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

Abstract. Although time-of-flight (ToF) cameras are becoming the sensor-of-choice for numerous 3D imaging applications in robotics, augmented reality (AR) and human-computer interfaces (HCI), they do not explicitly consider scene or camera motion. Consequently, current ToF cameras do not provide 3D motion information, and the estimated depth and intensity often suffer from significant motion artifacts in dynamic scenes. In this paper, we propose a novel ToF imaging method for dynamic scenes, with the goal of simultaneously estimating 3D geometry, intensity, and 3D motion using a single indirect ToF (I-ToF) camera. Our key observation is that we can estimate 3D motion, as well as motion artifact-free depth and intensity by designing optical-flow-like algorithms that operate on coded correlation images captured by an I-ToF camera. Through the integration of a multi-frequency I-ToF approach with burst imaging, we demonstrate high-quality *all-in-one* (3D geometry, intensity, 3D motion) imaging even in challenging low signal-to-noise ratio scenarios. We show the effectiveness of our approach through thorough simulations and real experiments conducted across a wide range of motion and imaging scenarios, including indoor and outdoor dynamic scenes.

Keywords: Time-of-flight imaging · 3D imaging in challenging conditions · 3D motion recovery · Imaging in dynamic scenes

1 A 3D World-in-Motion

Understanding a dynamic 3D world is a complex task, demanding an integrated grasp of geometry, intensity, and motion. While 3D geometry and intensity inform us about the identities and locations of scene objects, 3D motion provides insight into their actions. For instance, for a self-driving car, it is essential not only to detect neighboring vehicles, but also to estimate their motion for safe navigation. For a head-mounted camera on an AR headset, tracking the intricate 3D motion of fingers could enable seamless manipulation of virtual objects. Broadly, the ability to measure dense 3D motion, along with depths and intensities has several applications in robotics, AR, computer vision, and HCI.

Time-of-flight (ToF) cameras [3–5] are a popular sensing technology used to perceive the 3D world. They emit temporally coded light onto the scene and measure its depth and intensity from the reflected light, as shown in Fig. 1. Due to their low cost, low computational complexity, and compact form factors, ToF cameras have rapidly become a method of choice for many commercial 3D

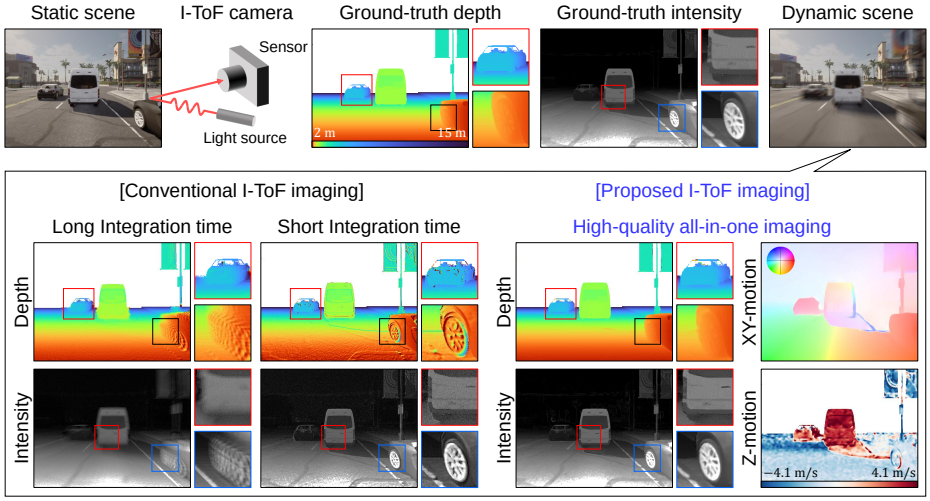

Fig. 1: High-quality all-in-one imaging with a single I-ToF camera. Conventional I-ToF cameras can recover correct 3D geometry and intensity, only for *static* scenes. For dynamic scenes, the depth and intensity estimates suffer from *motion artifacts*. Although the artifacts can be reduced with the short integration time, this results in noisy estimates. Our approach can estimate not only high-quality 3D geometry and intensity but also 3D motion of the dynamic scenes using a *single* I-ToF camera.

applications, including autonomous vehicles, cell phones (e.g., Apple iPhone), and HCI and AR/VR devices (e.g., Microsoft Azure Kinect and Hololens).

However, conventional ToF cameras do not explicitly account for scene or camera motion during capture. As a result, for dynamic scenes, the depth and intensity estimates often suffer from *motion artifacts* especially under rapid motion. Although the motion artifacts can be reduced with short capture times, it results in low signal-to-noise ratio (SNR). This raises the following questions: Is it possible to overcome this noise-vs-motion tradeoff and estimate artifact-free depth and intensity? Going further, although motion is often considered as nuisance due to motion-related artifacts, can we design techniques to actually recover high-resolution (both lateral and axial) 3D motion? To summarize, can we estimate high-quality 3D geometry, intensity, and 3D motion *simultaneously* with a *single* ToF camera for broader applications in the dynamic 3D world?

In this paper, we address these questions using indirect ToF (I-ToF) cameras. I-ToF cameras capture a set of correlation images sequentially to estimate depth and intensity (Sec. 3). For dynamic scenes, the correlation images are *not* aligned due to motion, leading to motion artifacts in the depth and intensity estimates (Fig. 1). Although several approaches [17, 20, 25, 27] have been proposed to reduce these motion artifacts, they typically require two out-of-phase correlation images to be captured simultaneously for motion compensation, which are *not* always available. Modern learning-based approaches for high-quality I-ToF imaging remove depth errors caused by shot noise and multi-path interference, assuming *static* scenes [6, 7, 10, 30, 34] or *no* motion between the correlation im-

ages for dynamic scenes [8], which leads to motion artifacts in dynamic scenarios. In contrast, our approach achieves motion artifact-free, high-quality depth and intensity estimates for the dynamic scenes without needing out-of-phase images captured simultaneously. Furthermore, our method allows for 3D motion estimates, which, to the best of our knowledge, is the first of its kind.

