradient of I_2;
  $(\Delta X, \Delta Y) \leftarrow$ high-quality XY-motion from ∇I_1 and ∇I_2;
  Align $C_{1,n}$, $n \in \{1, ..., N\}$ and $C_{2,n}$, $n \in \{1, ..., N\}$ along $(\Delta X, \Delta Y)$;
  $Z_1 \leftarrow$ high-quality depth from aligned $C_{1,n}$, $n \in \{1, ..., N\}$;
  $Z_2 \leftarrow$ high-quality depth from aligned $C_{2,n}$, $n \in \{1, ..., N\}$;
  $I_1 \leftarrow$ high-quality intensity from aligned $C_{1,n}$, $n \in \{1, ..., N\}$;
  $I_2 \leftarrow$ high-quality intensity from aligned $C_{2,n}$, $n \in \{1, ..., N\}$;
  $\Delta Z \leftarrow$ high-quality Z-motion from Z_1 and Z_2;
end

S9 Computational Efficiency

The most time-consuming parts of our approach are burst denoising and optical flow for XY-motion estimation (we use RAFT [13] for optical flow). Currently, comparing the runtime between modules is challenging because RAFT and other modules are implemented in Python and Matlab, respectively. RAFT and its lightweight version process 10 frames/s and 20 frames/s, respectively, with 1088 $\times$ 436 videos on a 1080Ti GPU [13]. There are many real-time optical flow methods that we can exploit. Although our unoptimized MATLAB implementation of burst denoising takes about 40 s, it is inherently an efficient and parallelizable algorithm that is now implemented on mobile devices.

S10 More Results for All-In-One I-ToF Imaging

Figs. S2 and S3 show comparisons between conventional and proposed I-ToF imaging for dynamic scenes through simulations and real experiments, respectively. While conventional I-ToF imaging suffers from motion artifacts or low SNR depending on the integration time, our approach can recover not only high-quality 3D geometry and intensity but also 3D motion using a single I-ToF camera.

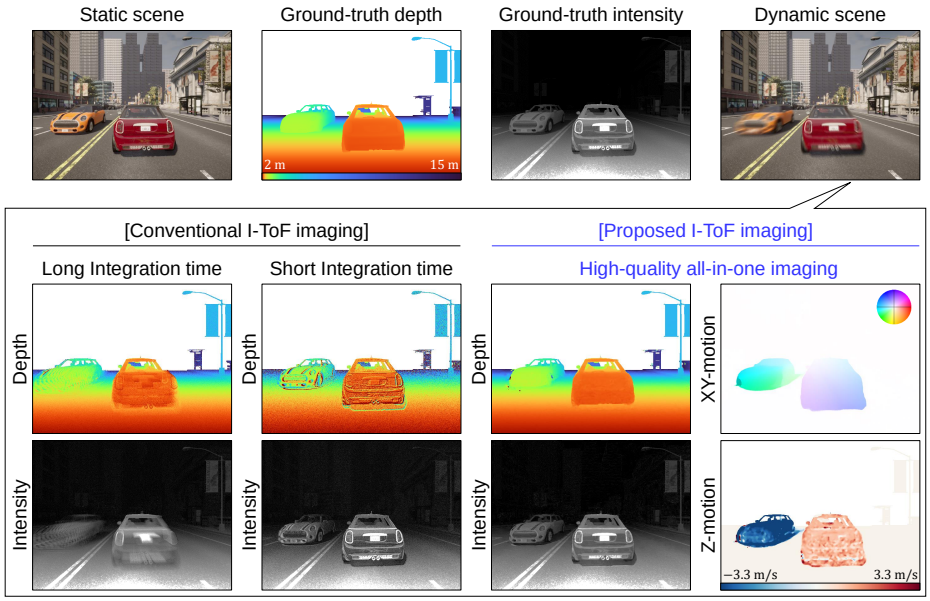

Fig. S2: Simulation results.

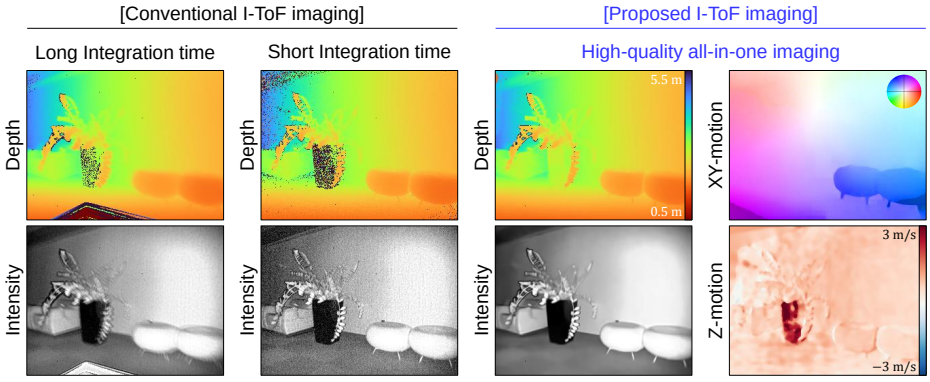

Fig. S3: Experimental results.

