https://doi.org/10.2352/EI.2025.37.7.ISS-287

A Novel Multimodal 3D Depth Sensing Device

Jian Ma, Shenge Wang, Matthieu Dupre, Ioannis Nousias, and Sergio Goma
Qualcomm Technologies, Inc., 5775 Morehouse Dr., San Diego, CA 92121, USA

Abstract

We introduce an innovative 3D depth sensing scheme that seamlessly integrates various depth sensing modalities and technologies into a single compact device. Our approach dynamically switches between depth sensing modes, including iTOF and structured light, enabling real-time data fusion of depth images. We successfully demonstrated iToF depth imaging without multipath interference (MPI), simultaneously achieving high image resolution (VGA) and high depth accuracy at a frame rate of 30 fps.

Introduction

Time of Flight (ToF) and Structured Light (SL) are the two prominent active depth sensing technologies currently available. ToF can be divided into two categories: direct Time of Flight (dToF) and indirect Time of Flight (iToF). dToF, iToF and SL have different advantages and disadvantages. The following is a brief summary:

dToF technology precisely detects distances by measuring the time it takes for light pulses to travel to an object and back. It relies on a single photon avalanche diode (SPAD) detector. dToF offers high accuracy and an extended detection range. Typically employed in LiDAR systems, it scans a pulsed laser beam point by point in space. This makes it ideal for applications requiring long-range detection without the need for high-resolution imaging. Examples include autonomous vehicles, aerial inspections, surveys, and mapping. Recently, low-resolution 2D SPAD arrays have emerged, potentially enhancing image resolution and frame rates.

iToF technology gauges distance by measuring the phase shift of modulated light within each pixel. It enables the creation of detailed 3D depth images with high imaging resolution, such as VGA or higher. Commercially available medium to high-resolution iToF sensors facilitate this capability.

To achieve the full sensor image resolution, diffused illumination is employed. However, this introduces multipath interference (MPI). MPI occurs when light from a projector reaches an iToF camera pixel via multiple routes. Typically, there exists a direct path plus a number of indirect paths caused by light scattering from multiple surfaces. Efforts have been made to mitigate MPI, although success has been limited [1]. These efforts include deep learning convolutional neural networks [2-3], MPI modeling [4], multiple modulation frequencies [5], ToF and stereo data fusion [6], and ToF and SL fusion [7].

Sparse spotlight illumination for iToF depth measurements can significantly reduce MPI because the indirect paths are often much weaker than the direct path. However, the resulting depth image is sparse, constrained by the number of illuminated spots (e.g., <5,000).

Structured Light (SL) depth detection operates by projecting a known pattern onto a scene and measuring the disparity of that pattern using a camera. Typically, this pattern consists of pseudo-random dots generated by a VCSEL array in conjunction with a diffractive optical element (DOE). Unlike iToF, SL does not suffer from multipath interference (MPI).

However, SL has limitations. Its effective range is typically restricted to a couple of meters due to the small disparity (<1 pixel) observed when the object is too distant. Additionally, SL patterns often involve multiple dots (e.g., ≥ 4x4) per code, which imposes constraints on the minimum detectable object size.

Our Approach

To address iToF MPI while preserving high imaging resolution, we propose integrating multiple depth sensing modalities into a compact device [8]. Our key innovation involves a switchable diffuser that toggles between two modes: diffuse and transparent, synchronized with the iToF camera frame rate. Figure 1 illustrates the device schematic and a functional prototype.

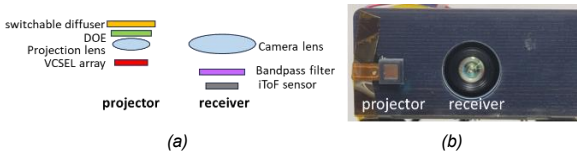

Figure 1. *A multimodal 3D depth sensing device. (a) The scheme of components arrangement; (b) A prototype*

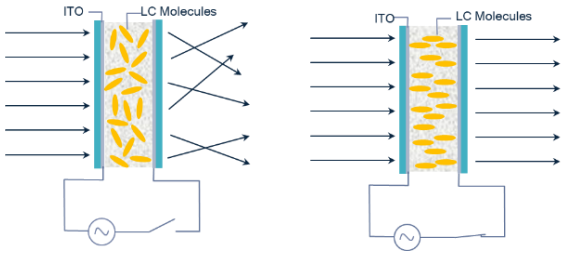

Figure 2. *A liquid crystal switchable diffuser*

The projector comprises a 940nm VCSEL array, a diffractive optical element (DOE), and a projection lens for pseudo-random dot projection. The VCSEL operates with modulated current pulses (~10ns) at a ~10% duty cycle. Placed in front of the DOE, the switchable diffuser enables the multimodal operation. The receiver consists of a camera lens and a 640x480 resolution iToF sensor with 5x5 um pixel pitches.