Motion and I-ToF imaging: The fundamental technical challenge in modeling and estimating motion in I-ToF imaging is that the raw correlation images are spatio-temporally coded, and thus *do not preserve brightness constancy*, an inherent assumption for classical optical flow methods [9]. All images within a correlation image set have *different* pixel values even for the same scene point because they are captured with different demodulation functions (Sec. 3). Our key observation is if we consider two correlation image sets captured sequentially in time, *the spatial gradient of the intensity image* estimated from each correlation image set (although misaligned due to motion) *still preserves its brightness along the true motion*. This observation requires an important motion constraint – the motion should be *small* and *linear* across a correlation image set – which can be satisfied with short integration times[3], albeit at the cost of *low* SNR. To overcome this noise-vs-motion tradeoff, we propose a novel multi-frequency I-ToF burst imaging method that *computationally* (not optically) increases the integration time of correlation images, thereby preventing motion artifacts. Obtaining high-quality depth and intensity estimates from the high-SNR correlation images then further improves the accuracy of motion estimates.

Implications: "All-in-one" imaging with a single I-ToF camera. Our approach achieves high-quality 3D geometry, intensity, and 3D motion with a single I-ToF camera by incorporating motion in I-ToF image-formation model from first principles, thereby addressing motion artifacts and low SNR, which have long been the limiting factors of I-ToF cameras. We demonstrate, via thorough simulations and hardware experiments, that our approach can reliably recover 3D geometry and intensity of both indoor and outdoor scenes in challenging imaging scenarios (strong ambient light, low scene albedo, high-speed non-rigid motion), and estimate dense and high-resolution 3D (both lateral and axial) motion. The proposed methods could enable holistic 3D inference in future vision systems by integrating geometry, intensity, and motion information.

2 Related Work

Optical flow and scene flow: Optical flow [18, 28] is a classical technique for measuring dense 2D XY-motion across images. Scene flow [37] is a dense 3D motion field (2D XY-motion + 1D Z-motion) for 3D scene points. Conventional scene flow approaches [21, 26, 35] typically use RGB-D cameras, where color information is used for XY-motion estimation and depth information is used for Z-motion estimation. However, these approaches assume that accurate depth information is available from the depth camera, which is not always true in the case of the dynamic scenes. Our goal is to recover correct depth, intensity, and motion with a single I-ToF camera for dynamic scenes.

[3] A short integration time is also required for instantaneous motion estimation.

Motion artifact reduction in I-ToF imaging: To reduce motion artifacts in I-ToF imaging, several methods [17, 20, 25, 27] capture two out-of-phase correlation images at the same time and get the brightness-conserving images from their sum. After obtaining the lateral (XY) motion between all temporally neighboring correlation images from the correlation-sum images, a depth map is recovered by warping the correlation images along the XY-motion. Unfortunately, these approaches are not applicable when the out-of-phase images are not available at the same time, and when the sum of these images could introduce sensor-dependent artifacts, which is the case for most commercial I-ToF cameras.

Axial motion estimation using I-ToF cameras: A few approaches [15, 19] have been proposed to estimate the axial (Z) motion using I-ToF cameras. These approaches measure the Doppler frequency shift, which is proportional to the object's radial velocity [15]. Although theoretically feasible, these approaches have limited scope in most practical conditions, where the Doppler shift is negligibly small as compared to the modulation frequency of the light source, making it challenging to robustly measure the Z-motion.

Burst imaging: Burst imaging methods [12–14, 16, 29, 31] create a high-quality image from a burst of underexposed noisy images by aligning and merging them along the pixel motion. This way, burst denoising can increase the capture time computationally without motion blur. We draw inspiration from burst imaging methodology for increasing the SNR of the I-ToF correlation images. High SNR correlation images result in high-quality depth and intensity estimates, even in challenging scenarios including low scene albedo and strong ambient light.

Multi-frequency schemes: In I-ToF imaging, higher modulation frequency increases depth accuracy but decreases measurable depth range (Sec. 3). Multi-frequency schemes overcome this trade-off by using two different frequencies [22, 32]. However, they often fail to recover a correct depth map in very low SNR imaging conditions. Our approach overcomes this limitation by integrating the multi-frequency scheme with burst denoising to recover high-quality depth even in these challenging imaging scenarios.

3 I-ToF Image Formation Model

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 $M(t)$ with period T_0. The light emitted by the light source travels to the scene of interest and is reflected back toward the sensor. Each sensor pixel $\mathbf{p}$ computes the correlation $C(\mathbf{p})$ between the radiance of the light incident on $\mathbf{p}$ and a periodic demodulation function $D(t)$ which has the same period as $M(t)$. Several modulation $M(t)$ and demodulation functions $D(t)$ can be used to compute $C(\mathbf{p})$. One example is to use sinusoids for $M(t)$ and $D(t)$[4] [23, 24, 33]:

$$M(t) = 1 + \cos(2\pi f_0 t), \qquad D(t) = \frac{1}{2} + \frac{1}{2}\cos(2\pi f_0 t), \tag{1}$$

[4] We assume a unipolar demodulation function ($0 \leq D(t) \leq 1$) for ease of noise analysis. The same analysis can be extended to a bipolar demodulation function ($-1 \leq D(t) \leq 1$). See the supplementary report.

where the modulation frequency $f_0 = 1/T_0$. In this case, $C(\mathbf{p})$ is

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

where T is the integration time, c is the speed of light, and Z is the scene depth between the camera and the scene point imaged at $\mathbf{p}$. e_s and e_a are the average number of photo-electrons generated at the sensor per unit time by the light source and the ambient light (e.g., sunlight), respectively. $\psi_n = 2\pi(n-1)/N$, $n \in \{1, ..., N\}$ is the phase shift of $D(t)$ by $N(\geq 3)$ times to decode three unknowns e_s, e_a, and Z from the N measured $C_n(\mathbf{p})$. See the supplementary report for the derivation of Eq. 2. It is important to note that the value of C_n changes according to ψ_n even for the same scene point.