S11 Example Application: Hand Gesture Classification

Fig. S4 shows 3D motion estimation results and corresponding motion histograms when a hand moves along the X-, Y-, and Z-directions. As shown in Fig. S4, each hand gesture is represented distinctively by motion histograms, allowing us to classify the hand gestures correctly based on the 3D motion estimation results from our approach.

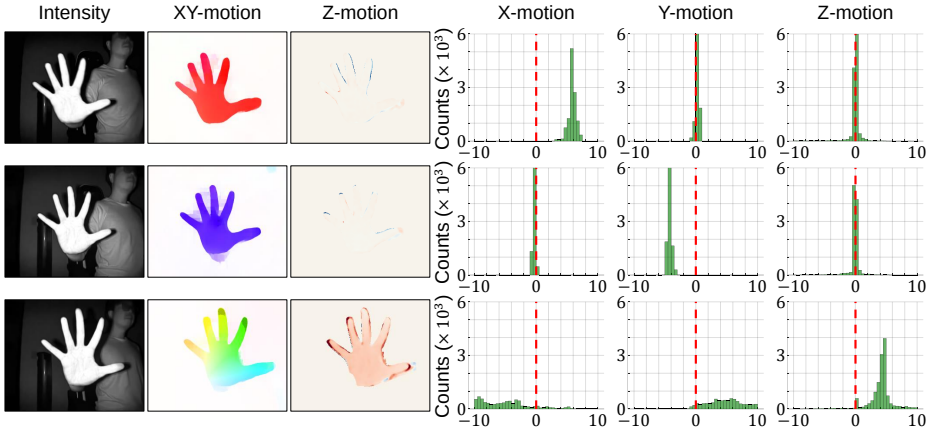

Fig. S4: Hand gesture classification.

References

1. Godard, C., Matzen, K., Uyttendaele, M.: Deep burst denoising. In: Proceedings of the European conference on computer vision (ECCV). pp. 538–554 (2018)
2. Hasinoff, S.W., Sharlet, D., Geiss, R., Adams, A., Barron, J.T., Kainz, F., Chen, J., Levoy, M.: Burst photography for high dynamic range and low-light imaging on mobile cameras. ACM Transactions on Graphics (ToG) **35**(6), 1–12 (2016)
3. Heide, F., Diamond, S., Nießner, M., Ragan-Kelley, J., Heidrich, W., Wetzstein, G.: Proximal: Efficient image optimization using proximal algorithms. ACM Transactions on Graphics (TOG) **35**(4), 1–15 (2016)
4. Heide, F., Heidrich, W., Hullin, M., Wetzstein, G.: Doppler time-of-flight imaging. ACM Transactions on Graphics (ToG) **34**(4), 1–11 (2015)
5. Heide, F., Steinberger, M., Tsai, Y.T., Rouf, M., Pajak, D., Reddy, D., Gallo, O., Liu, J., Heidrich, W., Egiazarian, K., et al.: Flexisp: A flexible camera image processing framework. ACM Transactions on Graphics (ToG) **33**(6), 1–13 (2014)
6. Jongenelen, A.P., Bailey, D.G., Payne, A.D., Dorrington, A.A., Carnegie, D.A.: Analysis of errors in tof range imaging with dual-frequency modulation. IEEE transactions on instrumentation and measurement **60**(5), 1861–1868 (2011)
7. Lange, R.: 3D ToF distance measurement with custom solid-state image sensors in cmos-ccd-technology. Ph.D. Thesis (2000)
8. Lange, R., Seitz, P., Biber, A., Lauxtermann, S.C.: Demodulation pixels in ccd and cmos technologies for time-of-flight ranging. vol. 3965 (2000)
9. Ma, S., Gupta, S., Ulku, A.C., Bruschini, C., Charbon, E., Gupta, M.: Quanta burst photography. ACM Transactions on Graphics (TOG) **39**(4), 79–1 (2020)
10. Mildenhall, B., Barron, J.T., Chen, J., Sharlet, D., Ng, R., Carroll, R.: Burst denoising with kernel prediction networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2502–2510 (2018)
11. Payne, A.D., Jongenelen, A.P., Dorrington, A.A., Cree, M.J., Carnegie, D.A.: Multiple frequency range imaging to remove measurement ambiguity. In: Optical 3-d measurement techniques (2009)
12. Payne, J.M.: An optical distance measuring instrument. Review of Scientific Instruments **44**(3) (1973)
13. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 402–419. Springer (2020)