The switchable diffuser, a liquid crystal (LC) device measuring 5 x 5 x 0.5 mm, operates based on the principle illustrated in Figure 2. Without applied voltage, LC molecules exhibit random orientation, leading to a diffuser-like effect due to the orientation-dependent refractive index distribution across the LC. However, when voltage is applied, all LC molecules align uniformly, rendering the LC transparent. The LC can fast switch between diffuse and clear states in less than 5 ms, facilitating multimodal operation across frames.

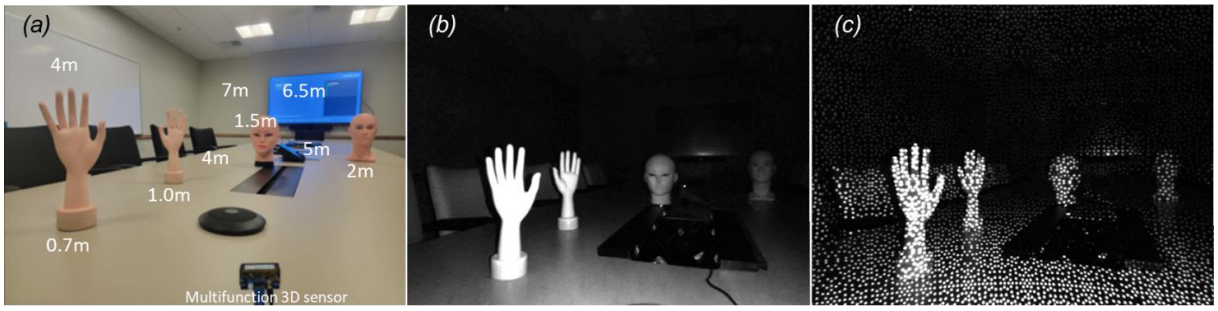

Figure 3. *(a) Objects distance to be measured. (b) d-iToF illumination. (c) s-iToF illumination.*

When the diffuser is in diffuse mode, the device functions as a regular diffuse iToF (d-iToF). Conversely, when the diffuser is in transparent mode, the device operates as a spot iToF (s-iToF) and simultaneously serves as a structured light (SL) system, contingent upon the algorithm employed for image processing. Figure 3 illustrates the illumination images corresponding to these two modes.

In addition to d-iToF, s-iToF, and SL, we have developed a novel depth algorithm called Spot Centroid Tracing (SCT). Unlike SL, which determines the depth via a group of dots (codeword), SCT determines the depth of each individual dot.

Consequently, our device captures four depth images across two consecutive frames: d-iToF depth in the first frame (with the diffuser on) and s-iToF, SCT, and SL depths in the second frame (with the diffuser off). Since we use the same iToF projector and receiver, all images are self-aligned.

To achieve highly accurate and high-resolution depth imaging while significantly reducing or eliminating multipath interference (MPI), we employ three data fusion options:

(1) d-iToF + s-iToF
(2) d-iToF + SCT
(3) d-iToF + SL

In essence, we combine a high-resolution depth image acquired using d-iToF with a precise, MPI-free low-resolution depth image (s-iToF, SCT, or SL).

Spot Centroid Tracing (SCT) Algorithm

SCT (Spot Centroid Tracing) is an algorithm recently developed by our group. It can be described as a structured light (SL) approach performed at the dot level, rather than at the codeword level. Unlike conventional SL patten matching algorithms (such as NCC or SGM) which are computational expensive, SCT efficiently computes most of its operations during calibration, resulting in a lightweight and real-time implementation. The key insight lies in determining the disparity (or depth) of each dot based on its centroid coordinates on the sensor. The SCT algorithm involves three steps:

(1) Begin by rasterizing the (x, y) pixels of the s-iToF image. If a pixel qualifies as a local maximum, it is identified as a dot. Then compute the centroid of this dot within a (5x5) window, achieving sub-pixel resolution [see Figure 4(a)]. Each spot corresponds to a unique trace [Figure 4 (b-c)], representing its trajectory on the sensor due to disparity.