Given a set of N correlation values (Eq. 2), the estimated scene depth Z and intensity I for the pixel $\mathbf{p}$ are given by:

$$\hat{Z} = \frac{c}{4\pi f_0}\tan^{-1}\left(\frac{\sum_{n=1}^{N} C_n \sin\psi_n}{\sum_{n=1}^{N} C_n \cos\psi_n}\right), \tag{3}$$

and

$$\hat{I} = \frac{1}{N}\sqrt{\left(\sum_{n=1}^{N} C_n \cos\psi_n\right)^2 + \left(\sum_{n=1}^{N} C_n \sin\psi_n\right)^2} \propto T e_s. \tag{4}$$

We drop $\mathbf{p}$ in $C_n(\mathbf{p})$ for brevity. The intensity I is proportional to the amount of incident signal photons, which is proportional to the scene albedo and inversely proportional to the squared depth. By computing Eqs. 3 and 4 for all pixels, we get a depth map and an intensity image. Since C_n is periodic, the measurable depth range $Z_{\max}$ without ambiguity is limited by

$$Z_{\max} = \frac{c}{2f_0}. \tag{5}$$

Fig. 2 (a) shows the set of correlation images of a *static* scene and the resulting depth map and intensity image.

SNR of depth and intensity estimates: Since C_n (Eq. 2) suffers from Poisson noise, the estimated Z and I by Eqs. 3 and 4 differ from the true Z and I. We can measure the quality of the Z and I estimates by the SNR, which is given as

$$\mathrm{SNR}_Z{}^5 = \frac{Z}{\sigma_Z} = \frac{2\pi f_0}{c}\frac{\sqrt{T}e_s Z}{\sqrt{e_s + e_a}} \tag{6}$$

and

$$\mathrm{SNR}_I = \frac{I}{\sigma_I} = \frac{\sqrt{T}e_s}{2\sqrt{e_s + e_a}}, \tag{7}$$

for the Z and I estimates, respectively, when $N = 4$ (see the supplementary report for the derivations). σ_Z and σ_I are standard deviations of the Z and I

[5] We assume $Z \neq 0$.

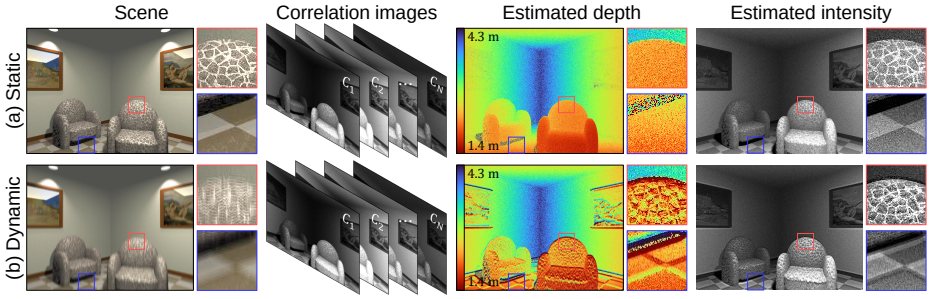

Fig. 2: I-ToF imaging. I-ToF cameras capture a set of correlation images of the scenes to estimate their depth and intensity images. Although I-ToF cameras provide correct depth and intensity information for (a) static scenes, they suffer from motion artifacts for (b) dynamic scenes due to misalignment between the correlation images.

estimates due to noise. Higher quality depth and intensity estimates are possible by increasing the integration time T and source strength e_s, and decreasing the ambient strength e_a. Increasing the modulation frequency f_0 enhances the SNR of depth estimates at the cost of a shorter measurable depth range (Eq. 5).

Artifacts due to motion: In addition to Poisson noise in the correlation images, scene or camera motion also prevents correct depth and intensity estimates. Eqs. 3 and 4 assume there is *no* motion while capturing the N correlation images. If the correlation images are not aligned due to motion, the depth and intensity images estimated by Eqs. 3 and 4 suffer from *motion artifact*, as shown in Fig. 2 (b). The motion artifacts are exacerbated with larger motion or longer integration time. One could reduce the motion artifacts by decreasing the integration time T, but it results in lower SNR, as indicated in Eqs. 6 and 7.

4 Resolving 3D Geometry, Intensity, and 3D Motion with a Single I-ToF Camera

Modeling motion in I-ToF imaging faces a fundamental challenge: the raw correlation images C_n ($n \in \{1, ..., N\}$) are spatio-temporally coded, and thus do not preserve brightness constancy, a widely used assumption for motion computation in conventional camera images. Since all correlation images in each set have *different* brightness values even for the same scene point (Fig. 2), it is challenging to accurately estimate the lateral XY-motion using traditional optical flow.

4.1 XY-Motion Estimation with Brightness-Varying Images

Our *key observation* is that if we consider *two* neighboring correlation image sets[6] under *small* and *linear* motion, we can estimate XY-motion precisely based on brightness conservation:

[6] Most I-ToF cameras provide a temporal stream of the correlation image sets.

