PhD Topic Proposal: Cross-Modal dTOF Robustification with RGB sensors

1. Introduction and Motivation

The proliferation of autonomous systems, from domestic robots to advanced driver-assistance systems and augmented reality, has created an immense demand for reliable, compact, and cost-effective 3D sensing. Direct Time-of-Flight (dTOF) sensors are emerging as a key enabling technology, offering direct depth measurements in a small, low-power package that overcomes the scale ambiguity of passive vision systems. They are used in a wide array of applications, from robotics to laser autofocus for phones.

However, the measurements, extracted from the dToF sensor raw outputs, suffer from significant limitations in real-world environments:

- **Low Spatial Resolution**: Often limited to a sparse grid (e.g., 8x8 zones), the data requires significant upsampling.
- **Measurement Noise**: Precision degrades with distance and ambient light, leading to temporal instability or "jitter".
- **Multipath Interference (MPI)**: Photons reflecting off multiple surfaces (e.g., in corners) before returning to the sensor create erroneous depth readings.
- **Material-Dependent Inaccuracies**: The performance is heavily influenced by the target surface. Very dark, shiny, or transparent surfaces can lead to highly inaccurate measurements.

Concurrently, high-resolution RGB sensors are ubiquitous and provide rich textural, structural, and semantic information. Recent research, such as the DELTAR model [1], has shown impressive results by fusing these two streams to produce high-quality, dense depth maps. These methods excel at depth super-resolution, effectively "filling in the gaps" of the sparse dTOF data.

However, a critical research gap remains. Current fusion models often implicitly trust the dTOF measurements as sparse but accurate anchors. They focus on *enhancing* the data's resolution rather than *validating* its correctness. When a dTOF measurement, extracted from the raw data, is fundamentally flawed, these models risk propagating the error, creating a detailed but inaccurate depth map. This proposal argues for a paradigm shift: from depth enhancement to **depth robustification**. The core objective is to develop a fusion framework where RGB data is used to actively diagnose, challenge, and correct error prone dTOF measurements, leading to a final depth output that is not only dense but also fundamentally more reliable.

2. Problem Statement

This research addresses the problem of robust depth estimation by fusing a low-resolution dTOF sensor and a high-resolution RGB camera. The primary goal is to produce a dense, accurate, and reliable depth map that is resilient to the inherent failure modes of dTOF technology.

"Robustness" in this context is defined as the system's ability to:

1. **Identify and Mitigate Multipath Interference:** Use geometric cues from the RGB image to disambiguate multiple photon returns registered by the dTOF sensor.
2. **Correct for Material-Dependent Errors:** Recognize challenging surface properties (e.g., specularity, low reflectivity) in the RGB domain and adjust the fusion strategy accordingly.
3. **Reduce Measurement Noise and Jitter:** Leverage strong structural priors from the RGB image to regularize and stabilize noisy depth readings.

3. State-of-the-Art and Research Gap

This research is situated at the intersection of several computer vision fields:

- **Depth Super-Resolution and Completion**: This is the most closely related area. Methods like PENet [2], PrDepth [3], and DELTAR [1] have pioneered techniques to generate dense depth maps from sparse inputs and an RGB guide image. DELTAR shows the power of using the full depth distribution from the sensor rather than just a single mean value. However, these methods are primarily designed for upsampling, not error correction.
- **Monocular Depth Estimation**: Techniques like AdaBins estimate depth from a single RGB image. While impressive, they suffer from scale ambiguity and can produce geometrically inconsistent results, especially in regions with ambiguous textures.
- **Sensor Fusion**: Deep learning has enabled fusion through feature concatenation or sophisticated attention-based mechanisms.

The **critical gap** this research will address is the lack of a framework that performs *cross-modal validation* as a core part of the fusion process. Furthermore, training such a system is hampered by the difficulty of obtaining dense, pixel-aligned ground truth for dTOF *errors*. This project will tackle both challenges by creating a novel simulation pipeline to train a new class of robust fusion networks.

4. Proposed Methodology

The project is centered on two core activities: first, creating a comprehensive training and validation dataset through advanced simulation and real-world capture; second, developing a novel neural network for robust fusion.