(2) The 1st lookup table (LUT) identifies which traces are likely to be present at a given pixel. Since dots are sparsely distributed in the sensor frame, each trace can be uniquely identified. Simultaneously, the depth of each spot is measured using s-iToF in the same frame. The s-iToF depth information serves to disambiguate between any residual potential traces.

(3) Utilizing a second lookup table (LUT), we calculate the depth corresponding to the trace obtained for the centroid located at coordinates (xc, yc).

To minimize ambiguity, we optimize the dot pattern, ensuring that each candidate is uniquely identifiable. Notably, SCT achieves remarkable accuracy, detecting objects as small as individual dots. SCT is robust against MPI and noise. However, akin to SL, its primary limitation lies in the restricted range due to sub-pixel uncertainty. With our ToF sensor's pixel size of 5 μm, the accurate depth range (error < ±0.5%) spans from 150 mm to 1000 mm.

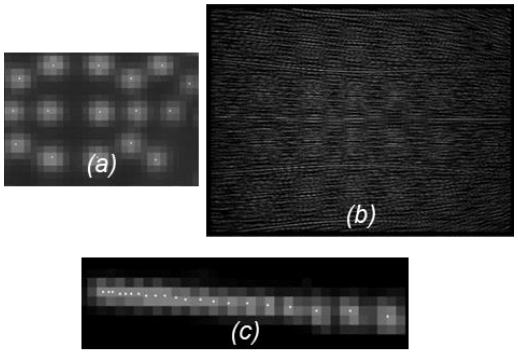

Figure 4. *SCT (spot centroid tracing). (a) Spot centroid is calculated with 5x5 pixels. (b) Traces of ~5,000 spots on an entire sensor. (c) an example of a trace with 21 positions for distances of 10-30 inches*

Multimodal Depth Imaging

The performance differences among d-iToF, s-iToF, SL, and SCT for objects without significant MPI are depicted in Figure 5. These differences correspond to the depth images of the objects shown in Figure 3 and are summarized:

d-iToF [Figure 5(a)] is capable of detecting the smallest objects and sharp edges.

s-iToF [Figure 5(b)] offers the longest detection range due to concentrated light energy in the spots. However, the resulting image is defined by sparse dots, making it challenging to detect sharp edges of objects.

SCT [Figure 3(c)] and SL [Figure 3(d)] both exhibit a shorter range. However, SL performs poorly in detecting small objects—for instance, it is unable to distinguish between fingers, whereas SCT can.

The performance differences for detecting objects with significant MPI are illustrated in Figure 6. The scene involves a 60-degree corner wall located at 500 mm from the device. Their characteristics are summarized:

SCT [Figure 6(a)] is free from MPI and can detect sharp corners. Consequently, it provides the most accurate depth