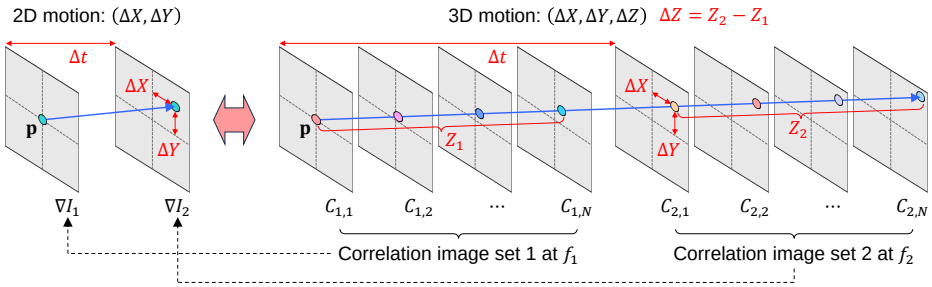

Fig. 3: XY- and Z-motion estimation with brightness-varying correlation images. All correlation images in each set have different pixel values (depicted as distinct colors) along the true XY-motion $(\Delta X, \Delta Y)$, posing a challenge for motion estimation. However, under the small and linear motion, 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.

Observation 1 Consider two correlation image sets captured successively in time, as shown in Fig. 3. If the 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.

See the supplementary report for the proof. Observation 1 holds regardless of whether unipolar or bipolar demodulation functions are used. Observation 1 says that even if all correlation images in each set have different brightness values along the true XY-motion, the spatial gradient of an intensity image (Eq. 4) obtained from each set *preserves* its value along the motion if the motion is *small* and *linear* (Fig. 3).[7] Note that the intensity image is blurred due to the motion. Observation 1 can be mathematically expressed as:

$$\frac{\partial |\nabla I|}{\partial X}\Delta X + \frac{\partial |\nabla I|}{\partial Y}\Delta Y + \frac{\partial |\nabla I|}{\partial t}\Delta t = 0, \tag{8}$$

where I (Eq. 4) is the blurred intensity image[8] and $\nabla = \left(\frac{\partial}{\partial X}, \frac{\partial}{\partial Y}\right)^T$ denotes the spatial gradient. $\frac{\partial}{\partial X}(\cdot)$, $\frac{\partial}{\partial Y}(\cdot)$, and $\frac{\partial}{\partial t}(\cdot)$ are the partial derivatives with respect to X, Y, and time, respectively. ΔX, ΔY, and Δt are the X-motion, Y-motion, and time step between the blurred intensity images as shown in Fig. 3.

Observation 1 is powerful because it allows us to exploit *any* off-the-shelf optical flow algorithm to estimate dense XY-motion by operating on spatial gradients of intensity images obtained from I-ToF correlation image sets. After we

[7] Due to motion, the absolute value of the estimated intensity image does not preserve its brightness even along the true motion. See the supplementary report.

[8] The intensity image I records the *signal* photons, while the image used in traditional optical flow records the *background* photons (e.g., sunlight reflected from the scene).

compute the XY-motion between the blurred intensity images, the finer grained XY-motion between successive correlation images can be simply obtained by interpolation (Fig. 3). Estimating the XY-motion using two correlation image sets has several benefits over conventional motion artifact reduction approaches: 1) While the conventional approaches estimate the motion between all neighboring correlation images independently, we can measure it more efficiently. 2) We can also estimate the Z-motion using depth difference along the XY-motion.

Motion Artifact-Free Depth and Intensity Estimates: After aligning two correlation image sets along the estimated XY-motion (Eq. 8), we can obtain motion artifact-free depth and intensity images for two correlation image sets. For more accurate estimates, we can additionally compensate for the Z-motion using two correlation image sets together. However, in practice, under the small motion constraint, we ignore the Z-motion *within* each correlation image set and estimate the depth and intensity estimates using Eqs. 3 and 4.

4.2 Z-Motion Estimation

Once the XY-motion is determined, we can get two aligned depth maps from the two correlation image sets, and then estimate the Z-motion from their difference (Fig. 3). Although the Z-motion is derived using two depth maps, it can be approximated well as instantaneous motion with a short integration time.

Theoretical comparisons with Doppler ToF: Doppler ToF imaging [15] estimates the Z-motion based on the Doppler effect. Given a scene with an axial velocity v, the emitted light undergoes a Doppler frequency shift when reflected from the scene. If the modulation frequency of the light signal is f_0, the frequency of the signal received at the sensor is $f_0 + \Delta f$, where $\Delta f = \frac{2v}{c} f_0$. Although Doppler ToF allows for instantaneous Z-motion estimation without measuring two depth values, it is challenging to measure Δf (thus axial velocity v) accurately under Poisson noise since Δf is negligibly small, compared to f_0 in practical conditions. Eqs. 9 and 10 are the theoretical standard deviations of the estimated axial velocity by depth difference ($\sigma_{v_{\Delta Z}}$) and Doppler ToF ($\sigma_{v_{\Delta f}}$), respectively.

$$\sigma_{v_{\Delta Z}} = \frac{c}{\sqrt{2}\pi f_0 \sqrt{T}\Delta t} \frac{\sqrt{e_s + e_a}}{e_s} \tag{9}$$

and

$$\sigma_{v_{\Delta f}} = \frac{2\pi c}{f_0\sqrt{T}} \frac{\sqrt{e_s + e_a}}{e_s} \frac{\sqrt{\frac{1}{\left(\Delta f - \frac{1}{T}\right)^2} + \frac{1}{\Delta f^2}}}{\left(\frac{1}{\Delta f - \frac{1}{T}} - \frac{1}{\Delta f}\right)^2} \frac{1}{|\sin\left(2\pi\Delta f T - \phi\right) + \sin\phi|}, \tag{10}$$

where $\phi = \frac{4\pi f_0 Z}{c}$. See Sec. 3 for a glossary of the other variables. See the supplementary report for the derivations of Eqs. 9 and 10.