1. Simulation and Data Acquisition Pipeline A primary obstacle to learning robust fusion is the lack of large-scale data that captures dTOF failure modes with corresponding ground truth. This research will create a high-fidelity simulation and data capture pipeline.

- **Physics-Based dTOF Simulation**: Using a modern rendering engine (e.g., Blender, Unreal Engine 5) with path tracing capabilities, we will develop a simulator that goes beyond rendering perfect depth maps. It will simulate the physical process of dTOF sensing by tracing photon paths, modeling their interaction with complex materials (specularity, subsurface scattering, absorption), and explicitly simulating the mechanism of multipath interference. This will generate a massive dataset of RGB images paired with realistic, error-prone dTOF histograms and dense ground truth depth.
- **Real-World Data Capture Rig**: To validate the simulator and capture real-world sensor noise characteristics, a physical data acquisition rig will be built. It will consist of a synchronized and geometrically calibrated setup including the dTOF sensor, a high-resolution RGB camera, and a high-fidelity ground truth sensor (e.g., a structured light scanner). This rig will be used to capture a diverse validation dataset of challenging real-world scenes.

2. Probabilistic dTOF Feature Extraction As demonstrated by DELTAR [1], the raw histogram of photon returns contains richer information than a single extracted depth value. We will use this full distribution as our input, designing a PointNet-like architecture to extract a powerful feature vector that encodes not just a single estimated depth value but also the measurement's confidence and potential multimodality.

3. Cross-Modal Anomaly Detection Module This is the central innovation in the network architecture. A Transformer-based module will be trained on our simulated dataset to

explicitly learn the relationship between RGB image patches and their corresponding dTOF distributions. By attending to both modalities simultaneously, this module will learn to identify patterns of disagreement indicative of sensor errors. The output will be an internal, spatially-varying **confidence map** that the network uses to weigh the trustworthiness of the dTOF readings.

4. Confidence-Guided Fusion Network The final depth map will be generated by a fusion network that takes three inputs: the RGB image features, the probabilistic dTOF features, and the internal dTOF confidence map. This network will learn to dynamically adjust its strategy based on this confidence. In high-confidence regions, it will heavily rely on the accurate dTOF data for metric precision. In low-confidence regions, it will down-weight the dTOF input and rely more on monocular depth cues from the RGB image, using surrounding trusted depth points for context and scale.

5. Research Objectives and Expected Contributions

Key Research Questions:

- How can the physical sensing process of a dTOF sensor, including its specific error modes like MPI, be accurately simulated to generate training data?
- What cross-modal patterns exist between RGB appearance and dTOF measurement errors, and how can a neural network trained on simulated data learn to detect them in real images?
- How can a fusion architecture explicitly model and utilize a measure of sensor confidence to dynamically arbitrate between sensor data and learned image priors?

Expected Contributions:

1. **A High-Fidelity dTOF Simulation Framework**: A novel and powerful tool for generating realistic training data for dTOF-based tasks, which will be a valuable contribution to the research community.
2. A **Benchmark Dataset**: A challenging dataset, combining simulated and real-world captures, designed to test the robustness of depth fusion algorithms.
3. **A Novel Robust Fusion Architecture**: An end-to-end network that pioneers confidence-aware fusion for robustifying dTOF sensors, moving beyond simple super-resolution.
4. **Advancement in Low-Cost 3D Sensing**: By significantly improving the reliability of commodity dTOF sensors, this research will broaden their applicability in robotics, AR, laser autofocus, etc...

Commenté [OP1]: Is this really something that ST wants?

Commenté [MA2R1]: changed

References

[1] Li, Y., et al.: DELTAR for Accurate Depth Estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). (2022).

[2] Hu, Mu, et al. "Penet: Towards precise and efficient image guided depth completion." *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021.

[3] Han, Chenggong, et al. "PRDepth: Pose Refinement Enhancement-Based Monocular Depth Estimation for Indoor Scenes." *IEEE Transactions on Instrumentation and Measurement* (2025).