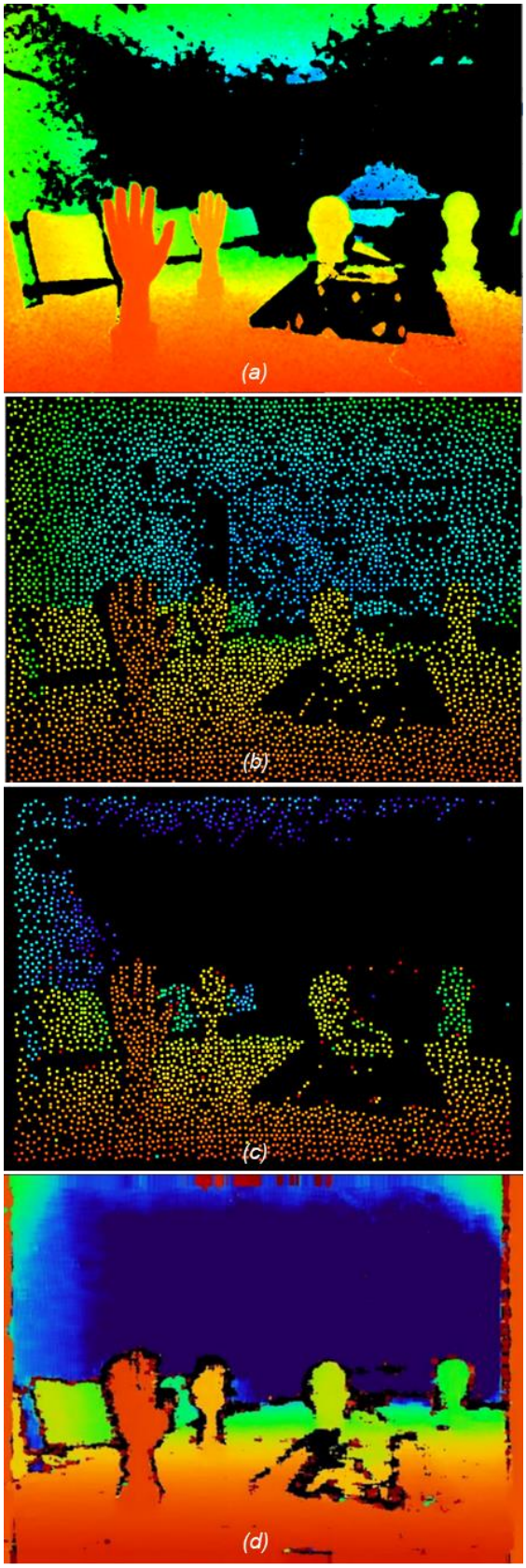

Figure 5. *The depth images of Figure 3. (a) Diffuse-iToF depth. (b) Spot-iToF depth. (c) SCT depth. (d) SL depth*

estimation. Its depth image is served as the ground truth for calculating the errors of the other depth images.

SL [Figure 6(b)] is also free from MPI, but it cannot detect sharp corners. On average, it exhibits a deviation of 14mm.

d-iToF [Figure 6(c)] exhibits the largest error, with an average deviation of 68mm, primarily attributed to MPI.

s-iToF [Figure 6(d)] demonstrates significantly improved performance compared to d-iToF primarily because it is less affected by MPI. However, MPI is not completely eliminated, resulting in a residual average error of 21mm.

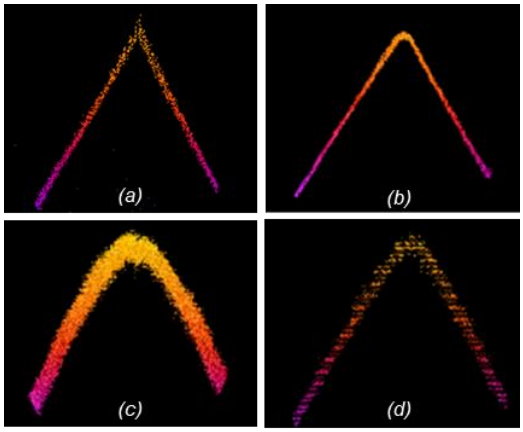

Figure 6. *Depth images of a 60º corner wall (top-view). (a) SCT, ground truth. (b) SL, average error: 14mm. (c) d-iToF, average error: 68mm. (d) s-iToF, average error: 21mm.*

Multimodal Data Fusion

To leverage the benefits of both high image resolution from d-iToF depth images and the good depth accuracy of SCT and s-iToF, we have implemented real-time data fusion. Specifically, for distances ≤1 meter, fusing d-iToF with SCT yields optimal results; and for distances >1 meter, combining d-iToF with s-iToF is preferred. Given the limited merit of SL compared to SCT, we do not present data fusion results involving SL.

The depth fusion algorithms are illustrated in Figures 7 and 8. Figure 7 presents the depth fusion flow diagram, which comprises three components: SCT, s-iToF, and d-iToF. Both SCT and s-iToF involve peak detection. SCT depth is determined using the LUT, while s-iToF and d-iToF depths are calculated based on the phases of the received light pulses. Additionally, the d-iToF depth image provides extra details, such as edges, shadows, and fine features.

High-accuracy depth measurements via spots are obtained in two ways: for distances up to 1 meter, SCT is used; for distances greater than 1 meter, s-iToF is employed. These measurements serve as anchor points for constructing the depth image.