Fig. 4 shows $\sigma_{v_{\Delta Z}}$ and $\sigma_{v_{\Delta f}}$ as a function of the source strength e_s, axial velocity v, modulation frequency f_0, and scene depth Z. When one of these parameters varies, the other parameters are fixed as $e_s = 5 \times 10^7$ e^-/s, $v =$

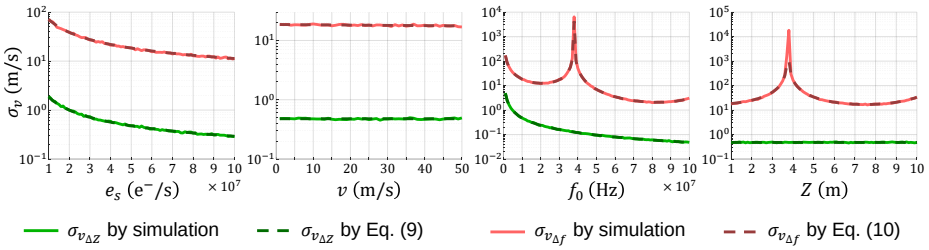

Fig. 4: Comparisons with Doppler ToF. Velocity standard deviations by Doppler ToF ($\sigma_{v_{\Delta f}}$) and by depth difference ($\sigma_{v_{\Delta Z}}$) are compared. Doppler ToF shows about 40 times higher standard deviation in the given practical conditions, and its estimation becomes very unreliable at certain modulation frequencies and depth values.

5 m/s, $f_0 = 10$ MHz, $T = 5$ ms, $\Delta t = 40$ ms, and $Z = 1$ m. We also include the simulation results[9] to verify our derivations. Under the given conditions, $\sigma_{v_{\Delta f}}$ is ~ 40 times higher than $\sigma_{v_{\Delta Z}}$. In addition, axial motion estimates from Doppler ToF have large noise when the term $|\sin(2\pi\Delta f T - \phi) + \sin\phi|$ in Eq. 10 converges to 0 (shown as peaks at certain f_0 and Z values in Fig. 4 and as horizontal error lines in Fig. 8). Estimating the Z-motion from the depth difference can also be challenging when the depth estimates are noisy. We discuss how to mitigate it in Sec. 5. See the supplementary report for further analysis, comparing the number of measurements between the proposed approach and Doppler ToF.

5 High-quality All-in-One I-ToF Imaging

As described in the previous section, Observation 1 allows reliable XY-motion estimation with brightness-varying correlation images, but it requires an important motion constraint: Motion should be *small* and *linear* while capturing two neighboring correlation image sets. This constraint can be satisfied by reducing the integration time, albeit at the cost of *low* SNR of the resulting depth and intensity estimates (Eqs. 6 and 7). Improving the SNR of these estimates is essential because inaccurate depth and intensity estimates lead to imprecise Z- and XY-motion estimates as well. How can we overcome the low SNR due to short integration time to achieve high-quality all-in-one I-ToF imaging?

5.1 Multi-Frequency Coding

We adopt a multi-frequency coding scheme [22] to increase the SNR of the depth and Z-motion estimates. As shown in Eqs. 5 and 6, the SNR of the depth estimates can be improved by increasing the modulation frequency at the cost of the reduced measurable depth range. We can achieve high-depth precision and a large depth range simultaneously by using multiple modulation frequencies. We use two different modulation frequencies f_1 and f_2 to capture two neighboring correlation image sets C_1 and C_2, respectively (Fig. 3). After obtaining

[9] We simulated velocity estimation from depth difference and Doppler ToF under Poisson noise. $\sigma_{v_{\Delta z}}$ and $\sigma_{v_{\Delta f}}$ were computed from 1000 repetitions.

two interim ambiguous depth maps from the two correlation image sets, a final unambiguous depth map is decoded from them.

However, *unlike* conventional multi-frequency schemes where *one* final depth map is recovered from two interim depth maps, we recover *two* depth maps from two correlation image sets to recover the Z-motion as well (see the supplementary report). Instead of using one high and one low frequency [32], we use two high frequencies [22] for two correlation image sets to achieve two high-SNR depth maps, and thus, a high-quality Z-motion estimate as well. Even with the two correlation image sets captured with different frequencies, we can still estimate the XY-motion based on Observation 1 (see the supplementary report).

Caveat of multi-frequency coding: Although the multi-frequency schemes improve the depth accuracy in certain conditions, they often fail under extremely low SNR scenarios. This is because severe noise in the interim depth estimates prevents correct depth decoding. For the multi-frequency schemes to perform well even in challenging imaging conditions, we need to further improve the SNR of the interim depth estimates in a complementary way.

5.2 Increasing Integration Time (and SNR) Computationally

The root cause of the low SNR in the depth and intensity estimates is the *short* integration time for accurate motion estimation. How can we increase the SNR of these estimates without optically extending the integration time? Burst imaging is a popular method for increasing the capture time *computationally* to enhance the SNR without introducing the motion artifact; it captures a burst of images, each with a short capture time, and aligns and merges them along the motion trajectory to increase the SNR. Burst denoising is computationally efficient enough to be implemented on smartphones.

We exploit burst imaging to enhance the SNR of the correlation images and thus, the resulting depth and intensity estimates. For a given reference correlation image, we define a burst of the correlation images from the stream of captured frames. Each burst comprises the correlation images with the same demodulation phase shift (ψ_n in Eq. 2) and the same modulation frequency (f_1 or f_2) to ensure consistent brightness for the same scene point. The correlation images in the burst are aligned and merged to increase the SNR of the reference image. See the supplementary report for more implementation details.

Integrating burst denoising with multi-frequency coding: Multi-frequency schemes and burst denoising improve depth estimation accuracy in complementary ways. Multi-frequency schemes increase the effective modulation frequency, while burst denoising extends the integration time computationally (Eq. 6). Therefore we can considerably improve the depth estimation performance by fusing them; when integrated with the multi-frequency scheme, burst denoising improves the quality of interim depth estimates and reduces decoding errors in the final depth estimates. As illustrated in Fig. 5, this hybrid approach demonstrates high depth estimation performance compared to using either method alone. See the supplementary report for the overall algorithm of our approach.