The green box labeled “Multimodal Depth Fusion” in Figure 7 is further detailed in Figure 8. To achieve a smooth and continuous 3D surface, triangulation-based data interpolation is applied between the depth points acquired by either SCT or s-iToF. To preserve edges and image details, the following approach is used:

- If there are no edges between the points, data interpolation between the triangular points is used to fill the surface.
- If there is an edge between the points, gradient depth mapping of d-iToF depth to the spot depth is used, with the edge serving as a smooth boundary in filling the peak depth

surface. In other words, the triangulation-based filling will not cross the edges.

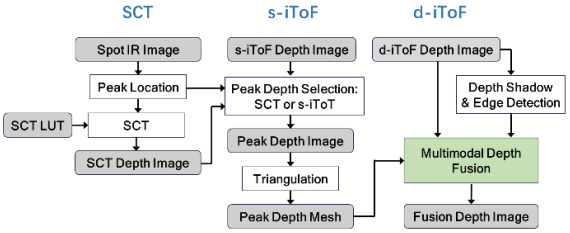

Figure 7. *Depth fusion algorithm flow diagram*

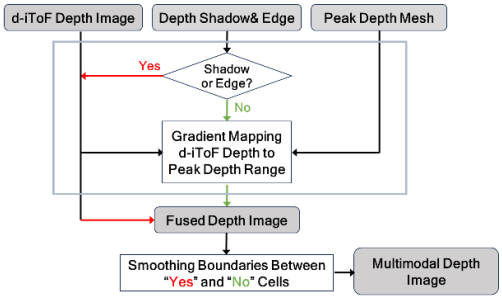

Figure 8. *Multimodal depth fusion algorithm flow diagram – the expansion of the green box in Figure 7.*

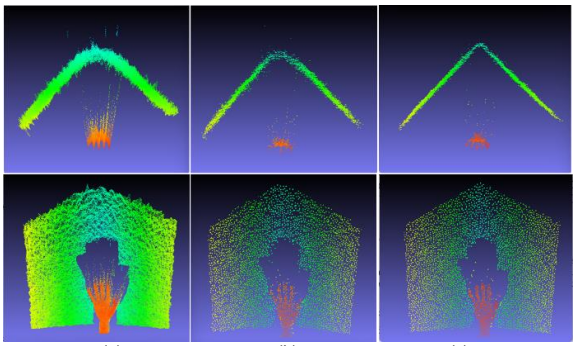

Figure 9. *Multimodal 3D point cloud images of a hand in front of a 90° MPI corner, top-view (top row), and front-view (bottom row). Obtained with: (a) d-iToF. (b) s-iToF. (c) SCT.*

Figures 9 and 10 showcase the effectiveness of our approach, particularly in achieving high-resolution, MPI-free iToF imaging. In this scenario, a mannequin hand is positioned approximately 300 mm in front of a 90° corner wall. The sensing device is situated about 400 mm away from the hand.

This setup intentionally combines challenging conditions: The 90° angled plane introduces strong MPI, and the fine-resolution object (fingers) adds complexity. In addition, edge scattering from the hand introduces false depth information.

Figure 9 (a) displays the 3D point cloud images obtained from d-iToF, revealing the curved/ballooned corner and strong edge scattering-induced false depth.

Figure 9 (b) shows the 3D point cloud images from s-iToF, where the corner shape is significantly improved, and edge scattering-induced false depth is reduced. However, the residual MPI effect is still visible in the corner shape.

Finally, Figure 9 (c) presents the most accurate depth image achieved using SCT. Two data fusion options were implemented, and the results are shown in Figure 10 (a) for the fusion of d-iToF with SCT, and Figure 10 (b) for the fusion of d-iToF with s-iToF. Here's what we observed:

The top-view shapes of the 90° walls in the fused 3D point cloud images closely resemble either SCT or s-iToF. This resemblance suggests that in both cases, MPI is either eliminated (when fused with SCT) or significantly reduced (when fused with s-iToF). Additionally, the false depth caused by edge scattering is also eliminated or substantially reduced.

Most importantly, the high-resolution details—such as the edges of fingers—remain nearly as well-preserved as in the original d-iToF image.

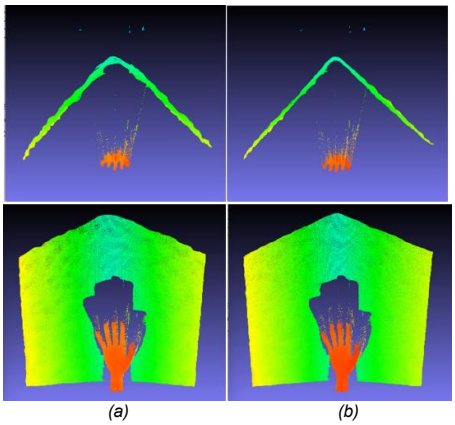

Figure 10. *Multimodal data fusion results of 3D point cloud image obtained with real-time data fusion of (a) d-iToF with SCT, and (b) d-iToF with s-iToF.*

Depth Accuracy

The device is calibrated with the process involving measuring the depth of a flat wall together with two commercial laser distance meters with ±1mm accuracy. The distance measured by the laser is served as the ground truth. Depth images were acquired for both d-iToF and s-iToF at varying distances. To account for nonlinearities and ensure planarization, a polynomial fitting was applied. The same s-iToF images were also utilized for the SCT calibrations.

The depth accuracy of the calibrated device was assessed under conditions without MPI, yielding a typical result depicted in Figure 11.

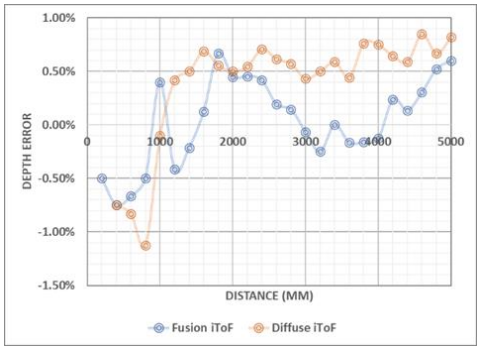

Figure 11. *Depth accuracy of d-iToF and multimodal data fused iToF.*

We are able to achieve an accuracy, with errors consistently within a tight margin of less than ±1% across distances ranging from 150mm to 5000mm, using our data fusion approach. It's interesting to point out that d-iToF exhibits positive errors for

distances beyond 1 meter. This phenomenon is likely attributed to MPI effects arising from the projection light scattering off the ceiling and floor.

Summary

In summary, we have successfully implemented a novel multimodal 3D depth sensing device that seamlessly integrates multiple depth sensing methods into a single compact device, without compromising on size or cost. Each modality can operate independently or in combinations, resulting in high-resolution, highly accurate, and MPI-free depth imaging through real-time data fusion.

References

[1] C. Bamji; et al, "A Review of Indirect Time-of-Flight Technologies", IEEE Transactions on Electron Devices; Vol. 69, Issue 6, June 2022, p2779-2793.

[2] K. Son; et al, "Learning to Remove Multipath Distortions in Time-of-Flight Range Images for a Robotic Arm Setup", 2016 IEEE International Conference on Robotics and Automation; 16-20 May 2016

[3] G. Agresti; et al, "Deep Learning for Multi-path Error Removal in ToF Sensors", Proceedings, Computer Vision – ECCV 2018 Workshops, P410-426.

[4] S. Fuchs, "Multipath Interference Compensation in Time-of-Flight Camera Images", 20th International Conference on Pattern Recognition, 23-26 August 2010

[5] A. Bhandari; et al, "Resolving Multi-path Interference in Time-of-Flight Imaging via Modulation Frequency Diversity and Sparse Regularization", Optics Letters, Vol. 39, Issue 6, April 2014

[6] C. D. Mutto, et al, "Locally Consistent ToF and Stereo Data Fusion", Computer Vision – ECCV 2012, Workshops and Demonstrations, p598-607.

[7] Gu, F.; et al, "Depth Recovery Method Based on the Fusion of Time-of-Flight and Dot-Coded Structured Light", Photonics 2022, 9(5), 333.

[8] B. Hseih, J. Ma, S. Goma, "Resolving multipath interference using a mixed active depth system", US Patent 11561085, filed: Nov 2019, published: Dec 2020, granted: Jan 2023

Dr. Jian Ma has been a Principal Engineer/Manager at Qualcomm since 2010. He is currently part of Qualcomm's QCT Multimedia R&D organization, where he is responsible for the development of novel camera systems and modules, advanced 3D depth sensing technologies, metalens optics for cameras and sensors, and innovative display technologies.