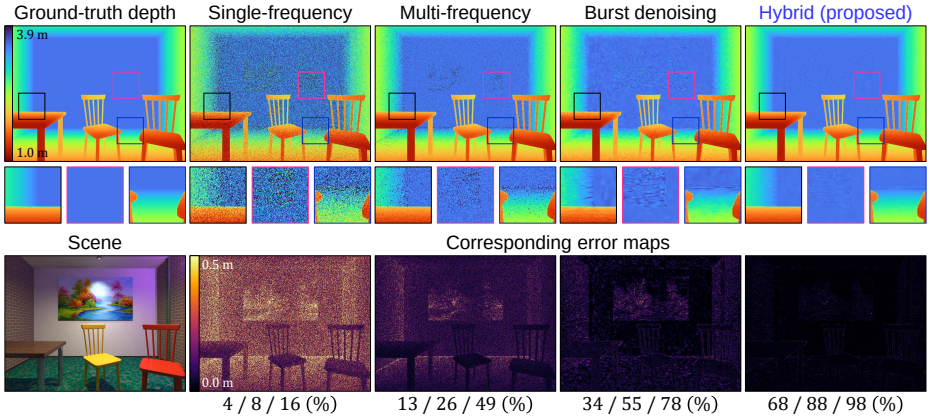

Fig. 5: Enhancing SNR of depth estimates. Although multi-frequency coding achieves lower depth errors than single-frequency coding, it fails to decode the correct depths under extremely low SNR conditions. The performance of multi-frequency coding can be improved when combined with burst denoising, which reduces the depth noise in a complementary way. The three numbers underneath each depth map show the percent fraction of inlier pixels that lie within 0.5, 1, and 2 % of the true depths.

6 Validation by Simulations

In this section, we validate the performance of our approach through simulations. This enables us to quantitatively compare our approach with the ground-truth and alternative methods. We can also simulate various motion scenarios and imaging parameters, such as modulation frequency, integration time, and lighting conditions. We model indoor scenes [1] using POVray, a ray tracing tool [2], and outdoor scenes using the CARLA simulator [11]. See the supplementary report for the parameter values used for all simulations.

Ground-truth comparisons of 3D geometry estimates: Fig. 6 shows the dynamic scenario (due to camera motion) and its ground-truth depth, along with the depth estimates using conventional and proposed approaches. While decreas-

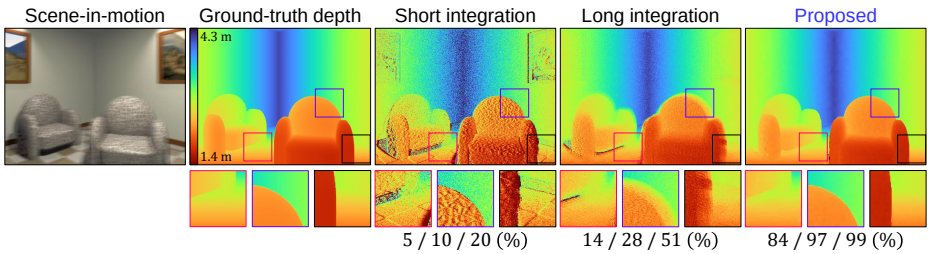

Fig. 6: Ground-truth comparisons of depth estimates. Depth estimates of the dynamic scene obtained via conventional I-ToF imaging suffer from noise or motion artifacts. In contrast, our approach recovers high-quality 3D geometry without such artifacts. The three numbers underneath each depth map show the percent fraction of inlier pixels that lie within 0.5, 1, and 2 % of the true depths.

Fig. 7: Resolving 3D motions in various motion scenarios. Our approach can estimate dense and high-quality 3D motions for various dynamic scenes. The exact motion scenarios can be identified correctly from our XY- and Z-motion estimates.

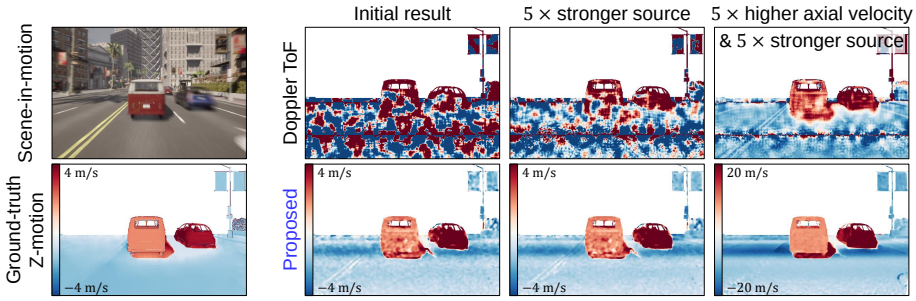

Fig. 8: Axial motion estimation comparison. Doppler ToF [15] cannot estimate small Z-motions ($\sim 4\,\mathrm{m/s}$) accurately since the corresponding Doppler frequency shifts ($< 1\,\mathrm{Hz}$) are negligibly small compared to the modulation frequency (MHz). In contrast, our approach can resolve even these small Z-motions reliably with lower source power.

ing the integration time can reduce motion artifacts in conventional methods, it leads to noisy estimates. The extended integration time reduces noise but introduces motion blur. In contrast, our approach effectively mitigates both noise and motion artifacts in the depth estimates.

3D motion estimation under various motion scenarios: We simulate various motion scenarios of an I-ToF camera attached to a moving car using the CARLA simulator [11]. Fig. 7 shows the 3D motion estimation results of various dynamic scenes using our approach. Our method reliably estimates 3D motions across different motion scenarios. We estimate XY-motion from the gradient of the I-ToF intensity image using RAFT [36] in our simulation and experimental results, but any state-of-the-art optical flow method may be employed.

Z-motion estimation comparisons: We compare Z-motion estimation performance between our approach and Doppler ToF [15], which measures instanta-

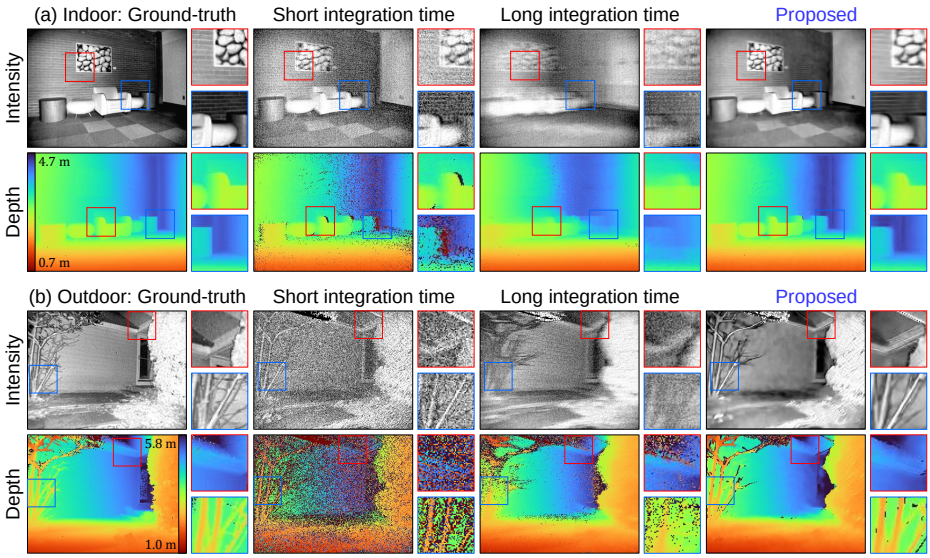

Fig. 9: Recovering 3D geometry and intensity of dynamic scenes. 3D geometry and intensity estimates by conventional I-ToF imaging suffer from low SNR and motion artifacts in both (a) indoor and (b) outdoor dynamic scenes. In contrast, our approach can recover high-quality and motion artifact-free estimates.

neous axial motion based on the Doppler effect. Fig. 8 shows a dynamic scene, its ground-truth Z-motion, and the estimated Z-motions by Doppler ToF and our approach. For the Doppler ToF estimates, we also applied a binning-based non-local means denoiser to increase the SNR, as described in [15]. As depicted in Fig. 8, Doppler ToF cannot robustly estimate small axial motions, such as 4 m/s, as the corresponding Doppler shift (<1 Hz) is negligibly small compared to the modulation frequency (in the MHz range). In contrast, our approach can reliably estimate even these relatively small axial motions with lower source power.

7 Hardware Prototyping and Experimental Results

We implemented our approach in hardware using the Chronoptics KeaB I-ToF camera, which provides access to the raw correlation images. We use two modulation frequencies, 40 MHz and 50 MHz, to capture two neighboring correlation image sets. We set the integration time to 2 ms and 3 ms for indoor and outdoor scenes, respectively. We employ 4 and 6 correlation images for each set for indoor and outdoor scenes, respectively. The camera resolution is 240×320 pixels.

Recovering 3D geometry and intensity of dynamic scenes: Fig. 9 (a) and (b) show recovery results for indoor and outdoor scenes, respectively. For comparison, we captured the correlation images with short integration times (indoor: 2 ms, outdoor: 3 ms) and long integration times (indoor: 18 ms, outdoor: 27 ms). The ground-truth data was obtained with the same, but static scenes by averaging 1000 correlation images captured with the short integration times. As

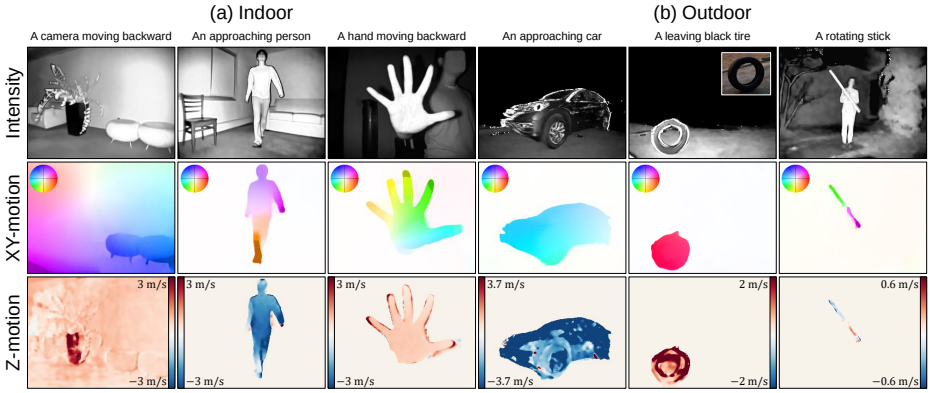

Fig. 10: Resolving 3D motions. Our approach can recover 3D motions reliably for (a) indoor and (b) outdoor dynamic scenes. Our approach can recover local and global motions under challenging conditions such as a low scene albedo and a thin object.

shown in Fig. 9, the estimates obtained with short integration times exhibit low SNR, while those obtained with long integration times suffer from motion artifacts. In contrast, our approach recovers high-SNR 3D geometry and intensity, free from motion artifacts, for both indoor and outdoor dynamic scenes.

Recovering 3D motion for indoor and outdoor scenes: Fig. 10 shows 3D motion estimation results for (a) indoor and (b) outdoor dynamic scenes. Our approach can recover both XY- and Z-motions reliably for both local scene motions and global camera motions, including in challenging scenarios such as a black tire (low scene albedo) and a fast rotation of a stick (intricate geometry). See the supplementary report for more results and visualizations.

8 Limitations and Future Outlook

Dependency on XY-motion estimation: Since our method resolves 3D geometry, intensity, and 3D motion of dynamic scenes after 2D motion estimation on the spatial gradients of the intensity images, its performance strongly depends on the results of XY-motion estimation, which exploits only intensity information. A promising line of future work is to leverage both geometry and intensity information for motion estimation, as available in the I-ToF correlation images.

Camera resolution and light source power: A higher spatial resolution of the I-ToF camera enables higher quality depth and intensity estimates as well as finer motion estimates. A stronger light source allows for long-range imaging even under high ambient light conditions. The prototype used for our experiments has a relatively small resolution (240×320 pixels) and limited available source power, thus constraining the performance of our approach. Recently, I-ToF cameras have seen significant improvements in spatial resolution (> 1 megapixels) and source power, which will enable higher quality all-in-one I-ToF imaging in the future.

Acknowledgement. This research was supported in part by NSF CAREER award 1943149, Cruise LLC, WARF, and ONR award N00014-24-1-2155.

References

1. http://www.ignorancia.org/index.php?page=lightsys, accessed: 2024-07-14
2. http://www.povray.org/, accessed: 2024-07-14
3. 3d depth sensing development kits, pmd. https://3d.pmdtec.com/en/3d-cameras/flexx2/, accessed: 2024-07-14
4. Azure kinect dk, microsoft. https://www.microsoft.com/en-us/d/azure-kinect-dk/8pp5vxmd9nhq?activetab=pivot:overviewtab, accessed: 2024-07-14
5. Time-of-flight sensors, texas instruments. https://www.ti.com/sensors/specialty-sensors/time-of-flight/products.html, accessed: 2024-07-14
6. Agresti, G., Schaefer, H., Sartor, P., Zanuttigh, P.: Unsupervised domain adaptation for tof data denoising with adversarial learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5584–5593 (2019)
7. Agresti, G., Zanuttigh, P.: Deep learning for multi-path error removal in tof sensors. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
8. Attal, B., Laidlaw, E., Gokaslan, A., Kim, C., Richardt, C., Tompkin, J., O'Toole, M.: Törf: Time-of-flight radiance fields for dynamic scene view synthesis. Advances in neural information processing systems **34**, 26289–26301 (2021)
9. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part IV 8. pp. 25–36. Springer (2004)
10. Dong, G., Zhang, Y., Xiong, Z.: Spatial hierarchy aware residual pyramid network for time-of-flight depth denoising. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16. pp. 35–50. Springer (2020)
11. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017)
12. Godard, C., Matzen, K., Uyttendaele, M.: Deep burst denoising. In: Proceedings of the European conference on computer vision (ECCV). pp. 538–554 (2018)
13. 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)
14. 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)
15. Heide, F., Heidrich, W., Hullin, M., Wetzstein, G.: Doppler time-of-flight imaging. ACM Transactions on Graphics (ToG) **34**(4), 1–11 (2015)
16. 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)
17. Hoegg, T., Lefloch, D., Kolb, A.: Real-time motion artifact compensation for pmd-tof images. In: Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications: Dagstuhl 2012 Seminar on Time-of-Flight Imaging and GCPR 2013 Workshop on Imaging New Modalities. pp. 273–288. Springer (2013)
18. Horn, B.K., Schunck, B.G.: Determining optical flow. Artificial intelligence **17**(1-3), 185–203 (1981)

19. Hu, Y., Miyashita, L., Ishikawa, M.: Differential frequency heterodyne time-of-flight imaging for instantaneous depth and velocity estimation. ACM Transactions on Graphics (TOG) **42**(1), 1–13 (2022)
20. Hussmann, S., Hermanski, A., Edeler, T.: Real-time motion artifact suppression in tof camera systems. IEEE Transactions on Instrumentation and Measurement **60**(5), 1682–1690 (2011)
21. Jaimez, M., Souiai, M., Gonzalez-Jimenez, J., Cremers, D.: A primal-dual framework for real-time dense rgb-d scene flow. In: 2015 IEEE international conference on robotics and automation (ICRA). pp. 98–104. IEEE (2015)
22. 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)
23. Lange, R.: 3D ToF distance measurement with custom solid-state image sensors in cmos-ccd-technology. Ph.D. Thesis (2000)
24. Lange, R., Seitz, P., Biber, A., Lauxtermann, S.C.: Demodulation pixels in ccd and cmos technologies for time-of-flight ranging. vol. 3965 (2000)
25. Lefloch, D., Hoegg, T., Kolb, A.: Real-time motion artifacts compensation of tof sensors data on gpu. In: Three-Dimensional Imaging, Visualization, and Display 2013. vol. 8738, pp. 166–172. SPIE (2013)
26. Letouzey, A., Petit, B., Boyer, E.: Scene flow from depth and color images. In: BMVC 2011-British Machine Vision Conference. pp. 46–1. BMVA Press (2011)
27. Lindner, M., Kolb, A.: Compensation of motion artifacts for time-of-flight cameras. In: Dynamic 3D Imaging: DAGM 2009 Workshop, Dyn3D 2009, Jena, Germany, September 9, 2009. Proceedings. pp. 16–27. Springer (2009)
28. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: IJCAI'81: 7th international joint conference on Artificial intelligence. vol. 2, pp. 674–679 (1981)
29. 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)
30. Marco, J., Hernandez, Q., Munoz, A., Dong, Y., Jarabo, A., Kim, M.H., Tong, X., Gutierrez, D.: Deeptof: off-the-shelf real-time correction of multipath interference in time-of-flight imaging. ACM Transactions on Graphics (ToG) **36**(6), 1–12 (2017)
31. 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)
32. 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)
33. Payne, J.M.: An optical distance measuring instrument. Review of Scientific Instruments **44**(3) (1973)
34. Su, S., Heide, F., Wetzstein, G., Heidrich, W.: Deep end-to-end time-of-flight imaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6383–6392 (2018)
35. Sun, D., Sudderth, E.B., Pfister, H.: Layered rgbd scene flow estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 548–556 (2015)
36. 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)

37. Vedula, S., Rander, P., Collins, R., Kanade, T.: Three-dimensional scene flow. IEEE transactions on pattern analysis and machine intelligence **27**(3), 475–480 (2005)